Even-odd crossover: a new crossover operator for improving the accuracy of students' performance prediction

Somia A. Shams[1], Asmaa Hekal Omar[1], Abeer S. Desuky[1], Mohammad T. Abou-Kreisha[2], Gaber A. Elsharawy[1]
[1]Department of Mathematics, Faculty of Science, Al-Azhar University, Cairo, Egypt
[2]Department of Mathematics, Faculty of Science, Al-Azhar University, Cairo, Egypt

Article Info

Article history:

Received Mar 22, 2022
Revised May 28, 2022
Accepted Jun 15, 2022

Keywords:

Crossover
Imbalanced data
Machine learning
Prediction
Students' performance

ABSTRACT

Prediction using machine learning has evolved due to its impact on providing valuable and intuitive feedback. It has covered a wide range of areas for predicting student' performance. Instructors can track student's dropout in a particular course at an early stage and try to improve students' performance. The problem of students' future performance prediction using advanced statistics and machine learning is a hard problem due to the imbalanced nature of the student data where the number of students who passed the exam is generally much higher than the number of students who failed the exam. This paper proposes a new type of crossover operator called Even-Odd crossover to generate new instances into the minority class to handle the imbalanced data problem. The experiments are implemented using three machine learning (ML) algorithms: random forest (RF), support vector machines (SVM), and K-Nearest-Neighbor (KNN) to ensure the efficiency of the proposed technique. The performance of the classifiers is evaluated using several performance measures. The efficient ability of the proposed method on solving the imbalance problem is proved by performing the experiments on 22 real-world datasets from different fields and four students' datasets. The proposed Even-Odd crossover shows superior performance compared to state-of-the-art resampling techniques.

Corresponding Author:

Somia A. Shams
Department of Mathematics, Faculty of Science, Al-Azhar University
Nasr City, Cairo 11754, Egypt
Email: somia.shams10@azhar.edu.eg

1. INTRODUCTION

Machine learning has been applied in many fields, such as prediction with categorize customer behavior in marketing or sales, detection fraud or omission in banks, the diagnosis of diseases in medicine, and more recently in education [1]. Data mining and machine learning are very useful in the field of education, especially to analyze the performance of students [2]. Machine learning techniques in educational data mining (EDM) aim to develop a model to discover hidden patterns and explore useful information from educational settings. Universities can use EDM to predict which students will pass or fail and have poor educational performance, to see who will pass examinations in certain subjects, and to obtain the percentage of graduates [3]. Since it is normal for the number of successful students to be greater than the number of failed students, a new problem arises is data imbalance that is one of the important challenges in data mining for dealing with data in classification. Imbalanced data sets are referred to the situation where there are too many examples in one category than the others [4]. Some classes may contain a large amount of data called majority classes and some may contain only a few instances of data called minority classes. Minority class samples are usually poorly predicted by various machine learning models since the supervised learning algorithm always pay attention to the samples of the majority category when the general classification model achieves high accuracy. The effect of the defect on classification is deadly, and the effect increases with the expansion of the task [5]. This problem appears in many real-world applications, such as healthcare sector, detection of oil spill, fraud detection in usage of credit cards, modeling of cultures, intrusion detection in networks, and categorization of texts [6]. Since the minority class is particularly important, so attention should be paid to it [7]. The optimal goal of many of machine learning algorithms is maximizing the overall accuracy, which is the percentage of correct predictions made by the classifier. This results in classification performance with high accuracy but very low sensitivity towards positive class.

Therefore, the optimal goal must be shifted towards maximizing sensitivity of minority class and majority class separately instead of with a focus on overall accuracy [8]. The classification in class imbalanced datasets has drawn great concern in the students' performance because often the classes of instances that are passed are significantly more than the classes of instances that are failed. To enhance the classification performance in this field, many methods have been made and still made [9]. Some rebalancing methods for pretreatment have been suggested in the past, especially in the aspects of artificial expansion of minority class examples (over-sample), resampling by decreasing the number of majority class examples (under-sample), or a combination of them [7]. The basic concept of sampling methods is to provide balanced stratification of imbalanced datasets. It is impossible to predict what the true class distribution should be. A very powerful technique called synthetic minority over-sampling technique (SMOTE) [10] was proposed for balancing imbalanced data sets. Cluster-SMOTE [11], another method in the category of technologies that focuses on specific stratigraphic areas, uses k- means grouping of minority class before applying SMOTE within existing clusters. A few years later, two new synthetic minority over-sampling were suggested, they are called borderline-SMOTe1 and borderline-SMOTe2 [12]. Safe-Level-SMOTE, Safe-Level-Synthetic Minority Over-sampling technique [13], assigns each positive case its safe level before delivery synthetic samples. Each synthetic sample is placed near the largest vault level so that all synthetic samples are created only in safe areas. A survey on methods for solving data imbalance problem for classification presented in [14].

Literature mainly focuses on two fronts: defining the most important features for predicting student performance and finding the best prediction method to improve prediction accuracy [15] common attributes used in predicting student performance, researchers discussed their factors and categorized them as either internal or external [16]. Attributes such as assignment marks, exams, class tests, and attendance are categorized as internal evaluation [17]. In terms of external evaluation, one needs to mention the student demographics such as gender, age, family background, special needs and interactions with learning environment [18] Several machine learning algorithms have been used to predict student performance. The effect of a classification algorithm is usually related to the characteristics of data. support vector machine (SVM) [19] is a widely used classification algorithm. For college academic performance three algorithms decision tree (DT), neural network (NN) and SVM were used to predict students' performance where data metrics included online time, internet frequency, internet volume online traffic and usage behaviors, which correlate with academic performance. Results showed that the most accurate is the SVM algorithm when predicting success and failure score (69-73%), followed by NN (68-71%) and DT (60-62%) [20]. Random Forest (RF) [21] was applied to predict which students would get bachelor's degree based on courses attended and completed in the first two semesters of the academic year, The dataset contains information regarding several courses taken by undergraduate students at a Canadian university [22], [23]. Authors in [24] focused on identifying dropout students using data mining (DM) approach in online application. They applied four algorithms, KNN, DT and navie baise (NB) and NN. KNN performed the best among all classifiers, with an accuracy of 87%. The proposed method in [25] based on investigating student learning performance from learning management system (LMS) data using five algorithms, The results showed the performance from the Random Forest as the best accuracy value is 90%. Proposal in [25] suggested solving the class imbalance problem at work in the future. This study proposed the even-odd crossover method for solving data imbalance problem. We apply this proposed method on 22 datasets with different imbalance ratios as well as the number and feature type, and four real world student datasets. Two of these student datasets are collected from students' grades system of the faculty of science, Al-Azhar University for two subjects of physics. The experiments are compared to show the superiority of our proposed method against various methods of resampling data using three classifiers random forest (RF), KNN, support vector machine (SVM). The performance of the model using various classifiers with our proposed show the best results compared to others resampling methods.

2. METHOD

This paper proposed a novel technique based on a new genetic algorithms' crossover operator named Even-Odd crossover. The proposed method is based on finding a new oversampling method based on crossover to create new minority class samples from the old samples. The process of crossover ensures the exchange of attribute values between the samples and thus creates new samples are like the old samples. It based mainly on exchanging the attribute values between samples at each even or odd location. the proposed method solves the problem of imbalanced data by generating new samples which make the data to be balanced. The function of the remainder of the division has been used in the position of each attribute in the samples, if the remainder of the division equal zero, the value between the samples is exchanged otherwise the value remains as showed in Figure 1 that presents all method steps. Figure 2(a) shows two samples each of them has 6 attributes and 2(b) two new samples after applying the proposed method.

```
Algorithm1:   Even-Odd crossover (proposed method)
Begin
          Initialize the parameters, variable size n,
          v→ First vector (v1,v2,…vn) to be crossed over.
          w→ First vector (w1,w2,…wn) to be crossed over.
             For I from 1 to n     do
                {  If  I is even        Then
                     swap the values of vi, wi }
                   End If
          End  For
          return v, w
End
```

Figure 1. Algorithm of the proposed method

Sample 1: $value_1$,$value_2$,$value_3$,$value_4$,$value_5$,$value_6$
Sample 2: $value_7$,$value_8$,$value_9$,$value_{10}$,$value_{11}$,$value_{12}$

(a)

New-sample1: $value_1$,$value_8$,$value_3$,$value_{10}$,$value_5$,$value_{12}$
New-sample2: $value_7$,$value_2$,$value_9$,$value_4$,$value_{11}$,$value_6$

(b)

Figure 2. The effect of proposed method on any two samples, (a) two samples with six attributes and (b) even-odd crossover operation (proposed method)

2.1. Datasets description

The experiments were applied using two real-world educational datasets that are collected from the students' grading system of the Faculty of Science, Al-Azhar University for two subjects of physics. Data was collected of the students using faculty reports and questionnaires, with the collection of their grades from the faculty's grading system. The dataset consists of 92 learners in 20 features with no missing values and one response variable. The features were classified into three main divisions: (a) demographic features that consisted of gender, birthplace and residence while studying (b) academic status attributes such as studying system and (c) behavioral attributes such as external courses, practical presence, father's job, mother's job and total income. The response variable had two classes namely pass or fail. Table 1 describes the data.

To ensure the validation of the proposed method, the experiments are firstly applied on 22 datasets (Table 2) with different imbalance ratios as well as the number and types of features and two real student datasets (Table 3) that collected from two secondary schools of Portuguese (Gabriel Pereira (GP) and Mousinho da Silveira (MS)). The dataset contains attributes for students like academic grades, social attributes, demographic attributes, and school-related attributes. Data was collected from the students using the school reports and questionnaires and used in recent paper [26].

Table 1. Describes these two datasets.

Datasets	Instances	No of attributes	No of Majority	No of Minority
Ph211	92	20	79	13
Ph212	92	20	83	9

Table 2. Description of 22 real-world datasets

Datasets	Instances	No. of attributes	No. of Majority	No. of Minority
abalone(5-other)	4174	9	4059	115
abalone9-18	731	9	689	42
abalone19	4174	9	4142	32
Ecoli1	336	8	259	77
Ecoli2	336	8	284	52
Ecoli3	336	8	301	35
Ecoli4	336	8	316	20
ecoli-0-1-3-7_vs_2-6	281	8	274	7
Page-blocks13vs2	472	11	444	28
Page-blocks(4-other)	5472	11	5385	87
poker-8_vs_6	1477	11	1460	17
ThoraricSurgery	470	28	400	70
Transfusion	748	5	570	178
Vehicle1	846	19	629	217
Vehicle3	846	19	634	212
yeast3	1484	9	1321	163
yeast4	1484	9	1433	51
yeast5	1484	9	1440	44
Yeast6	1484	9	1449	35
yeast-1_vs_7	459	8	429	30
yeast-2_vs_4	514	9	463	51
Yeast (POX-others)	1484	9	1464	20

Table 3. Description of two real student datasets

Datasets	Instances	No of attributes	No of Majority	No of Minority
student-port-binary	649	33	549	100
student-mat-binary	395	33	265	130

2.2. Data preprocessing

The input data are pre-processed by cleaning up the missing values, and converting the nominal data into numeric data, converting the output data into a binary class (0 means failure,1 means success), and split the data set into two parts: training and testing data sets (75%: 25%) with the same ratio of minority and majority class and without any feature selection for any data designation.

2.3. Proposed methods

This paper solves the imbalanced data problem using the proposed Even-odd crossover oversampling method. and compares the results with various methods of resampling such as (SMOTE, SLSMOTE, Cluster-SMOTE, Bor-SMOTE). The best method of resampling and the best classifier are selected after comparing all methods. The experiments are applied on two stages: first stage, all the classifiers (RF, KNN, SVM) are applied on the imbalanced data to show the effect of the imbalanced data problem on the models' performance. Second stage, all the classifiers are implemented on balanced data that generated by resampling methods to obtain a better perception of the effectiveness of the resampling methods as ways to solve the imbalanced problem.
Phases of Proposed method:
Phase 1: in this phase Precision, Recall, F-score, G-mean are calculated for imbalanced data using Random Forest (RF), KNN and SVM classifiers. Parameters that were used are k value in KNN=3, k-fold=10. Number of bags in RF nBag=100.
Phase 2: in this phase the same performance measures are calculated for RF, KNN, SVM classifiers applied on the resampled data with the state-of-the-art oversampling algorithms (SMOTE, SLSMOTE, Cluster-SMOTE, Bor-SMOTE).
Phase 3: KNN, SVM, RF algorithms is applied in this phase on the resampled data using the proposed Even-odd crossover oversampling method.
Phase 4: in this phase, comparison is done between phase 1, phase 2 and phase 3 results.
Phase 5: The best method is selected and applied in real datasets.

2.4. Performance evaluation

Classification performance evaluation is an important reference for classification algorithms [27]. Rating indexes applied to classify balanced data are no longer appropriate for classifying imbalanced data. Additional classification performance evaluation indicators, such as Precision, Recall, F-score and G-mean, are often used to measure the imbalanced classification performance [28]. If TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives; the classification performance measures are as follows:

$$pression = \frac{TP}{TP+FP} \tag{1}$$

$$recall = \frac{TP}{TP+FN} \tag{2}$$

$$F - score = 2 * \frac{precision * recall}{precision + recall} \tag{3}$$

$$G - mean = \sqrt{\frac{TP}{(TP+FN)} * \frac{TN}{(FP+TN)}} \tag{4}$$

The performance measures are used to evaluate the performance of original, SMOTE, SLSMOTE, Cluster-SMOTE, Bor-SMOTE, and proposed method using classifiers.

3. RESULTS AND DISCUSSION

Our experiments were performed to find the classification performance measures (Accuracy, Precision, Recall, F-score and G-mean) for the three classifiers-RF, KNN and SVM- applied firstly on the 22 real-world datasets from different fields then two real datasets of students [26] and finally two real students' datasets (Ph211, Ph212) collected from the faculty reports and questionnaires, and their grades from the faculty grading system. With imbalanced datasets, often increases in Recall come at the cost of decreases in Precision, since in order to increase the TP for the minority class, also the number of FP is often increased, resulting in reduced Precision. F-score provides a way to combine both Recall and Precision into a single score that achieves both properties and provides a way to express them with a single measure that can give a good indication to the classification of imbalanced data [28], [29].

First, each classifier is applied to the imbalanced data and performance measures are calculated for the (original) test dataset, the second stage is applied to the training data by performing (SMOTE, SLSMOTE, Cluster-SMOTE, Bor-SMOTE and proposed method) to balance the data and recalculate the performance measures for the same test dataset. When applying RF to abalone(5-other), abolone19, Ecoli1, and Page-blocks(4-other) datasets, proposed method achieved the best result in performance measures except for accuracy which achieved by Bor-SMOTE and the same result when applying KNN and SVM in first two mentioned datasets but the third when applying KNN and when applying SVM, best Precision is achieved by Bor-SMOTE and best accuracy is achieved by Cluster-SMOTE unlike when applying SVM to third dataset best performance in all measure is achieved by the proposed method and when apply KNN to last, the proposed method achieved the best result in performance measures except accuracy achieved by Bor-SMOTE.

In abalone 9-18 and yeast 4, when applying the three classifiers, proposed method achieved the best result in Recall, F-score and G-mean but Precision achieved the best value by Cluster-SMOTE . Bor-SMOTE achieved best result in accuracy with applying RF and with the two other classifier SL-SMOTE achieve best result in Accuracy and Precision. When apply RF, KNN and SVM in Ecoli2, Page-blocks13vs2 datasets proposed method achieved best performance except accuracy and Precision is achieved by Bor-SMOTE.

In Ecoli3 dataset with applying three classifiers, proposed method achieved best performance in Recall, F-score, and G-mean but accuracy and Precision is achieved by SLSMOTE with applying RF and Precisions is achieved by Bor-SMOTE and accuracy is achieved by SLSMOTE with KNN, and best accuracy is achieved by Cluster-SMOTE and for Precision is achieved by proposed method. When apply RF, KNN and SVM in Ecoli4 proposed method achieved best performance except accuracy and Precision is achieved by Cluster-SMOTE.

In ecoli-0-1-3-7_vs_2-6 dataset when apply RF and SVM, proposed method achieved the best performance in Precision, Recall, F-score, G-mean but best accuracy is achieved by Bor-SMOTE but when apply KNN best accuracy and Precision is achieved by Bor-SMOTE or Cluster-SMOTE. Poker-8_vs_6 dataset when applying RF and KNN, the proposed method achieved best performance in Precision, Recall, F-score, G-mean but best accuracy is achieved by Bor-SMOTE unlike applying SVM proposed method achieved only best accuracy.

In ThoraricSurgery dataset, the proposed method achieved best performance measure with applying RF and SVM except accuracy is achieved using Cluster-SMOTE. also, when applying KNN classifier proposed method achieved the best performance measures except accuracy is achieved by Bor-SMOTE and Precision is achieved by SMOTE. In Transfusion database, Recall, G-mean, and F-score are achieved by proposed method with applying three classifiers, but best accuracy and Precision are achieved by Cluster-SMOTE with applying SVM.

In Vehicle1 and Vehicle3 datasets, when applying RF, KNN, and SVM performance measures are achieved by the proposed method except accuracy is achieved by SLSMOTE with applying RF and best Precision is achieved by Cluster-SMOTE with applying SVM in the second dataset. In yeast3 dataset when applying the three classifiers, best accuracy and Precision are achieved by SLSMOTE and Bor-SMOTE respectively but RF achieved the best result in Recall, F-score, and G-mean .When applying three classifiers to yeast5 dataset, the best accuracy and Precision are achieved by SVM with applying Cluster-SMOTE, but Recall, F-score, and G-mean is achieved by the proposed method when applying RF.

In Yeast6, yeast-1_vs_7, and yeast-2_vs_4 are achieved best accuracy and precision at most with Bor-SMOTE or Cluster-SMOTE but Recall, G-mean, and F-score achieved with the proposed method.Yeast(POX-others) achieved the best accuracy and Precision with SLSMOTE when applying RF and KNN but Recall, F-score, and G-mean are achieved the best result with the proposed method with SVM.

Figures 3-5 show that the highest performance is achieved by the proposed method in Recall, F-score and G-mean in all datasets, then SMOTE and SLSMOTE in 14 and 11 datasets respectively. In other hand the worest result achieved by Bor-SMOTE and Cluster-SMOTE. Figure 3 shows the proposed method achieved best Precision in 11 datasets and best accuracy in only two datasets.

When appling all methods of resampling data the result shows RF achieved the best performance than SVM and KNN as RF achieved the best accuracy, Precicion, Recall, F-score, and G-mean in 15, 12, 17, 16, and 18 datasets respectively.

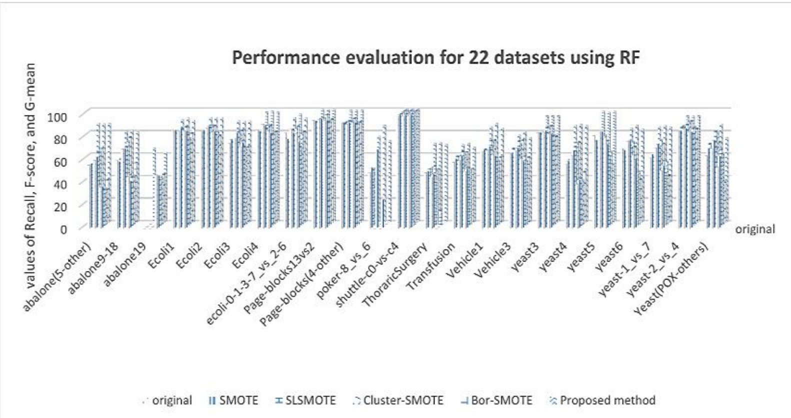

Figure 3. Performance evaluation of the proposed method and the best results of the state-of-the-art oversampling algorithms using RF on 22 datasets

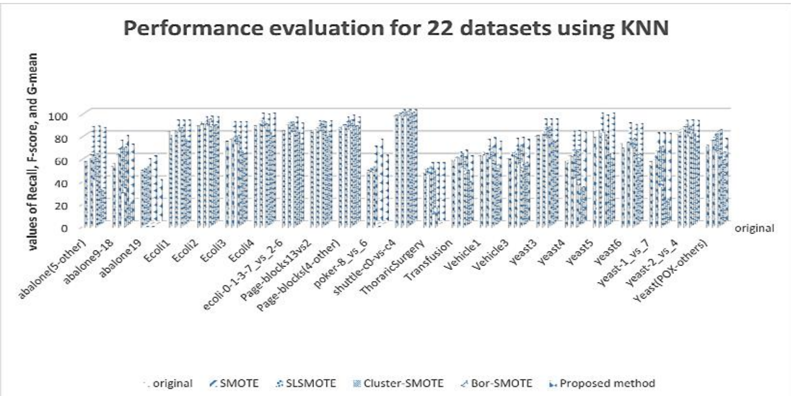

Figure 4. Performance evaluation of the proposed method and the best results of the state-of-the-art oversampling algorithms using KNN on 22 datasets

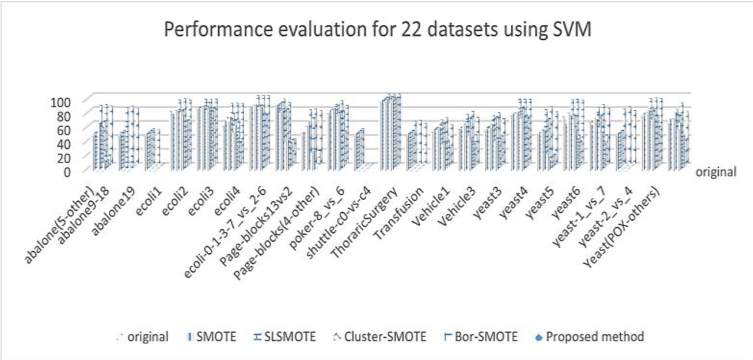

Figure 5. Performance evaluation of the proposed method and the best results of the state-of-the-art oversampling algorithms using SVM on 22 datasets

Experiments were also performed to calculate the performance measures on two real datasets of students used in recent paper [26]. In the experiments, same steps are applied. The results that show the efficiency of the proposed method compared to the other the state-of-the-art methods using RF, KNN, and SVM classifiers are presented in Figures 6, 7, and 8 respectivly.

Figure 6 shows that the proposed method has the highest performance in Precision, Recall, F-score and G-mean using RF classifier. Figure 7 shows that the proposed method has the highest performance in Precision, Recall, F-score and G-mean using KNN classifier on student-port-binary datasets, it also has highest performance in Recall and G-mean using KNN classifier on sapfile-binary dataset. Figure 8 shows that the proposed method has the highest performance in Recall, F-score and G-mean using SVM classifier on student-port-binary datasets, it has highest performance in Recall and G-mean using SVM classifier on sapfile-binary dataset. Tables 4-6 show the performance evaluation for two real students' datasets (Ph211, Ph212) collected from the faculty reports and questionnaires, and their grades from the faculty grading system. In the experiments, same steps are applied.

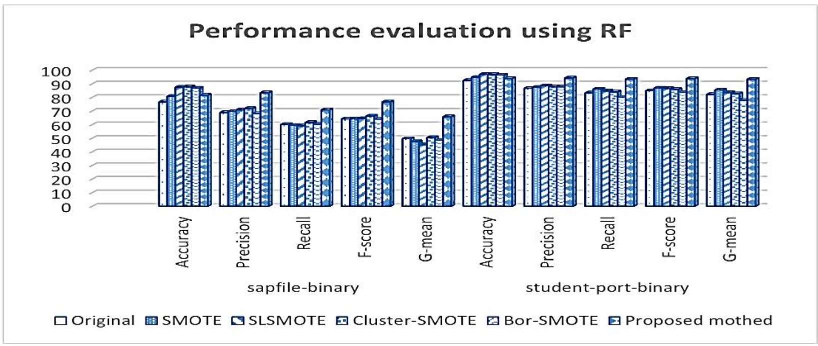

Figure 6. Performance evaluation of proposed method and the best results of the state-of-the-art oversampling algorithms on two real student datasets using RF

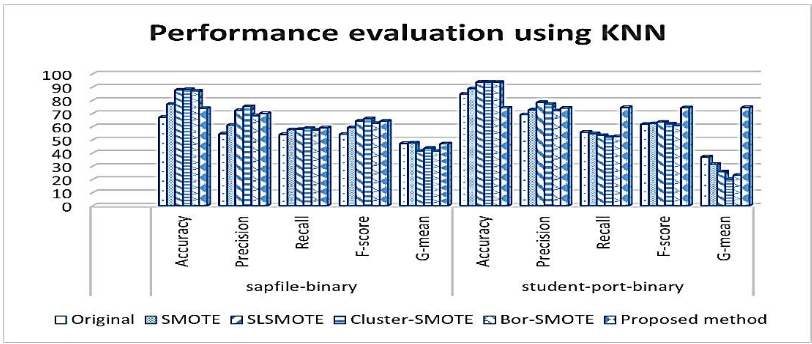

Figure 7. Performance evaluation of proposed method and the best results of the state-of-the-art oversampling algorithms on two real student datasets using KNN

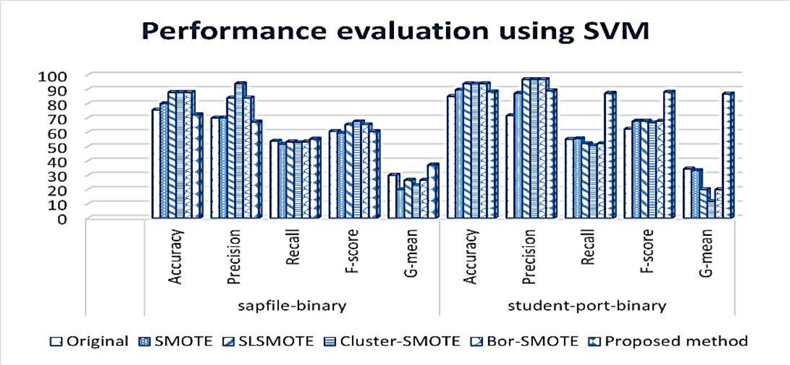

Figure 8. Performance evaluation of proposed method and the best results of the state-of-the-art oversampling algorithms on two real student datasets using SVM

Table 4. Performance evaluation for original, SMOTE, SLSMOTE, Cluster-SMOTE, Bor- SMOTE, and proposed method on collected dataset (Ph211, Ph212) using RF

Dataset	RF	original	SMOTE	SLSMOTE	Cluster-SMOTE	Bor-SMOTE	Proposed method
Ph211	Accuracy	84.8889	89.8077	92.9739	92.9739	92.9739	75.5
	Precision	63.6782	70.1613	47.0414	47.0414	47.0414	77.037
	Recall	55.7936	53.4037	49.3789	49.3789	49.3789	75.0572
	F-score	59.4757	60.6462	48.1818	48.1818	48.1818	76.0342
	G-mean	38.4713	27.612	0	0	0	74.108
Ph212	Accuracy	90.2222	93.0769	95.6433	95.6433	96.1988	81.6667
	Precision	NAN	NAN	48.0769	48.0769	NAN	83.7171
	Recall	50	50	49.7159	49.7159	50	84.1667
	F-score	NAN	NAN	48.8827	48.8827	NAN	83.9413
	G-mean	0	0	0	0	0	83.666

Table 5. Performance evaluation for original, SMOTE, SLSMOTE, Cluster-SMOTE, Bor- SMOTE, and proposed method on collected dataset (Ph211, Ph212) using KNN

Dataset	KNN	original	SMOTE	SLSMOTE	Cluster-SMOTE	Bor-SMOTE	Proposed method
Ph211	Accuracy	78.3333	86.6667	91.2092	92.9739	92.3856	55
	Precision	48.3266	44.6721	46.988	47.0414	47.0238	53.3333
	Recall	48.7829	48.2301	48.4472	49.3789	49.0683	52.74
	F-score	48.5537	46.383	47.7064	48.1818	48.0243	53.038
	G-mean	26.2932	0	0	0	0	48.3125
Ph212	Accuracy	89.1111	92.3077	96.1988	96.1988	96.1988	65
	Precision	45.0549	46.5116	NAN	NAN	NAN	65.5229
	Recall	49.3976	49.5868	50	50	50	65.8333
	F-score	47.1264	48	NAN	NAN	NAN	65.6777
	G-mean	0	0	0	0	0	65.8281

Table 6. Performance evaluation for original, SMOTE, SLSMOTE, Cluster-SMOTE, Bor- SMOTE, and proposed method on collected dataset (Ph211, Ph212) using SVM

Dataset	SVM	original	SMOTE	SLSMOTE	Cluster-SMOTE	Bor-SMOTE	Proposed method
Ph211	Accuracy	86	89.7436	**94.1503**	**94.1503**	**94.1503**	61.5
	Precision	NaN	NaN	NaN	NaN	NaN	**61.4118**
	Recall	50	50	50	50	50	**61.0984**
	F-score	NaN	NaN	NaN	NaN	NaN	**61.2547**
	G-mean	0	0	0	0	0	**60.5089**
Ph212	Accuracy	90.2222	93.0769	**96.1988**	**96.1988**	**96.1988**	70.8333
	Precision	NAN	NAN	NAN	NAN	NAN	**71.2418**
	Recall	50	50	50	50	50	**71.6667**
	F-score	NAN	NAN	NAN	NAN	NAN	**71.4536**
	G-mean	0	0	0	0	0	**71.6473**

The results show the efficiency of the proposed method in term of Precision, Recall, F-score and G-mean compared to the other the state-of-the-art methods. Figure 9 shows that the proposed method using RF achieves the best performance than SVM and KNN.

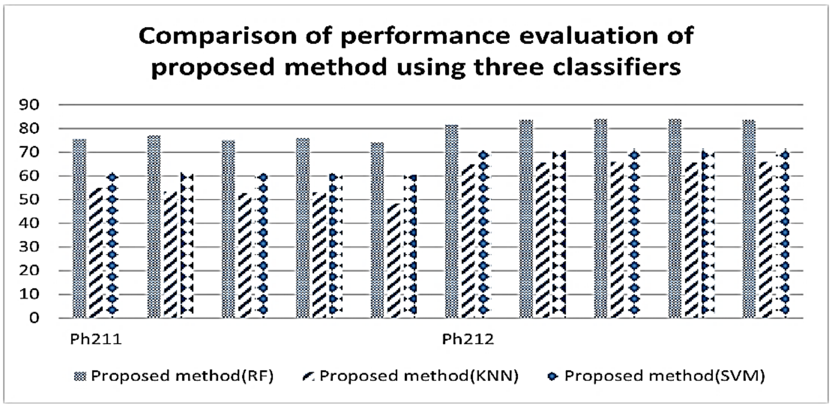

Figure 9. Comparison of performance evaluation of proposed method on collected datasets (Ph211, Ph212) using three classifiers

4. CONCLUSION

Student's performance prediction is an important process to improve the educational quality which is vital to help students to improve their academic performance. In this paper, we proposed Even-Odd crossover to solve the imbalance problem in student datasets. The imbalanced data sets are classified with SVM, RF, and K-nearest neighbor classification algorithm. The experiments applied first to 22 real-world datasets from different fields with different imbalance ratios and various distributions to ensure the proposal validation. Then, four students' imbalanced data sets were used. We collected two of these students' educational datasets from student grading system of the faculty of science, Al-Azhar university. The experimental results show that the Even-Odd crossover oversampling method has a superior performance regarding rebalanced data classification compared to other state-of-the-art oversampling methods SMOTE, SLSMOTE, Cluster-SMOTE, and Bor-SMOTE.

REFERENCES

[1] M. A. T. Rony, M. M. Hassan, E. Ahmed, A. Karim, S. Azam and D. S. A. A. Reza, "Identifying Long-Term Deposit Customers: A Machine Learning Approach," *2021 2nd International Informatics and Software Engineering Conference (IISEC)*, 2021, pp. 1-6, doi: 10.1109/IISEC54230.2021.9672452.

[2] M. Kumar and A. J. Singh, "Performance Analysis of Students Using Machine Learning & Data Mining Approach," *Predicting Students Academic Performance using Data Mining Techniques: A Case Study*, 2019.

[3] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student Performance Prediction Model based on Supervised Machine Learning Algorithms," in *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 3, Nov. 2020, doi: 10.1088/1757-899X/928/3/032019.

[4] D. Jahin, I. J. Emu, S. Akter, M. J. A. Patwary, M. A. S. Bhuiyan and M. H. Miraz, "A Novel Oversampling Technique to Solve Class Imbalance Problem: A Case Study of Students' Grades Evaluation," *2021 International Conference on Computing, Networking, Telecommunications & Engineering Sciences Applications (CoNTESA)*, 2021, pp. 69-75, doi: 10.1109/CoNTESA52813.2021.9657151.

[5] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 4, pp. 687–719, Jun. 2009, doi: 10.1142/S0218001409007326.

[6] H. Patel, D. Singh Rajput, G. Thippa Reddy, C. Iwendi, A. Kashif Bashir, and O. Jo, "A review on classification of imbalanced data for wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 16, no. 4, Apr. 2020. doi: 10.1177/1550147720916404.

[7] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of Classification Methods on Unbalanced Data Sets," *IEEE Access*, vol. 9, pp. 64606–64628, 2021, doi: 10.1109/ACCESS.2021.3074243.

[8] E. AT, A. M, A.-M. F, and S. M, "Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method," *Global Journal of Technology and Optimization*, vol. 01, no. S1, 2016, doi: 10.4172/2229-8711.s1111.

[9] A. Mahani and A. R. B. Ali, "Classification Problem in Imbalanced Datasets," *Recent Trends in Computational Intelligence*, 2019, doi: 10.5772/intechopen.89603.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321-357, 2002, doi: 10.1613/jair.953.

[11] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine Learning Research*, vol. 7, pp. 1-30, 2006.
[12] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," *International Conference on Intelligent Computing*, vol. 3644, pp. 878-887, 2005, doi: 10.1007/1153805991.
[13] R. Goebel, J. Siekmann, and W. Wahlster, "Lecture Notes in Artificial Intelligence," in *Proceedings 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand*, vol. 6863, April 2009.
[14] A. Singh and A. Purohit, "A Survey on Methods for Solving Data Imbalance Problem for Classification," *International Journal of Computer Applications*, vol. 127, no. 15, 2015, doi: 10.5120/ijca2015906677.
[15] H. Zeineddine, U. Braendle, and A. Farah, "Enhancing Prediction of Student Success: Automated Machine Learning Approach," *Computers & Electrical Engineering*, vol. 89, p. 106903, doi: 10.1016/j.compeleceng.2020.106903.
[16] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," in *Procedia Computer Science*, vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.
[17] Z. K. Papamitsiou, V. Terzis, and A. A. Economides, "Temporal learning analytics for computer based testing," in *ACM International Conference Proceeding Series*, pp. 31–35, 2014, doi: 10.1145/2567574.2567609.
[18] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores," *Computational Intelligence and Neuroscience*, vol. 2018, 2018, doi: 10.1155/2018/6347186.
[19] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.
[20] K. Kim, H. S. Kim, J. Shim, and J. S. Park, "A study in the early prediction of ict literacy ratings using sustainability in data mining techniques," *Sustainability (Switzerland)*, vol. 13, no. 4, pp. 1–11, Feb. 2021, doi: 10.3390/su13042141.
[21] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[22] C. Beaulac and J. S. Rosenthal, "Predicting University Students' Academic Success and Major Using Random Forests," *Research in Higher Education*, vol. 60, no. 7, pp. 1048–1064, Nov. 2019, doi: 10.1007/s11162-019-09546-y.
[23] G. Czibula, A. Mihai, and L. M. Crivei, "S PRAR: A novel relational association rule mining classification model applied for academic performance prediction," in *Procedia Computer Science*, 2019, vol. 159, pp. 20–29. doi: 10.1016/j.procs.2019.09.156.
[24] E. Yukselturk, S. Ozekes, and Y. K. Türel, "Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program," *European Journal of Open, Distance and E-Learning*, vol. 17, no. 1, pp. 118–133, Jul. 2014, doi: 10.2478/eurodl-2014-0008.
[25] A. Dutt and M. Akmar Ismail, "Can We Predict Student Learning Performance from LMS data? A Classification Approach," *Proceedings of the 3rd International Conference on Current Issues in Education (ICCIE 2018)*, 2019, doi: 10.2991/iccie-18.2019.5.
[26] A. S, Y. Salal, and M. Kumar, "Educational Data Mining: Student Performance Prediction in Academic," *Conference: International Conference on Engineering , Technology and Management for Sustainable Development(ICETMSD-2018)At: DEV BHOOMI GROUP OF INSTITUTIONS, Dev Bhoomi Campus, Chakrata Road Navgaon, Manduwala, Dehradun- 248007,* 2019.
[27] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Applied Sciences (Switzerland)*, vol. 11, no. 1, pp. 1–28, Jan. 01, 2021. doi: 10.3390/app11010237.
[28] W. Xie, G. Liang, Z. Dong, B. Tan, and B. Zhang, "An Improved Oversampling Algorithm Based on the Samples' Selection Strategy for Classifying Imbalanced Data," *Mathematical Problems in Engineering*, vol. 2019, 2019, doi: 10.1155/2019/3526539.
[29] S. Belarouci and M. A. Chikh, "Medical imbalanced data classification," *Advances in Science Technology and Engineering Systems Journal*, vol. 2, no. 3, pp. 116-124, doi: 10.25046/aj020316.

BIOGRAPHIES OF AUTHORS

Somia A. Shams administrator in Computer Science, Faculty of Science, Al-Azhar University, Cairo, Egypt. She can be contacted at email: somia.shams10@azhar.edu.eg.

Asmaa Hekal Omar received the B.Sc., M.Sc., and Ph.D. degrees from the Mathematics Department, Faculty of Science, Al-Azhar University, Cairo, Egypt, in 2002, 2009, and 2013, respectively, where she is currently pursuing the Ph.D. degree in computer science. She has coauthored publications in different journals. Her research interests include machine learning and optimization. She can be contacted at email: Asmaahekl@azhar.edu.eg.

Abeer S. Desuky received the B.Sc. degree in science, in 2003, and the M.Sc. and Ph.D. degrees in computer science, in 2008 and 2012, respectively. She is currently an Associate Professor in computer science with the Mathematics Department, Faculty of Science, Al-Azhar University, Cairo, Egypt. She has published several research papers in the field of AI, machine learning, meta-heuristic optimization, and data mining and analysis. She is also a supervisor of some master's and Ph.D. theses. She is a reviewer in many Scopus-indexed journals, such as IEEE Access and Peerj. She can be contacted at email: abeerdesuky@azhar.edu.eg.

Mohammad T. Abou-Kreisha received the B.Sc., pre-M.Sc., M.Sc., and Ph.D. degrees in Computational Mathematics from the faculty of Science, Cairo University, Egypt, in 1997, 1999, 2003, and 2008 respectively. He was involved as an administrator of Computer Science, Assistant of lecturer, Lecturer, Associate Professor of Computer Science in the Mathematical Department, Faculty of Science, Al-Azhar University, Cairo, Egypt, in 1999, 2003, 2008, and 2017 respectively. And from 2010 till now, he has been an Executive Manager of the E-Learning Program for Islamic and Arabic Sciences, Al-Azhar University. His teaching and research interest are Introduction to computer science, multimedia processing, discrete mathematics, artificial intelligent, algorithms, cryptography, data structure, database, file processing, java, asp.net, c#, vb, matlab, php-mysql, and dynamic web site development.He can be contacted ae email: mt@alazhar.edu.eg.

Gaber A Elsharawy Professor of computer science at Faculty of science, Al Azhar University. Ph.D. in Computer and Systems Engineering, Faculty of Engineering, Al Azhar University. M.Sc.in computer system U.S. Air Force University, Air Force Institute of Technology (AFIT), Dayton, Ohio, USA. Author of many publications in the fields of Database management systems, artificial intelligent, modeling & simulation, and programming languages. He can be contacted ae email: Gaber.Ahmed@buc.edu.eg.