Enhancing Classification Through Multi-view Synthesis in Multi-Population Ensemble Genetic Programming

Mohammad Sadegh Khorshidi
msadegh.Khorshidi.ak@gmail.com
PhD Candidate, Faculty of Engineering & IT, University of Technology Sydney
Sydney, NSW, AU

Navid Yazdanjue
Navid.Yazdanjue@gmail.com
PhD Candidate, Faculty of Engineering & IT, University of Technology Sydney
Sydney, NSW, AU

Hassan Gharoun
hassan.gharoun@student.uts.edu.au
PhD Candidate, Faculty of Engineering & IT, University of Technology Sydney
Sydney, NSW, AU

Danial Yazdani
danial.yazdani@gmail.com
Research Associate, Faculty of Engineering & IT, University of Technology Sydney
Sydney, NSW, AU

Mohammad Reza Nikoo
m.reza@squ.edu.om
Associate Professor, Department of Civil & Architectural Engineering, Sultan Qaboos University
Muscat, Muscat, Oman

Fang Chen
fang.chen@uts.edu.au
Distinguished Professor, Faculty of Engineering & IT, University of Technology Sydney
Sydney, NSW, AU

Amir H. Gandomi
a.h.gandomi@uts.edu.au
Professor, Faculty of Engineering & IT, University of Technology Sydney
Sydney, NSW, AU
Distinguished Professor, University Research and Innovation Center (EKIK), Óbuda University
Budapest, HU

ABSTRACT

This study proposes a genetic programming (GP) approach for classification, integrating cooperative co-evolution with multi-view synthesis. Addressing the challenges of high-dimensional data, we enhance GP by distributing features across multiple populations, each evolving concurrently and cooperatively. Akin to multi-view ensemble learning, the segmentation of the feature set improves classifier performance by processing disparate data "views". Individuals comprise multiple genes, with a SoftMax function synthesizing gene outputs. An ensemble method combines decisions across individuals from different populations, augmenting classification accuracy and robustness. Instead of exploring the entire search space, this ensemble approach divides the search space to multiple smaller subspaces that are easier to explore and ensures that each population specializes in different aspects of the problem space. Empirical tests on multiple datasets show that the classifier obtained from proposed approach outperforms the one obtained from a single-population GP executed for the entire feature set.

GECCO 24, July 14-18, 2024, Melbourne, Australia

ACM ISBN 979-8-4007-0494-9/24/07…$15.00
https://doi.org/10.1145/3638530.3664172

CCS CONCEPTS

• **Computing methodologies** → **Genetic programming**; *Ensemble methods.*

KEYWORDS

Pattern recognition and classification, Genetic programming, Multi-population models, Multi-view Ensemble Learning.

ACM Reference Format:
Mohammad Sadegh Khorshidi, Navid Yazdanjue, Hassan Gharoun, Danial Yazdani, Mohammad Reza Nikoo, Fang Chen, and Amir H. Gandomi. 2024. Enhancing Classification Through Multi-view Synthesis in Multi-Population Ensemble Genetic Programming. In *Proceedings of ACM Conference (GECCO 24).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3638530.3664172

1 INTRODUCTION

The quest for efficient and interpretable models is driven by the need for informed decision-making in various domain knowledges [9]. Genetic Programming (GP) is an evolutionary learning method that seeks to address this challenge by evolving computer programs, or individuals, to solve complex problems without predefined structures. Tree-based GP's representation system, typically structured as expressions comprising functional and terminal nodes, facilitates the exploration of complex relationships between target and input variables. This exploration, however, becomes increasingly challenging as the dimensionality of the dataset expands, demanding

a more extensive search through potential combinations of nodes [3].

To enhance GP's efficiency and mitigate computational demands, the development of Multi-tree Genetic Programming (MTGP) represents a significant advancement. The MTGP evolves multiple genes within a single individual, allowing for the construction of more complex models through the aggregation of multiple independent programs or trees [8]. Despite MTGP's advancements, we face the persistent challenge of dimensionality in searching for interpretable models, common across machine learning algorithms, which can hinder efficient solution space exploration and elevate computational costs. Addressing this issue, dimensionality reduction techniques, including feature selection and extraction, have been explored, though they risk information loss and may oversimplify complex feature dynamics [4, 6].

Multi-view learning (MVL) and its extension, Multi-View Ensemble Learning (MVEL), present innovative strategies to leverage data from diverse origins or perspectives, enhancing model performance through techniques like bagging, boosting, or stacking. These approaches, utilizing either natural views derived directly from varied data sources or artificial views generated through data transformation, aim to mitigate over-fitting risks and improve model adaptability in handling complex, high-dimensional datasets.

In the domain of evolutionary machine-learning the ensemble learning approaches have been explored extensively [2]. However, the novel MVEL method is yet to be investigated. In this study, we propose the Multi-Population Ensemble GP (MPEGP) which serves as a framework based on the concept of tree-based GP. The MPEGP incorporates the benefits of the MVEL in search for an interpretable classifier. The contribution of the present study in the domain of knowledge can be summarized as follows:

- Utilizing a multi-population approach to emulate multi-view learning, enabling independent evolution of populations across diverse feature subsets representing distinctive "views" of a dataset.
- Using simple SoftMax function to preserve simplicity and interpretability of obtained ensemble models.
- Using an adaptive gradient descent fine-tuning strategy to adaptively learn and adjust genes' contributions in predictive ability of individuals in response to the evolving populations.
- Integration of multiple views through ensembling synthesized gene outputs that enhances classification accuracy and model robustness.
- Selecting elite individuals based on fitness both in isolation and in conjunction with individuals from other populations to preserve diversity and create cooperation between populations.

The rest of the paper is organized as follows: section 2 presents the proposed method, section 3 describes the benchmarking and experimental setup to evaluate the proposed method's efficiency, section 4 discusses the obtained results, and section 5 concludes the paper.

2 METHODOLOGY

2.1 Multi-Population Genetic Programming

As stated before, in learning tasks, an increase in the number of features results in exponential growth of the search space—an effect commonly referred to as the "curse of dimensionality". One principal strategy to overcome this issue is MVEL, which segments the dataset vertically, dividing it into multiple subsets of features that provide unique perspectives or "view" of the dataset. Each subset is then used to train distinct classifiers. This technique effectively partitions the original expansive search space into several smaller, more manageable subspaces, simplifying the learning task and reducing the problem's complexity. Subsequently, the classifiers' output class label probabilities are synthesized using an ensemble method to achieve a complete picture of the target class label.

In this paper, we propose the Multi-Population Ensemble Genetic Programming (MPEGP) that incorporates the power of MVEL within tree-based GP. While this approach allows for division of the search space and enhance the ability of GP to handle large datasets without increasing the risk of over-fitting, it preserves the interpretability of obtained ensembles. It is noteworthy that current framework is designed for classification tasks; nevertheless, it can be adapted to address regression and feature learning tasks. Figure 1 provides an overview of MPEGP. As illustrated, the features characterizing an object (for instance, a cylinder) are divided into multiple views capturing different angles (in this case, two views from the top and side, symbolized by a circle and rectangle, respectively). These views could be artificially constructed from a single dataset, or obtained from multiple sources. Moreover, in Fig. 1 only two views are visualized; however, the MPEGP framework is capable of handling more than two views.

Subsequently, the subsets of features representing views are assigned to two separate population (i.e. $Population_1$ and $Population_2$) of individuals. Each individual within the populations represents a candidate solution, encoded as a set with variable number of genes. The genes correspond to the arithmetic combination (expression tree) of features subsets (views) that are evolved by genetic operators within generations. The maximum number of genes per individual is predefined, with individuals initially generated randomly by one of the tree initialization methods. The predictive classifier for each individual consists of a combination of its gene outputs. These outputs are aggregated with a SoftMax function ($SoftMax_1$ and $Softmax_2$), fine-tuned using an adaptive gradient descent algorithm, to yield a probability distribution (pdf) over potential class labels ($P(Y)$). The fitness of each individual, $Fitness_1$ and $Fitness_2$, is calculated based on the closeness of this probability distribution to the true class labels. Here, we use cross-entropy as the measure of fitness for the individuals and the ensemble.

Next, an ensemble SoftMax function, $SoftMax_{en}$ combines the predicted probabilities obtained from the pairs of the first tier classification models and undergoes a similar fine-tuning process. To control the number of ensembles, the predicted probabilities from an individual $Individual_{1,j}$ only pairs with the corresponding individual $Individual_{2,j}$. This ensemble outputs the ensemble pdf of class labels. The fitness of the ensemble ($Fitness_{en}$) can be calculated from the resulting ensemble pdf to assess the ensemble predictive performance. Following fitness evaluation, individuals within each

population are sorted based on their fitness scores. Selection for reproduction is conducted separately within each population based on $Fitness_1$ and $Fitness_2$. Standard genetic operations—such as crossover and mutation—are performed to generate new offspring, ensuring diversity and the propagation of high-fitness genetic material within each population.

An elite selection strategy is employed to ensure that the top-performing individuals, according to a user-defined fraction, are passed directly to the next generation. This selection is based on both individual fitness scores within each population and the ensemble fitness score. This dual-criterion approach allows for the preservation of individuals that excel in their predictive capacity both in isolation and in conjunction with the other population. The introduction of ensemble fitness into the elite selection process facilitates a cooperative co-evolution between the populations. Individuals that may not exhibit optimal fitness within their respective populations but contribute significantly to the ensemble's predictive capability are retained. This mechanism promotes genetic diversity and encourages a broader exploration of the solution space. The adaptive gradient descent algorithm is critical to the fine-tuning of gene weights, effectively determining the contribution of each gene to the prediction outcome and ensures dynamic adjustment of weights in response to the evolving populations.

The evolutionary process continues over multiple generations until a convergence criterion is met. Convergence is assessed based on the stability of fitness scores across generations (stall generation), or a maximum number of generations is reached. The final model is selected based on the highest ensemble fitness score (lowest ensemble cross-entropy), representing the optimal combination of predictive capacity from both populations. This structured approach enables the MPEGP to handle high-dimensional data without increasing the risk of over-fitting, simplifies the learning task and enhances predictive accuracy and robustness.

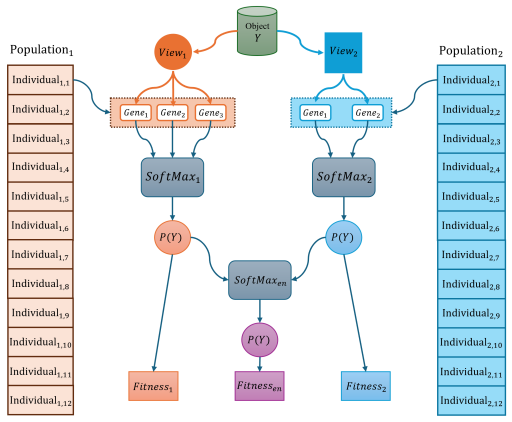

Figure 1: Overview of Multi-Population Ensemble Genetic Programming.

3 EXPERIMENTAL DESIGN

3.1 Data Description

The evaluation of the (MPEGP) method is conducted using four publicly available datasets, including Activity Recognition Using Wearable Physiological Measurements (ARWPM) [1], Gene Expression Cancer RNA-Seq Data Set (GECR) [5], Gas Sensor Array Drift Dataset at Different Concentrations (GSAD) [10], and Smartphone-Based Recognition of Human Activities and Postural Transitions Data Set (HAPT) [7]. Table 1 provides the characteristics of the selected datasets. These datasets were selected to include both ends of the spectrum of the number of features, classes and instances. This selection of datasets can justify the consistency of obtained results for datasets with different characteristics.

Table 1: Datasets' description.

Dataset	# Instances	# Features	# Classes
ARWPM	4, 480	533	4
GECR	801	20, 531	5
GSAD	13, 910	129	6
HAPT	10, 929	561	12

3.2 Experimental Setup

To evaluate the MPEGP performance against a benchmark model, we consider the benchmark model to be the single-population of MPEGP. In other words, all features within the datasets are assigned to the first population of individuals in MPEGP and the number of populations in MPEGP is set to 1. Moreover, the ensembling module of MPEGP that combines the output class label pdfs of multiple populations is disabled. Hereby, this version of MPEGP is called "baseline GP". On the other hand, two distinct views of the four datasets are generated at the beginning of each run using the training data via the SPFP algorithm proposed in [6]. In contrast to the baseline GP, the features of views are assigned to two separate population of individuals for MPEGP. Thus, the results obtained from MPEGP is compared against the baseline GP. The individuals in baseline GP can contain all features within a dataset, while the individuals in MPEGP can contain only a subset of the features. This allows for a fair assessment of MPEGP performance. For the experiments, we performed 20 independent runs of MPEGP and the baseline GP. For each of the 20 independent runs, the model evolves over a maximum of 150 generations.

Individuals in MPEGP and baseline GP are limited to 5 and 10 maximum genes, respectively, to balance computational feasibility and solution complexity. The population comprises 25 individuals, with the maximum depth of each tree capped at 10. Initial tree structures are generated using the 'Half-and-Half' approach for both models. The models adopt a tournament selection method with a size of two. An elite fraction of 0.05 is also considered for both models to preserve top-performing solutions. For genetic operations, the functional terminals include basic arithmetic functions: addition (+), subtraction (-), multiplication (×), and division (/). The crossover and mutation probabilities are set at 0.8 and 0.1, respectively. The inclusion probability for constant terminals in trees is set at 0.1, adding more versatility to the evolved expressions.

4 RESULTS AND DISCUSSION

This comparative study was structured around 20 independent evaluations. In each evaluation, the dataset underwent a random division, allocating 60% for training, 10% for validation, and the remaining 30% for testing purposes. Throughout these evaluations, a key metric of interest was the best fitness achieved by the ensembles (referred to as ensemble fitness) alongside the best fitness scores within each of the two populations (designated as the fitness of population 1 and 2, respectively). These fitness metrics were stored at each generation within the evaluations. Figure S1 visually presents the temporal evolution of these fitness metrics for a randomly selected run, color-coded as blue for ensemble fitness, red for fitness of population one, and orange for fitness of population two. For more comprehensive visualization of all runs, please refer to the Supplementary Document.

An important consideration in this analysis is the non-correspondence of the best fitness values across the populations and the ensemble. Theoretically, it might be expected that in certain generations, the fitness of one population could momentarily exceed that of the ensemble, particularly if the corresponding individual from the other population exhibits suboptimal performance. Nonetheless, this anticipated occurrence was not witnessed during any of the runs across all datasets, indicating the fitness of an ensemble should always be equal or better that its constituents.

Further insight into the dynamics of individual and ensemble performance is provided by Figure S2, which explores the fitness ranking of individuals constituting the fittest ensemble within their populations for a run randomly selected from the 20 trials. Notably, the individuals contributing to the fittest ensemble were often not the top performers within their respective populations. Moreover, their simultaneous attainment of the top rank within their populations was exceedingly rare. This observation underscores a critical facet of the MPEGP's design, which effectively promotes the combination of individuals that form the most potent ensemble, thereby preserving genetic diversity and enhancing the system's exploratory capabilities beyond that of single-population models.

To evaluate the performance differences between MPEGP and baseline GP, we applied the Wilcoxon rank-sum test, aiming to determine if the outcomes during the training, validation, and testing phases significantly diverge at $\alpha = 0.05$ significance level. This non-parametric statistical test compares two independent samples to ascertain if there is a significant difference in their population mean ranks. The null hypothesis posits no difference between the samples. A p-value less than 0.05 indicates rejection of the null hypothesis, with the sample exhibiting the lower mean deemed superior. Table 2 compiles the mean, standard deviation, and Wilcoxon p-values for these stages across the AWRPM, GECR, GSAD, and HAPT datasets.

The summarized results in the table, demonstrate the superiority of the MPEGP-derived models over the baseline GP across all metrics, with the exception of the training and validation phases for the GECR dataset. The limited size of the validation dataset (80 instances) for GECR, combined with the baseline GP's susceptibility to over-fitting, led to significant variability in the cross-entropy and statistically insignificant difference between the validation performance of MPEGP comaperd to baseline GP models on GECR dataset. From the testing phase results of the GECR dataset for the baseline GP, it is evident that the obtained classifier is prone to over-fitting.

Table 2: Mean and standard deviation of cross-entropy for training, validation, and testing phases across 20 runs of MPEGP and baseline GP on ARWPM, GECR, GSAD, and HAPT datasets. Included are the p-values from the Wilcoxon rank-sum test, with superior performance highlighted in bold.

		ARWPM	GECR	GSAD	HAPT
MPEGP	Train	**0.0723 ± 0.0104**	0.0516 ± 0.032	**0.0098 ± 0.003**	**0.0195 ± 0.0028**
	Validation	**0.1018 ± 0.0237**	0.0886 ± 0.0454	**0.0212 ± 0.0065**	**0.0269 ± 0.0037**
	Test	**0.1057 ± 0.0138**	**0.1613 ± 0.0837**	**0.026 ± 0.0063**	**0.0281 ± 0.0028**
Baseline GP	Train	0.1271 ± 0.0184	**0.0175 ± 0.0187**	0.0479 ± 0.0194	0.0344 ± 0.0097
	Validation	0.1342 ± 0.0168	0.1569 ± 0.142	0.0514 ± 0.0198	0.0344 ± 0.0095
	Test	0.1357 ± 0.0181	0.3316 ± 0.1829	0.054 ± 0.0198	0.0362 ± 0.0095
P-Value	Train	9.17×10^{-8}	2.1×10^{-4}	6.8×10^{-8}	1.23×10^{-7}
	Validation	4.68×10^{-5}	0.57	1.38×10^{-6}	1.2×10^{-4}
	Test	1.41×10^{-5}	5.1×10^{-4}	5.23×10^{-7}	6.22×10^{-4}

5 CONCLUSION

The method proposed in this paper synthesizes an individual's gene outputs and ensembles the predictions of individuals across multiple populations to achieve a robust and interpretable model of the target object. The results show that MPEGP outperforms single-population GP of the same variant. This approach allows for cooperation and co-evolution among individuals of different populations by conserving individuals that create high performing ensembles. While there is room for investigation of more aspects of proposed MPEGP, the future works can focus on more sophisticated variants of GP (e.g. strongly typed GP) for image and text datasets.

ACKNOWLEDGEMENT

This work was supported by the Australian Government through the Australian Research Council under Project DE210101808.

REFERENCES

[1] 2019. Activity recognition using wearable physiological measurements. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5RK6V.
[2] Ying Bi, Bing Xue, and Mengjie Zhang. 2020. Genetic programming with a new representation to automatically learn features and evolve ensembles for image classification. *IEEE transactions on cybernetics* 51, 4 (2020), 1769–1783.
[3] Qi Chen, Mengjie Zhang, and Bing Xue. 2017. Feature selection to improve generalization of genetic programming for high-dimensional symbolic regression. *IEEE Transactions on Evolutionary Computation* 21, 5 (2017), 792–806.
[4] Qinglan Fan, Ying Bi, Bing Xue, and Mengjie Zhang. 2024. Multi-Tree Genetic Programming for Learning Color and Multi-Scale Features in Image Classification. *IEEE Transactions on Evolutionary Computation* (2024).
[5] Samuele Fiorini. 2016. gene expression cancer RNA-Seq. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5R88H.
[6] Mohammad Sadegh Khorshidi, Navid Yazdanjue, Hassan Gharoun, Danial Yazdani, Mohammad Reza Nikoo, Fang Chen, and Amir H Gandomi. 2024. Semantic-Preserving Feature Partitioning for Multi-View Ensemble Learning. *arXiv e-prints* (2024), arXiv–2401.
[7] Anguita Davide Oneto Luca Reyes-Ortiz, Jorge and Xavier Parra. 2015. Smartphone-Based Recognition of Human Activities and Postural Transitions. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C54G7M.
[8] Dominic P Searson, David E Leahy, and Mark J Willis. 2010. GPTIPS: an open source genetic programming toolbox for multigene symbolic regression. In *Proceedings of the International multiconference of engineers and computer scientists*, Vol. 1. Citeseer, 77–80.
[9] Chaonan Shen and Kai Zhang. 2022. Two-stage improved Grey Wolf optimization algorithm for feature selection on high-dimensional classification. *Complex & Intelligent Systems* (2022), 1–21.
[10] Alexander Vergara. 2013. Gas Sensor Array Drift Dataset at Different Concentrations. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5MK6M.