ISSN 1533-9211

DOI 10.5281/zenodo.7388773

GP-MOMS BASED CLASSIFICATION OF MULTI-CLASS IMBALANCED DATA

RADHIKA KOTECHA

Department of Information Technology, K. J. Somaiya Institute of Engineering and Information Technology, University of Mumbai, India. Email: kotecha.radhika7@gmail.com

Abstract:

Prediction of rarely occurring patterns is challenging but crucial for several real-world applications like healthcare, fraud detection, etc. However, for datasets with imbalanced class distribution, the traditional techniques in Machine Learning focus mainly on frequently occurring patterns, and exhibit poor performance in classifying instances of underrepresented classes present in minority. Further, most research in this field focuses on binary classes only. But, several applications of interest involve multiple classes, which is much more complex than learning from bi-class imbalanced datasets. Hence, the proposed work addresses the issue of multi-class imbalanced data classification through a generic framework suitable for all application areas. Firstly, the work extends the bi-class evaluation measures to multi-class datasets for unbiased performance analysis. Further, a sampling and Genetic Programming based approach named GP-MOMS is proposed for efficient classification of multi-class imbalanced data, especially the rare patterns. Performance comparison with related benchmark techniques on standard datasets proves the efficacy of the proposed approach, which is presented in this work.

Keywords: Classification, Minority Classes, Imbalanced Datasets, Multi-Class, Genetic Programming, Sampling.

1. INTRODUCTION

Learning patterns from imbalanced data remains one of the major challenges experienced in several real-world applications. For classification problems in Machine Learning, the data is considered “class imbalanced” when training examples from at least one class are rare. The class(es) represented by small number of examples are known as minority class(es) whereas the other class(es) that make up the rest are called majority class(es). Some examples of problems that encounter such embedded difficulty of imbalanced data include medical diagnosis (only a minority of patients will be having a disease), fraud detection (fraudulent transactions occur less), loan approval (only few applicants will default), etc. Protein classification, churn prediction, fault diagnosis, action recognition from videos, anomaly detection, etc. are other common problems affected by uneven class distributions [1]–[8].

The uneven distribution causes the classifier to exhibit poor learning of underrepresented classes than of the prevalent classes. This in turn results into higher misclassification of test examples representing minority classes as compared to the majority classes. However, in most of the real-world classification tasks, the key classes-of-interest are the minority classes. Hence, accurate classification of such minority class examples is equally and in fact more important and crucial than classifying examples from other majority classes.

Traditional classifiers ignore the class distribution and simply aim to produce maximum accuracy, given some dataset. That is, they ignore the minority classes. For example, if a dataset contains 95% records with class A and 5% samples with class B, the traditional classifiers

ISSN 1533-9211
DOI 10.5281/zenodo.7388773

might correctly classify all 95% records of the majority class A, but misclassify all the records of minority class B. The resultant accuracy would be 95%, but all the examples of rare — but class-of-interest, that is, class B will remain misclassified. Major reasons behind such biased behaviour of traditional classifiers are:

1. When tiny clusters of minority class instances are present, they are most likely to be ignored or identified as noise / outliers — thus resulting into increased classification error.
2. Modern algorithms that utilize classifier performance for guiding the global pattern learning generally employ accuracy or coverage of instances as a performance measure — which again provides an advantage to the majority class.

Thus, traditional classifiers are not apt for such imbalanced data and some approaches to deal with it are proposed over the last decade. These approaches can be divided into three main categories:

1. Data-level approaches: External approaches that aim to transform the original imbalanced dataset into an artificially class-balanced set by under-sampling the majority class examples and over-sampling the minority class samples.
2. Algorithmic-level approaches: Internal approaches wherein existing algorithms are adapted to reinforce the learning towards the minority classes with an objective to be more attuned to class imbalance issues.
3. Hybrid approaches: These incorporate approaches at both data-level and algorithmic-level combined, considering higher misclassification costs for minority class instances, and hence attempting to minimize misclassification errors.

However, all these approaches have multiple limitations and confined applicability. Moreover, the performance of these existing techniques degrades when applied to multi-class imbalanced classification problems [9], [10]. This results into a major issue as most real-world tasks include multiple classes rather than only binary divisions.

Hence, the paper aims to produce accurate classification results on multi-class classification problems with imbalanced datasets. The paper proposes the use of the evolutionary algorithms – specifically, Genetic Programming (GP) for construction of a classifier that deals with such multi-class imbalanced datasets. The proposed approach codifies and evolves individuals (patterns) for GP using interpretable decision trees, as well as optimizes the classifier performance on minority class instances.

The rest of the paper is organized as follows: Section 2 discusses the related work for imbalanced data classification. Section 3 presents measures for evaluation of classifier performance on imbalanced multi-class datasets. Further, section 4 proposes GP-MOMS based approach to classification of imbalanced multi-class datasets. Several experimental results that demonstrate the performance of the proposed algorithm are stated in Section 5. Conclusion and future work are presented in Section 6.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7388773

2. RELATED WORK

Prior to proposing the GP-MOMS based proposed approach, existing solutions that attempt to address the class imbalance issues were studied. These approaches are used as a base for further experimental analysis.

In the specialized literature [11], several under-sampling, and over-sampling methods have been applied to tackle the problem of class imbalance. These resampling techniques methods are preferred as they are independent of the underlying classifier. However, under-sampling causes a loss of potentially important patterns whereas over-sampling often results into over-fitting.

To address the above issues, some variant approaches for resampling have been proposed. One of the popular methods is - Synthetic Minority Oversampling Technique (SMOTE) [12], which synthetically generates new minority class instances by interpolating prevalent minority class instances using k-Nearest Neighbour. The k nearest neighbours is chosen based on the amount of oversampling required. However, SMOTE does not consider the minority class distribution and suffers from over-generalization in some cases. Its variants with ability of adaptive sampling like MSMOTE [13], Borderline-SMOTE [14], Adaptive Synthetic Sampling [15], Safe-Level-SMOTE [16], WEMOTE [17], MLSMOTE [18], etc. are also implemented by researchers. These approaches have demonstrated improved performance, but are observed to have high time-complexity.

Several studies have been done to identify and address the learning difficulties with traditional classifiers. In [19], the authors propose using Support Vector Machine (SVM) to eliminate redundant majority class instances. Authors in [20] propose discarding majority class examples far from the decision boundaries. An approach to combine over-sampling and under-sampling is proposed in [21], wherein the goal is to generate a balanced dataset by assuring a trade-off between loss of information and addition of synthetic examples. A noise-filtered under-sampling approach is proposed in [22] to discard noisy instances, followed by resampling of the remaining minority class instances.

Many researchers [23]–[26] have applied ensemble methods for learning from class imbalanced data. Hybridization of ensemble methods with sampling or data-level approaches result into improved classification and have proved to be robust for challenging imbalanced data. However, as majority of these approaches are heuristic-based, the performance of ensembles are unstable for minority classes. A taxonomy depicting application of ensemble learning for imbalanced data is presented in [27]. One of the branches focus on developing cost-sensitive ensembles wherein variants of Boosting are used to guide the minimization of misclassification costs. The other branch focuses on methods that embed data pre-processing with ensemble learning – Bagging, Boosting, and Hybrid ensembles. Methods like RUS Boost, SMOTE Boost, etc. have proven to be simpler and efficient approaches; however, there are scopes for performance improvement.

Krawczyk et al. [28] discuss the open challenges in classifying imbalanced data in real-time with computational efficiency, and provide future research directions for data stream mining,

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7388773

classification, clustering, regression, big data analytics, etc. It is analysed that along with disparity in instances per class, the classifier performance is also affected due to occurrence of difficult minority class instances. The authors have also highlighted the need to address outliers or noise in minority class instances.

Alberto Fernández et al. [[29]] discuss the issue of learning specifically from imbalanced Big Data, focusing on challenge of volume. The authors analyse various MapReduce-based algorithmic implementations as well as the behaviour of standard data pre-processing techniques such as under-sampling. The challenges in each have been presented for further research.

Lopez et al. [30] identify 06 issues in conjunction with class imbalanced ratio that affect the performance of the classifier, which are: a) class overlapping, b) lack of information in training data, c) noisy data, d) presence of small disjuncts, e) differences in the data distribution for the training and test data, and f) management of borderline instances. The article highlights how these issues result into a biased classification and present directions for future research focusing on data properties.

Authors in [31] present feature selection as an approach to address class imbalanced issue – wherein features that can accurately classify minority class instances are chosen and provided as input to the algorithm. Decision trees that have implicit feature selection are chosen with Weighted Gini Index (WGI) as an attribute selection measure. Performance comparison with λ^2, traditional Gini Index, and F-statistic demonstrate the efficacy of the proposed WGI based approach. However, determining the optimal weights for WGI remains an open issue. The approach also needs analysis for cases where multiple attributes have the same values for the selection measure.

Recently, Deep Learning classifiers have demonstrated promising results for addressing several issues like vanishing gradient, high dimensional data, etc. The features implicitly embedded via multiple layers of deep classifiers are used to obtain separability between the classes. However, when the training dataset is imbalanced, the deep features may not be adequately embedded to establish the required borderline between various classes. Many researchers have applied variants of Generative Adversarial Networks (GANs) [32]–[36] to address this issue. In [37], instead of the employing adversarial relationship between classifier and GAN, the generator is trained in cooperation with the classifier to produce minority class instances that gradually expand the minority decision region, and enhance classification of imbalanced data. However, the computational complexity of the approach is high.

As the goal of this work is to produce optimal classification for all the instances – majority and minority classes of the imbalanced data, evolutionary algorithms, and specifically Genetic Programming (GP) [38] has also been referred. GP or in general evolutionary algorithms are search heuristics based on the notion of natural selection; and are used to produce optimal or near-optimal solutions to complex problems. GP maintains a pool of population representing candidate solutions to the given problem. Each solution is assigned a fitness value based on their performance over a fitness function. Some of the individuals are directly reproduced for

The Seybold REPORT
ISSN 1533-9211
DOI 10.5281/zenodo.7388773

the next generations whereas other selected individuals undergo crossover and mutation to produce next generation of individuals. Solutions with higher fitness have more chances to be selected and to mate for producing even fitter solutions. Over the generations, better solutions are evolved until a termination criterion of interest is reached. GP has shown remarkable results on numerous classification related tasks. Not only classifier induction, GP has been successfully applied on various pre-processing as well as post-processing classification tasks too. A detailed explanation on GP and its application in classification can be found in [39].

Further, few researchers [40] have attempted to apply GP for the class imbalance problem, but they targeted only binary classification — using discriminant functions to codify individuals. GP has also been employed for multi-class classification, however, its effect on imbalanced dataset is not yet analysed extensively. Based on detailed literature studied and briefed herein, the proposed work employs Genetic Programming based approach along with the strengths of the data level solutions for effective classification of imbalanced data.

3. EVALUATION IN IMBALANCED DOMAINS

The challenge in classification of imbalanced data is the requirement of precise evaluation of classifier performance on minority class instances. Hence, this section presents different measures for evaluating classifier performance and reviews their suitability for imbalanced datasets. Further, evaluation measures available for bi-class datasets have been extended to multi-class.

Given some n-class problem, the performance of a classifier can be computed using an n x n confusion matrix as shown in Table 1. All the evaluation measures are stated using this confusion matrix, which is a standard approach for describing classification results.

Table 1: Confusion Matrix for N-Class Problem

		Predicted Class			
		Class 1	**Class 2**	**. . .**	**Class n**
Actual Class	Class 1	$\mu_{1,1}$	$\mu_{1,2}$		$\mu_{1,n}$
	Class 2	$\mu_{2,1}$	$\mu_{2,2}$		$\mu_{2,n}$
	. .	. .	. .		. .
	.	.	.		.
	Class n	$\mu_{n,1}$	$\mu_{n,2}$		$\mu_{n,n}$

Here, $\mu_{i,j}$ represents the number of instances of i^{th} class that are predicted as belonging to j^{th} class. It is clear that $\mu_{i,i}$ represents the true prediction (true positive) of instances of i^{th} class. The popular 04 evaluation measures extended for multi-class classification are as follows:

A) Accuracy

The traditional measure to evaluate the performance of a classifier is accuracy. It is defined as the number of instances correctly classified by the classifier as a proportion of total number of

ISSN 1533-9211
DOI 10.5281/zenodo.7388773

instances seen by the classifier. Using the confusion matrix of Table 1, accuracy [41], [42] can be represented as in Equation (1):

$$\text{Accuracy} = \frac{\sum_{i=1}^{n} \mu_{i,i}}{\sum_{i,j=1}^{n} \mu_{i,j}} \qquad (1)$$

However, this overall accuracy alone is not apt for imbalanced datasets as it gives equal importance to all the training examples irrespective of the fact that the examples of minority class are significantly less than the instances of majority class. In addition, as shown in [40], a classifier may give poor accuracy on the minority class but high accuracy on majority class and hence giving a high overall accuracy due to the sway of majority class training examples.

B) G-Mean

In case of class imbalanced datasets, the performance of all classes should equally contribute in forming the evaluation measure. For binary classification cases, G-Mean, that is, the geometric mean of accuracies is measured separately on both the classes. Using the confusion matrix of Table 1, Equation (2) defines G-Mean as an evaluation measure of classifier performance on imbalanced multi-class datasets.

$$\text{G} - \text{Mean} = \left(\prod_{i=1}^{n} \frac{\mu_{i,i}}{\sum_{j=1}^{n} \mu_{i,j}} \right)^{1/n} \qquad (2)$$

As G-Mean takes into consideration the accuracy of each class, it is suitable for evaluating balanced performance on imbalanced datasets.

C) Average Accuracy / Recall

The average accuracy is another good indicator of overall accuracy measure of classifier performance on class imbalanced datasets. Existing solutions for multi-class classification from imbalanced datasets have used average accuracy on bi-class scenario only. Expanding it to multi-class problems, this paper proposes Equation (3) that shows how balanced accuracy can be obtained using the confusion matrix of Table 1. To elaborate, the accuracy of any class i, also known as Recall of class i, refers to the correctly classified instances of class i in relation to the total instances in class i. As there is a need to have better accuracy on all the classes, average accuracy uses the average of classification accuracy obtained for every existing class. Average accuracy is also referred as final Recall score.

$$\text{Average Accuracy / Recall} = \frac{1}{n} \sum_{i=1}^{n} \frac{\mu_{i,i}}{\sum_{j=1}^{n} \mu_{i,j}} \qquad (3)$$

Average accuracy when used as an evaluation measure, gives efficient conclusions about classifier accuracy as it takes into consideration the performance of every class. In [40], weighted average accuracy is proposed for binary classification tasks. This approach of using weighted accuracy allows giving more weight to minority classes.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7388773

D) Precision

Precision of a class i refers to the correctly classified instances of class i in relation to all the instances predicted as belonging to class i. For multi-class classification, considering equal importance of all majority as well as minority classes, the average of Precision for all the classes is considered to calculate the final Precision score, as shown in Equation (4).

$$\mathrm{Precision} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mu_{i,i}}{\sum_{j=1}^{n}\mu_{j,i}} \qquad (4)$$

As Precision considers class-wise result, it is well suited for imbalanced data classification.

E) F-Measure

Another significant measure is F-Measure, which is the harmonic mean of Precision and Recall. F-measure can be evaluated overall or individually for each class by considering Precision and Recall of the respective class. Higher value of F-Measure indicates better performance of the model on the target class. The overall F-Measure metric is defined in Equation (5).

$$\mathrm{F-Measure} = \frac{2 * \mathrm{Recall} * \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}} \qquad (5)$$

Like Precision and Recall, as F-Measure also considers classification cost for each class, it is best recommended for imbalanced data classification.

Several other measures like ROC, Optimized Precision, etc. exist for evaluating the performance of classifiers. However, the paper limits itself to the usage of above detailed measures in the proposed approach as they provide sufficient level of evaluation. In the next section, a Sampling and Genetic Programming based approach is proposed to evolve decision tree classifier by considering G-mean and Average Accuracy measures as fitness criterion.

4. PROPOSED METHODOLOGY

As stated in earlier sections, traditional classifiers perform well on majority classes examples. Moreover, a lot of work has been done to improve accuracies of such traditional classifiers. Most of these works contribute in increasing overall accuracy of majority classes only. The objective of the paper is to obtain accurate classification on minority class examples also. In order to accomplish this objective, the proposed approach focuses on obtaining high accuracy on minority classes at the time of induction itself. Specifically, a hybrid (data and algorithm level) approach that takes imbalanced data as input and produces a classifier with high accuracy on minority classes as well. The proposed approach is presented in Figure 1 and has 05 phases, as elaborated below.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7388773

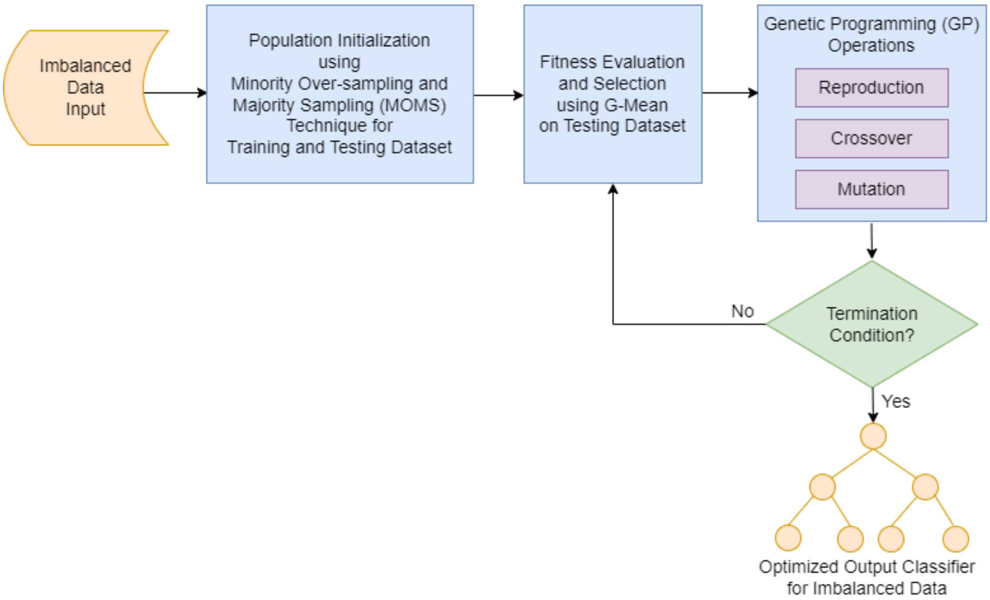


Figure 1: Proposed GP-MOMS Approach for Classification of Multi-class Imbalanced Data

A) Population Initialization

The work proposes Minority Over-sampling and Majority Sampling (MOMS) technique for population initialization. Primarily, the imbalanced dataset that is input for classification is to be divided into training and testing set. For each of these sets, instances from the minority classes are over-sampled and instances belonging to the majority classes are sampled randomly, without replacement. This MOMS technique assures that both the training and testing set have sufficient instances of minority and majority classes. The training set is in turn divided into random non-overlapping subsets for building population of classifiers for GP.

Decision trees are one of the most efficient classification algorithms in literature. Other classification algorithms like kNN, Bayesian Classifiers, etc. also provide good accuracy, but are not comprehensible. On the other hand, along with high accuracy, decision trees also provide comprehensibility, which is the fundamental requirement in real-world applications. Further, GP uses tree structure to codify individuals. Hence, the proposed methodology uses decision trees as base classifiers, that is, to represent individuals of the population. Here, we aim at utilizing GP for classifier induction, that is, to produce a classifier that performs well on multi-class imbalanced datasets.

For initial population of size k, the training dataset is divided into k datasets, each of which is used to produce a decision tree (hence k trees) for the initial population.

B) Fitness Evaluation and Selection

One key aspect of GP is that the fitness function guides the search process. That is, the solutions with higher fitness either are forwarded to the next generation directly by reproduction or are

ISSN 1533-9211

DOI 10.5281/zenodo.7388773

made to undergo crossover and mutation to produce solutions with even better fitness. The proposed approach uses GP to evolve a decision tree classifier with G-Mean and F-measure as a fitness measure. With this approach, solutions that perform well on every class (majority as well as minority) will be preferred over the ones that are biased. Given any multi-class imbalanced dataset, then as per the nature of GP, the final classifier evolved will be the one that performs the best among all the individuals in the population.

C) Genetic Programming Operations

Based on the fitness values, these classification trees are made to undergo reproduction, crossover and mutation to produce the population for new generation. With decision trees as individuals in the population, crossover and mutation are simple to perform – making the proposed approach efficient and effective. Insights on application of GP for crossover and mutation on decision tree classifiers are presented in [43].

D) Termination Condition

As the proposed approach follows GP algorithm, it continues producing newer generations until the termination criterion is reached, which is either obtaining 100% accuracy on all the classes or meeting maximum number of generations, or no improvement in performance for 03 consecutive generations.

E) Output Classification

The best solution obtained when the algorithm terminates, that is, the solution (decision tree classifier) with the highest fitness in the last generation is designated as the solution.

5. EXPERIMENTAL RESULTS

To verify the effectiveness of the proposed GP-MOMS approach, we conducted several experiments with 07 imbalanced datasets with varying number of classes and from various real domains. The datasets have been taken from Machine Learning Repository of the University of California at Irvine [44] and Kaggle datasets [45]. The details about these datasets, along with the ratios of imbalance between the classes are presented in Table 2.

ISSN 1533-9211
DOI 10.5281/zenodo.7388773

Table 2: Composition of Datasets

Dataset	No. of Attributes	No. of Classes	No. of Instances	Imbalanced Ratio
Car Evaluation [44]	6	4	1728	70:22:4:4
Nursery [44]	8	4	12960	33:3:33:31
Glass Identification [44]	10	6	214	33:35:8:6:4:14
Page Blocks Classification [44]	10	5	5473	90:6:1:1:2
Credit Card Fraud Detection [45], [46]	28	2	284807	99:1
Diabetic Retinopathy [47]	64	5	8400	73:12:8:4:3
Thyroid Disease Prediction [44]	30	8	9178	2:6:5:6:4:3:1:73

Implementations of the proposed GP-MOMS approach as well as all the baseline methods used for comparison have been done in MATLAB R2021a. In the experimentations, 10-fold cross-validation has been used for training and testing. Some initial runs were carried out to figure out values of GP parameters used in implementation of the proposed approach. Table 3 presents the finalized GP parameters used for experiments.

Table 3: GP Parameters

Parameter	Value
Initial Population Size	100
Maximum Generations	50
Method of Selecting Individual for GP operation	Tournament Selection with size 7
Crossover Rate	70%
Mutation Rate	5%
Elitism Rate	25%
Fitness Function	G-Mean & F-Measure

The performance of the proposed GP-based approach is compared with 04 other specialized classification techniques. The algorithm level approaches used for comparison are hybridized with data level approaches to predict the minority class instances precisely based on the best-reported outcomes in the literature. These baseline methods used for comparison are: 1) Random Sampling (RS) with Decision Tree classifier, 2) SMOTE with Decision Tree classifier, 3) SMOTE with Bagging Decision Tree classifier ensemble, 4) SMOTE with Random Forest ensemble. For all approaches, including the proposed approach, the algorithm used for inducing decision tree is CART.

Based on the analysis in Section 3, the evaluation metrics used for performance comparison and analysis of the proposed approach are G-Mean and F-Measure as they are best suited for evaluation of imbalanced data classification. The results for evaluation of the said classifiers using G-Mean and F-Measure are presented in Figures 2 and 3 respectively.

ISSN 1533-9211

DOI 10.5281/zenodo.7388773

Figure 2: Performance Comparison of Proposed GP-MOMS Approach with Baseline Methods using G-Mean

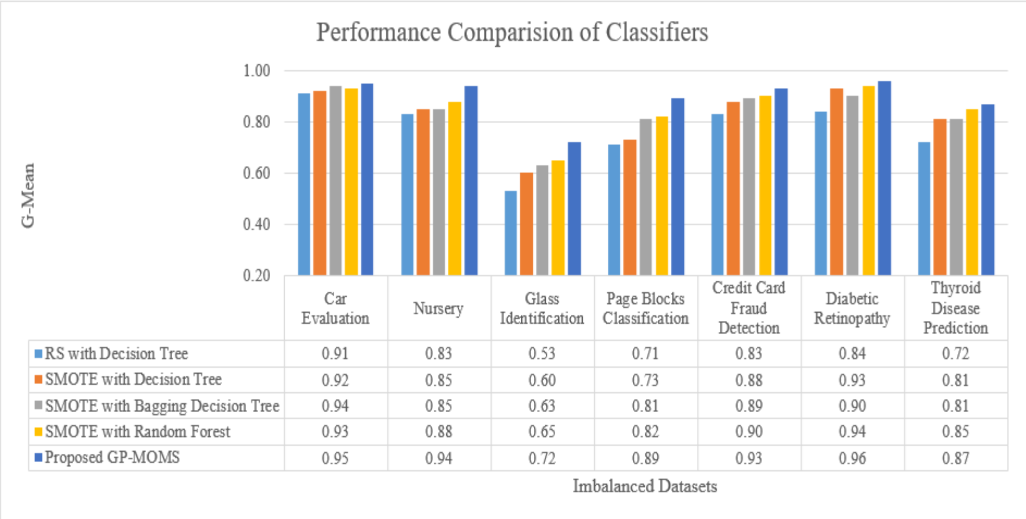
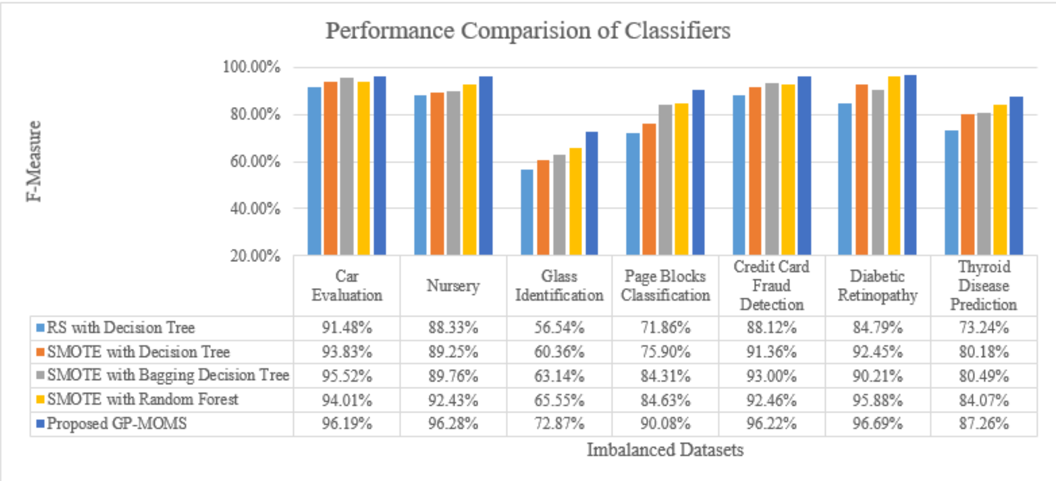


Figure 3: Performance Comparison of Proposed GP-MOMS Approach with Baseline Methods using F-Measure



Experiments on the above datasets confirm that the proposed approach has been successful in increasing the G-Mean as well as F-Measure of classifier on all the datasets. While Random Sampling with Decision Tree is a simpler approach, the efficiency on minority classes increases when Decision Tree is coupled with SMOTE. As SMOTE synthetically generates the targeted minority class instances, the classifier gets a larger base for learning and an increased performance is seen over Random Sampling. Hence, SMOTE is preferred over RS for all further implementations and comparisons. Further, it is observed that the performance of single Decision Tree classifier is low on datasets with more number of classes. Instead, an ensemble of Decision Tree classifiers, developed through the Bagging approach is improving the

The Seybold REPORT
ISSN 1533-9211
DOI 10.5281/zenodo.7388773

performance, both using G-Mean and F-Measure. Furthermore, when the numbers of attributes increase, the Random Forest ensemble performs better — as evident from the classification results for Diabetic Retinopathy dataset as well as Thyroid Disease Prediction dataset.

It can be analysed from Figure 2 and 3 that the proposed GP-MOMS approach outperforms all the baseline methods used for comparison of multi-class imbalanced data classification. These results prove that the MOMS approach of oversampling minority class instances and randomly sampling majority class instances is promising. When hybridized with GP, the classification results improve significantly, as the fitness function prefers individuals with higher G-Mean and F-Measure. As G-Mean and F-Measure are higher, it is evident that the model has precisely learned the classification pattern for minority class instances. These classification results using the proposed GP-MOMS approach have been efficient than the traditional baseline methods used for comparison on all the considered imbalanced datasets, which justifies the contributions of the proposed work.

6. CONCLUSIONS AND FUTURE WORK

The paper focuses the issue of learning classification patterns for instances of minority classes, which is a requirement for several real-world problems. The paper explores the traditional evaluation measures for classifiers used often for bi-class datasets, and re-designs them for multi-class applications. As compared to the other measures, G-Mean and F-Measure are analysed to produce fair evaluation of a classifier performance on minority classes, along with majority classes. Based on these outcomes and in-depth literature review of data level approaches and algorithm level approaches, a novel GP-MOMS based hybrid approach is proposed. Herein, the population for GP is initialized by oversampling minority class instances and randomly sampling the instances of remaining classes, which provides larger learning pool for the classifier to predict the minority class instances correctly. Further, G-Mean and F-Measure are used as fitness functions for evolving the GP population in every generation. As these measures promote classifiers that are unbiased and inclined towards minority as well as other classes, the final classifier evolved through GP is optimal. Thus, the works successfully address the problem of misclassification of minority class instances from an imbalanced dataset. However, with the fast changing world and continuous changes in trends, the underlying mapping of input data to the target class attribute may change. Efficient classification with such concept drift requires analysis and re-designing of classifiers, which is a subject for future work.

REFERENCES

[1] B. Krawczyk, “Learning from imbalanced data: open challenges and future directions,” Progress in Artificial Intelligence, vol. 5, no. 4. 2016. doi: 10.1007/s13748-016-0094-0.

[2] Z. Yang, W. H. Tang, A. Shintemirov, and Q. H. Wu, “Association rule mining-based dissolved gas analysis for fault diagnosis of power transformers,” IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, vol. 39, no. 6, 2009, doi: 10.1109/TSMCC.2009.2021989.

ISSN 1533-9211

DOI 10.5281/zenodo.7388773

[3] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, "Toward credible evaluation of anomaly-based intrusion-detection methods," IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, vol. 40, no. 5, 2010, doi: 10.1109/TSMCC.2010.2048428.

[4] A. Al-Shahib, R. Breitling, and D. Gilbert, "Feature selection and the class imbalance problem in predicting protein function from sequence," Appl Bioinformatics, vol. 4, no. 3, 2005, doi: 10.2165/00822942-200504030-00004.

[5] A. Azaria, A. Richardson, S. Kraus, and V. S. Subrahmanian, "Behavioral analysis of insider threat: A survey and bootstrapped prediction in imbalanced data," IEEE Transactions on Computational Social Systems, vol. 1, no. 2. 2014. doi: 10.1109/TCSS.2014.2377811.

[6] X. Gao, Z. Chen, S. Tang, Y. Zhang, and J. Li, "Adaptive weighted imbalance learning with application to abnormal activity recognition," Neuro computing, vol. 173, 2016, doi: 10.1016/j.neucom.2015.09.064.

[7] B. Krawczyk, M. Galar, Ł. Jeleń, and F. Herrera, "Evolutionary under sampling boosting for imbalanced classification of breast cancer malignancy," Applied Soft Computing Journal, vol. 38, 2016, doi: 10.1016/j.asoc.2015.08.060.

[8] M. J. Siers and M. Z. Islam, "Software defect prediction using a cost sensitive decision forest and voting, and a potential solution to the class imbalance problem," Inf Syst, vol. 51, 2015, doi: 10.1016/j.is.2015.02.006.

[9] Wang Shuo and Yao Xin, "Multiclass Imbalance Problems: Analysis and Potential Solutions," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 4, pp. 1119–1130, 2012.

[10] T. R. Hoens, Q. Qian, N. v. Chawla, and Z. H. Zhou, "Building decision trees for the multi-class imbalance problem," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2012, vol. 7301 LNAI, no. PART 1. doi: 10.1007/978-3-642-30217-6_11.

[11] L. Abdi and S. Hashemi, "To Combat Multi-Class Imbalanced Problems by Means of Over-Sampling Techniques," in IEEE Transactions on Knowledge and Data Engineering, 2016, vol. 28, no. 1. Doi: 10.1109/TKDE.2015.2458858.

[12] 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, 2002, doi: 10.1613/jair.953.

[13] S. Hu, Y. Liang, L. Ma, and Y. He, "MSMOTE: Improving classification performance when training data is imbalanced," in 2nd International Workshop on Computer Science and Engineering, WCSE 2009, 2009, vol. 2. doi: 10.1109/WCSE.2009.756.

[14] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in Lecture Notes in Computer Science, 2005, vol. 3644, no. PART I. doi: 10.1007/11538059_91.

[15] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in Proceedings of the International Joint Conference on Neural Networks, 2008. doi: 10.1109/IJCNN.2008.4633969.

[16] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2009, vol. 5476 LNAI. doi: 10.1007/978-3-642-01307-2_43.

[17] R. Xu, T. Chen, Y. Xia, Q. Lu, B. Liu, and X. Wang, "Word Embedding Composition for Data Imbalances in Sentiment and Emotion Classification," Cognit Comput, vol. 7, no. 2, 2015, doi: 10.1007/s12559-015-9319-y.

ISSN 1533-9211

DOI 10.5281/zenodo.7388773

[18] F. Charte, A. J. Rivera, M. J. Del Jesus, and F. Herrera, “MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation,” Knowl Based Syst, vol. 89, 2015, doi: 10.1016/j.knosys.2015.07.019.

[19] Y. Tang, Y. Q. Zhang, and N. v. Chawla, “SVMs modeling for highly imbalanced classification,” IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 39, no. 1, 2009, doi: 10.1109/TSMCB.2008.2002909.

[20] A. Anand, G. Pugalenthi, G. B. Fogel, and P. N. Suganthan, “An approach for classification of highly imbalanced data using weighting and under sampling,” Amino Acids, vol. 39, no. 5, 2010, doi: 10.1007/s00726-010-0595-2.

[21] E. Ramentol, Y. Caballero, R. Bello, and F. Herrera, “SMOTE-RSB*: A hybrid preprocessing approach based on oversampling and under sampling for high imbalanced data-sets using SMOTE and rough sets theory,” Knowl Inf Syst, vol. 33, no. 2, 2012, doi: 10.1007/s10115-011-0465-6.

[22] Q. Kang, X. S. Chen, S. S. Li, and M. C. Zhou, “A Noise-Filtered Under-Sampling Scheme for Imbalanced Classification,” IEEE Trans Cybern, vol. 47, no. 12, 2017, doi: 10.1109/TCYB.2016.2606104.

[23] J. Błaszczyński and J. Stefanowski, “Neighborhood sampling in bagging for imbalanced data,” Neuro computing, vol. 150, no. PB, 2015, doi: 10.1016/j.neucom.2014.07.064.

[24] B. Krawczyk, M. Woźniak, and G. Schaefer, “Cost-sensitive decision tree ensembles for effective imbalanced classification,” Applied Soft Computing Journal, vol. 14, no. PART C, 2014, doi: 10.1016/j.asoc.2013.08.014.

[25] S. Wang and X. Yao, “Diversity analysis on imbalanced data sets by using ensemble models,” in 2009 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009 - Proceedings, 2009. doi: 10.1109/CIDM.2009.4938667.

[26] L. Cen, “Classifying imbalanced data uses a Bagging Ensemble Variation (BEV),” in Proceedings of the Annual Southeast Conference, 2007, vol. 2007. doi: 10.1145/1233341.1233378.

[27] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, “A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches,” IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, vol. 42, no. 4. 2012. doi: 10.1109/TSMCC.2011.2161285.

[28] B. Krawczyk, “Learning from imbalanced data: open challenges and future directions,” Progress in Artificial Intelligence, vol. 5, no. 4. 2016. doi: 10.1007/s13748-016-0094-0.

[29] A. Fernández, S. del Río, N. v. Chawla, and F. Herrera, “An insight into imbalanced Big Data classification: outcomes and challenges,” Complex & Intelligent Systems, vol. 3, no. 2, 2017, doi: 10.1007/s40747-017-0037-9.

[30] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics,” Inf Sci (N Y), vol. 250, 2013, doi: 10.1016/j.ins.2013.07.007.

[31] H. Liu, M. Zhou, and Q. Liu, “An embedded feature selection method for imbalanced data classification,” IEEE/CAA Journal of Automatic a Sinica, vol. 6, no. 3, 2019, doi: 10.1109/JAS.2019.1911447.

[32] C. Wang, Z. Yu, H. Zheng, N. Wang, and B. Zheng, “CGAN-plankton: Towards large-scale imbalanced class generation and fine-grained classification,” in Proceedings - International Conference on Image Processing, ICIP, 2018, vol. 2017-September. doi: 10.1109/ICIP.2017.8296402.

[33] G. Douzas and F. Bacao, “Effective data generation for imbalanced learning using conditional generative adversarial networks,” Expert Syst Appl, vol. 91, 2018, doi: 10.1016/j.eswa.2017.09.030.

DOI 10.5281/zenodo.7388773

[34] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin, “Emotion classification with data augmentation using generative adversarial networks,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, vol. 10939 LNAI. doi: 10.1007/978-3-319-93040-4_28.

[35] I. J. Goodfellow et al., “Generative adversarial nets,” in Advances in Neural Information Processing Systems, 2014, vol. 3, no. January. doi: 10.3156/jsoft.29.5_177_2.

[36] Y. Zhang, “Deep Generative Model for Multi-Class Imbalanced Learning,” ProQuest Dissertations and Theses, 2018.

[37] H. S. Choi, D. Jung, S. Kim, and S. Yoon, “Imbalanced Data Classification via Cooperative Interaction between Classifier and Generator,” IEEE Trans Neural Netw Learn Syst, vol. 33, no. 8, 2022, doi: 10.1109/TNNLS.2021.3052243.

[38] J. R. Koza, “Genetic programming as a means for programming computers by natural selection,” Stat Comput, vol. 4, no. 2, 1994, doi: 10.1007/BF00175355.

[39] P. G. Espejo, S. Ventura, and F. Herrera, “A survey on the application of genetic programming to classification,” IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, vol. 40, no. 2. 2010. doi: 10.1109/TSMCC.2009.2033566.

[40] U. Bhowan, M. Johnston, M. Zhang, and X. Yao, “Evolving diverse ensembles using genetic programming for classification with unbalanced data,” IEEE Transactions on Evolutionary Computation, vol. 17, no. 3, 2013, doi: 10.1109/TEVC.2012.2199119.

[41] Y. Sun, M. S. Kamel, and Y. Wang, “Boosting for learning multiple classes with imbalances class distribution,” in Proceedings - IEEE International Conference on Data Mining, ICDM, 2006. doi: 10.1109/ICDM.2006.29.

[42] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. 2012. doi: 10.1016/C2009-0-61819-5.

[43] S. Garg and R. Kotecha, “Genetic Programming based Evolution of Classification Trees for Decision Support in Banking Sector,” Int J Knowl Eng Soft Data Paradig, vol. 5, no. 3–4, 2016, doi: 10.1504/ijkesdp.2016.10004137.

[44] D. Dua and C. Graff, “UCI Machine Learning Repository: Data Sets,” Irvine, CA: University of California, School of Information and Computer Science, 2019.

[45] Kaggle, “Kaggle: Your Home for Data Science,” Website, 2021.

[46] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, “Credit card fraud detection: A realistic modeling and a novel learning strategy,” IEEE Trans Neural Netw Learn Syst, vol. 29, no. 8, 2018, doi: 10.1109/TNNLS.2017.2736643.

[47] C. healthcare foundation Eye PACS, “Diabetic Retinopathy Detection | Kaggle dataset,” Kaggle. 2018.