GOS: A Genetic OverSampling Algorithm for Classification of Quranic Verses

Bassam Arkok
Kulliyyah of Information and Communication Technology
International Islamic University Malaysia, Malaysia
bassam.arkok@gmail.com

Akram M. Zeki
Kulliyyah of Information and Communication Technology
International Islamic University Malaysia, Malaysia
akramzeki@iium.edu.my

2022 13th International Conference on Information and Communication Systems (ICICS) | 978-1-6654-8097-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICS55353.2022.9811224

***Abstract*—Imbalanced classes problem is a problem in many datasets in real applications, where one class "minority class" contain few numbers of samples and the other "majority class" contain many numbers of samples. It is difficult to build a training model to classify the imbalanced classes correctly due to tending the accuracy of classification to the majority class. In this paper, a new technique is called "GOS: a Genetic OverSampling algorithm", is proposed using a genetic algorithm. A genetic algorithm is applied to oversample the imbalanced datasets and to improve the performance of imbalanced classification. This improvement is achieved due to adjusting the locations of samples in the minority class in the optimal places. According to the experimental results obtained, the GOS algorithm outperformed other techniques used widely in the imbalanced classification field.**

Keywords— Imbalanced Classification, Re-sampling techniques, Quranic Topics, Genetic Algorithm

I. INTRODUCTION

The Qur'an is the first source of religious text for 1.6 billion Muslims worldwide. It is a revelation from Allah SWT to Prophet Muhammad (Peace Be Upon Him) 1443 years ago in classical Arabic language. It is divided into 114 chapters (Surah) and 6236 verses (Ayah). The Qur'an contains around 77,000 words which are organised into 114 varying-sized chapters called Surah. Each chapter contains varying-sized verses or Ayat (in total over 6200 verses). The Qur'an was revealed verbally from Allah to Prophet Mohammed (PBUH) through the angel Jibril. The Qur'an is the most important miracle that allows us to believe in the message of Prophet Mohammad (PBUH) to conclude a series of the prophets' divine messages from Prophet Adam (PBUH) until Prophet Mohammed (PBUH). Furthermore, the Qur'an contains valuable information and gives answers and solutions to many problems facing mankind, and it is free from discrepancies and contradictions. In various places, the Qur'an has challenged mankind to author a book or a chapter of a book that would resemble the Qur'an in content and style. Scholars in the past fifteen centuries had been authoring books highlighting various linguistic, stylistic, scientific, rhetorical, and hidden discoveries from the Qur'an. These scholars relied on their knowledge and familiarity with the Qur'an as there were no computational tools available. The Qur'an is characterised by vast information in unstructured and scattered, yet conceptually-related verses. Furthermore, Quranic verses are presented as either instruction or narrative and have an underlying deep connection that joins the entire text into one whole concept [1]. The essential ideas and meaning of the Qur'an overlap from chapter to chapter and verse to verse, thus finding out the implied connections would need time and more in-depth study to discover the hidden topics [2] as there are many topics found. These topics can be obtained from books that index the Quranic verses according to their content. These topics were extracted manually by the Muslim scholars that their books are usually similar in terms of indexing, hence common topics can be found. However, many uncommon topics have been found due to the addition by scholars to their indexes. Therefore, all Qur'anic topics are challenging to collect manually in one index. The manual extraction of Quranic topics is difficult and time-consuming. Therefore, many Islamic scholars have attempted to classify the Qur'anic text automatically to reveal certain information according to specific topics. This classification can be called "Topical classification of the Qur'anic text" which assigns one or more Qur'anic topics automatically according to their content. The Qur'anic classification can be considered one of the critical research in Machine Learning due to the differences in the number of verses. As it explores imbalanced classification, the classification performance will be weak when the number of samples in the classes is unequal. Imbalanced classification has been implemented on many imbalanced datasets because of its capability to classify imbalanced classes that cannot be solved by traditional classification. Hence, imbalanced classification can be applied on Quranic topics as the Qur'an contains both a high number of verses and a low number of verses.

Therefore, this paper aims to classify the qur'anic verses based on imbalanced classification techniques. The second objective is to design a new oversampling method using a genetic algorithm to locate the samples of the minority classes in the optimal places and then oversample them for the rebalancing issue. To sum up, the main contribution of this research is to apply genetic algorithms to the qur'anic verses to obtain better performance for the Qur'anic classification. Also, the genetic algorithm in this research applied a new fitness function which depends on the selection of the best locations in the minority classes.

II. LITERATURE REVIEW

Many studies had applied the GA processes to create artificial samples to oversample the minority class. For example, [3] proposed a new oversampling method based on K mean to cluster the samples of minority class and use GA to generate new examples of the minority class. [4] applied GA directly to gain a new instance for the minority class. [5] used SVM to generate support vectors and a draft hyperplane, then, implemented GA to create new data points in the classification margin or sensible area. Finally, SVM was used again to determine the best hyperplane of the data points created. [6] presented a novel SVG classification approach to detect splice sites. The proposed method gets new artificial instances from SVs obtained in the first stage and includes just the samples that improve SVM performance in the data set as the GA is used to evaluate and get better instances at each iteration. [7] proposed a new genetic algorithm based on SMOTE (GASMOTE) algorithm. The proposed approach

uses different sampling rates for other samples in the minority class and determines a combination of the optimal sampling rates. GA is applied here to find optimised sampling rates and create a new dataset through oversampling using the optimised rates. In another study [8], GenSample was proposed to oversample the samples of minority classes using GA. GenSample creates examples of synthetic minority based on the difficulty in learning a sample point as the performance will be improved when the oversampling is conducted via this point. The algorithm is terminated when the desired imbalanced ratio is achieved, or the performance is reduced by adding a new synthetic data point to the dataset.

All above studies applied genetic algorithms to oversample the samples of minority classes in many datasets but no study used the genetic algorithms to overcome the problem of imbalanced classification based on the Quranic verses. So, this paper aimed to propose a new technique that depends on the operators of genetic algorithms to rebalance the quranic topics that have a difference in the number of their verses. The proposed method is called "A Genetic OverSampling algorithm". GOS method aimed also to optimize the minority class by relocating the samples of the minority class to the optimal places. the optimal locations of samples during the generation may improve the performance of imbalanced classification.

III. GOS: A GENETIC OVERSAMPLING ALGORITHM

The proposed algorithm that is called "GOS: a Genetic OverSampling algorithm", is explained in this section to overcome the problem of imbalanced classification. This algorithm is aimed to improve the performance of imbalanced classification by applying GA's operators. The operators of genetic algorithms are used in this paper to locate the samples of the minority class in the optimal places. Also, GA's operators are applied to rebalance the imbalanced datasets via the oversampling process for the samples of the minority class.

The pseudo-code of the GOS method and the GOS's procedures are as follows:

1- Let n= the number of samples of the majority class.

2- Select samples from the minority class that satisfy the fitness function condition to assign them as the current population.

3- Apply the crossover operator for the current population to generate the offspring samples.

4- Assign the samples of off-springs that satisfy the fitness function condition to assign them to the current population.

5- Apply mutation operator for the samples of off-springs that did not satisfy the fitness function condition to assign them for the current population when they achieve the fitness function condition.

5- Go to step 3 if the optimal samples are not equal to n else stop the GOS algorithm.

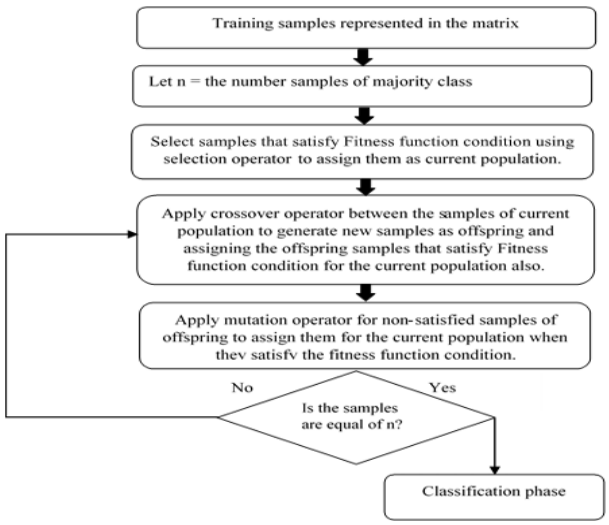

Fig. 1 clarifies the genetic procedures of the GOS method

IV. GA'S OPERATORS IN GOS

Operators of the genetic algorithm were used in this work to design new oversampling methods. GOS method applied these operators to obtain optimised training datasets due to the genetic search to get the optimal samples. First, GOS removes the non-optimal samples in the minority class, after that the oversampling process is implemented for this class to be equal to the number of samples in the original majority class.

The genetic processes of the proposed method "GOS" are as follows:

A. Fitness Function

To define the fitness function for the proposed GOS method, the mean or centre sample for the training dataset is explored. Then, the fitness value F for each sample is computed by measuring the distance between the mean sample and the other samples. Euclidean distance is used to calculate the distances between the mean sample and the samples.
According to the experiments implemented for all collected datasets, the value of x ranged from 1.5 to 9.5, which increased by 0.5 in each iteration. Searching for the best F value to assign it for the training dataset was done by generating many optimised copies for the training datasets based on GA operators using F values. The best F value with the higher performance was chosen.
As is noticed in the Literature review section, the previous studies used different fitness functions to evaluate the genetic samples. While In this research, the locations factor was used, as a new method, to measure the samples' distances in the minority classes from the mean sample. According to this condition, the best locations of samples will be selected; which are supposed that contribute to improving the imbalanced classification performance.

B. Selection Operator

The samples are selected from the training dataset/current population based on their fitness function condition. In this research, the examples were selected when their fitness

values were less than or equal to the F value. The selected samples were passed to the next phase to mate them and generate new examples as offspring. The instances not chosen in this phase were forwarded to the mutation operator later to be modified and rechecked.

C. Crossover Operator

Single point crossover operator was conducted for the selected samples to generate the offspring. The offspring samples were assigned to the current population when their fitness values were less than or equal to the F value. Then, the crossover operator was prepared for the new population again to generate new samples. To reduce overgeneration in the oversampled samples, the mating process among the parent samples is conducted uniquely. Therefore, the samples of parents will not be mated throughout the mating period again. This policy is done via the division of parents into two parts, and each parent marries a sample that they have met from the other part only. For more control of this mating between parent samples, a variable *shift_k* is responsible for controlling the mating. The control variable is set to 0 and incremented to 1 in every iteration to ensure the implementation of non-repeated mating. This controlling solves two cases of the mating where the samples have increased after the process of mating and the samples have not increased where there are no new samples in the current population.

D. Mutation Operator

Inversion Mutation was applied to the unsatisfied samples of the fitness function condition during the selection process over time.

The processed samples were tested again to select them for the next population when their fitness values were less or equal to the condition of the fitness function. The genetic proceses are repeated until the Qura'nic datasets are balanced.

As is noticed in the genetic algorithms, they are computationally expensive i.e. time-consuming. This is the limitation of the proposed technique, but on the other side, the Qur'anic classification will be improved a higher.

V. Experimental Procedure

Here, the experimental procedures of this work will be presented in the following subsections.

A. Datasets

The GOS algorithm was evaluated on 8 binary Quranic datasets with different imbalance ratios. The Quranic datasets were applied by other studies [9-11] that were extracted from the Quranic Index collected by Dr Abu Akhir [12]. Table I presents these datasets with their imbalanced ratios and the number of verses in the majority and minority classes.

TABLE I. DESCRIBES THE QURANIC DATASETS USED

	Datasets	# of majority class	# of minority class	# of features	IR
1	Islam_Faith	1120	761	117	1.47
2	Prayer_Zakat	51	27	19	1.89
3	Tawheed_Shirk	241	80	48	3.01
4	Labor_Science	369	74	90	4.99
5	Q_stories_Political_R	176	32	76	5.5
6	Prophecy_Religion	128	35	35	3.66
7	Organization of financial Relations_Call_to_Allah	130	81	51	1.6
8	Jihad_Fasting	46	29	32	1.59

B. Division of Datasets

The collected datasets were divided into 70% as training data and 30% as testing data. the imbalanced issue of classes was taken into the consideration, it's mean that 70% of each class was taken for the training set and 30% of each class for the testing set.

C. Applied Classifiers

9 classifiers were applied with the proposed oversampling method "GOS" to pick the best classifier and to prove the robustness of the proposed method. The required classifiers are: KNN, Random Tree, Naive Bayes, J48, LibSVM, Random Forest, SMO, and Simple CART, Voted perceptron.

D. Evaluation Metrics

The performance of imbalanced classification techniques is evaluated by many metrics, four of these metrics were used in this research to evaluate the performance of the GOS method and the other resampling techniques. The overall accuracy metric is no longer valid to evaluate the performance of imbalanced classification techniques [13-15]. The overall accuracy metric evaluates the performance of classification depending on the accuracy of the majority and minority classes together. So, this metric will be not a good choice if the accuracies of these classes are very different. Thus, the performance of classifiers will tend to the samples of majority classes while the minority classes will obtain poor results when the overall accuracy metric is used.

The used evaluation metrics are:

- Sensitivity/Recall (also called the True Positive Rate, the accuracy of positive samples, or recall): measures the proportion of the positive sample, which is determined correctly.

$$\text{Sensitivity/Recall}=TP/(TP+FN) \quad (1)$$

- Specificity (also called the True Negative Rate and accuracy of negative samples): measures the proportion of the negative instances identified correctly.

$$\text{Specificity}=TN/(TN+FP) \quad (2)$$

- G-mean: G-Mean is a geometric mean for recall and specificity. For the binary classes, when accuracies of the classification are balanced, the accuracy will be maximized for each of them by G-mean.

G-mean= √ (Sensitivity × Specificity) (4)

• MCC: Matthews's correlation coefficient (MCC) considers the false and true positives and the negatives. It is a correlation coefficient among the predicted binary classifications, and the truly observed that it returns a value in (−1 and +1). A coefficient of +1 indicates that the prediction is perfect, while 0 value indicates that the prediction is worse, while −1 indicates total disagreement between the observation and prediction

MCC=(TP × TN-FP × FN)/ √ ((TP+FP) × (TP+FN) × (TN+FP) × (TN+FN)) (5)

VI. Experimental Results

GOS method is compared to the other techniques in this section. Table 5.3 shows the comparison results of GOS with SMOTE, ROS, and RUS methods in the four-evaluation metrics: Sensitivity/Recall, Specificity, G_Mean, and MCC. The experimental results contain the average results of each classifier in all datasets based on the used resampling method. For example, the results of SMOTE method comprise the results average of SMOTE method in all datasets of nine classifiers. In Table II, it is shown the expriemental results obtained for the proposed method and the other techniques.

TABLE II. PRESENTS THE EXPRIEMENTAL RESULTS OBTAINED

	Sensitivity/ Recall	**Specificity**	**G-Mean**	**MCC**
AVG. SMOTE	**0.86**	0.74	0.79	0.59
AVG. RUS	0.79	0.76	0.77	0.53
AVG. ROS	0.84	0.70	0.76	0.55
AVG. GOS	0.84	**0.79**	**0.81**	**0.62**

As seen in the above table, GOS had a higher performance among the others. It outperformed the others in all metrics. GOS improved by 3% from the second-best method in the performance in Specificity and MCC metrics. It improved by 2% in the Sensitivity/Recall and G-Mean metric. Figure 2 illustrates GOS' improvement and outperformance than the other methods. GOS outperformed the others in Specificity, G-mean, and MCC metrics.

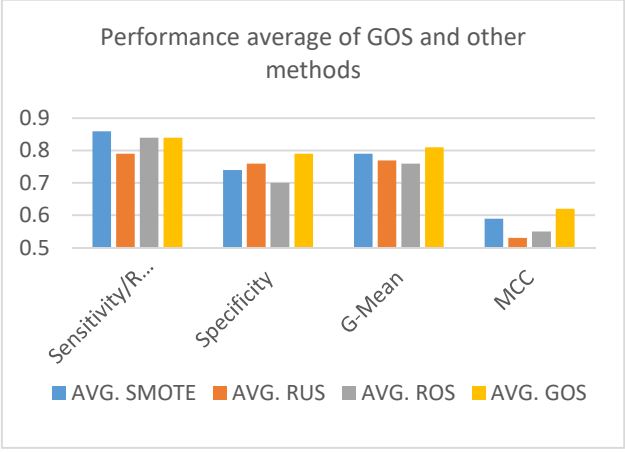

Fig.2 illustrates GOS's improvement and slight outperformance than the other methods

From the above figure, it is noticed that the GOS method outperformed the other resampling methods remarkably in all metrics except in the Sensitivity/Recall metric. SMOTE method was the best method in the Sensitivity/Recall metric, but it obtained a poor performance method in the Specificity metric. It means that SMOTE method is biased to the majority classes and paid less attention to the minority classes. Also, we can see that the GOS method outperformed the others significantly in the MCC metric.

VII. Conclusion

Imbalanced classification has been used widely in real data, which is applied to classify imbalanced datasets. The Quranic verses can be considered imbalanced datasets because the quranic topics vary in the number of their verses. Therefore, the problem of imbalanced classification can occur through the quranic classification. On the other side, the genetic operators have been exploited to overcome the problem of imbalanced classification to rebalance the imbalanced dataset. This study proposed a new oversample "GOS" based on the genetic algorithm to relocate the samples of the minority class in the optimal places and oversample these samples at the same time. The proposed method outperformed other resampling techniques that are used widely in the imbalanced learning techniques. GOS and the other resampling methods were evaluated by Sensitivity/Recall, Specificity, G_Mean, and MCC.

Acknowledgment

This paper was supported by International Islamic University of Malaysia under (FRGS19-083-0691) research project.

References

[1] Siddiqui, M.A., S.M. Faraz, and S.A. Sattar. Discovering the thematic structure of the Quran using probabilistic topic model. in 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences. 2013. IEEE.

[2] Salloum, S.A., et al., A survey of Arabic text mining, in Intelligent Natural Language Processing: Trends and Applications. 2018, Springer. p. 417-431.

[3] Benchaji, I., S. Douzi, and B. El Ouahidi. Using genetic algorithm to improve classification of imbalanced datasets for credit card fraud detection. in International Conference on Advanced Information Technology, Services and Systems. 2018. Springer.

[4] Beckmann, M., B.S.L. de Lima, and N.F. Ebecken. Genetic algorithms as a pre processing strategy for imbalanced datasets. in Proceedings of the 13th annual conference companion on Genetic and evolutionary computation. 2011.

[5] Cervantes, J., X. Li, and W. Yu. Using genetic algorithm to improve classification accuracy on imbalanced data. in 2013 IEEE International Conference on Systems, Man, and Cybernetics. 2013. IEEE.

[6] Cervantes, J., et al. A New Approach to Detect Splice-Sites Based on Support Vector Machines and a Genetic Algorithm. in Iberoamerican Congress on Pattern Recognition. 2013. Springer.

[7] Jiang, K., J. Lu, and K. Xia, A novel algorithm for imbalance data classification based on genetic algorithm improved SMOTE. Arabian journal for science and engineering, 2016. 41(8): p. 3255-3266.

[8] Karia, V., et al., GenSample: A Genetic Algorithm for Oversampling in Imbalanced Datasets. arXiv preprint arXiv:1910.10806, 2019.

[9] Arkok, B. and A.M. Zeki. Classification of Quranic Topics Using Ensemble Learning. in 2021 8th International Conference on Computer and Communication Engineering (ICCCE). 2021. IEEE.

[10] Arkok, B. and A.M. Zeki, Classification of Qur'anic topics based on imbalanced classification. Indonesian Journal of Electrical Engineering and Computer Science, 2021. 22(2): p. 678-687.

[11] Arkok, B. and A.M. Zeki. Classification of Quranic Topics Using SMOTE Technique. in 2021 International Conference of Modern

Trends in Information and Communication Technology Industry (MTICTI). 2021. IEEE.

[12] Al-Khair, A.A. and M.A. Kabbani, Quran teacher intonation. The Tunisian Company for Distribution, 2003.

[13] Tang, Y., et al., SVMs modeling for highly imbalanced classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2009. 39(1): p. 281-288.

[14] Patel, H. and G.S. Thakur, Classification of imbalanced data using a modified fuzzy-neighbor weighted approach. International Journal of Intelligent Engineering and Systems, 2017. 10(1): p. 56-64.

[15] Al-Azani, S. and E.-S.M. El-Alfy. Using Word Embedding and Ensemble Learning for Highly Imbalanced Data Sentiment Analysis in Short Arabic Text. in ANT/SEIT. 2017.