This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/CEC65147.2025.11042993, 2025 IEEE Congress on Evolutionary Computation (CEC)

A Multi-Objective Genetic Programming with Size Diversity for Symbolic Regression Problem

1st Yujie Zhang
School of Communication and Information Engineering
Chongqing University of Post and Telecommunications
Chongqing,China
0009-0004-9381-0279

2nd Guoquan Li*
School of Communication and Information Engineering
Chongqing University of Posts and Telecommunications
Chongqing,China
0000-0001-8022-743X

3rd Zhengwen Huang
Department of Electronic and Electrical Engineering
Brunel University London
London,U.K
0000-0003-2426-242X

4th Jinshuo Jia
School of Communication and Information Engineering
Chongqing University of Post and Telecommunications
Chongqing,China
S220101055@stu.cqupt.edu.cn

5th Xiang Li
School of Communication and Information Engineering
Chongqing University of Post and Telecommunications
Chongqing,China
S230101088@stu.cqupt.edu.cn

6th Donghui Peng
School of Communication and Information Engineering
Chongqing University of Post and Telecommunications
Chongqing,China
S230132102@stu.cqupt.edu.cn

***Abstract*—Genetic programming has been positioned as a fit-for-purpose approach for symbolic regression. Researchers tend to select algorithms that produce a model with low complexity and high accuracy. Multi-objective genetic programming (MOGP) is a promising approach for finding appropriate models by considering tradeoffs between accuracy and complexity. The MOGP has gained significant attention for non-dominated sorting genetic algorithm II (NSGA-II). However, NSGA-II tends to excessively select individuals of lower complexity, making NSGA-II inefficient in real world applications. SD can be a strategy to promote the evolutionary process by adapting selection pressures for individuals of various size. It deals with the excessive tendency to select low complexity individuals in NSGA-II.We also introduce a practical industrial case of defect detection for dispensing machines. By modeling the dispensing volume of the fluid dispensing systems, defects in the dispensing machine can be detected under different external environmental factors.For the validation of SD, other MOGP algorithms are compared with the improved NSGA-II algorithm, NSGA-II with SD. By comparing multi-objective optimization methods tested on seven general datasets and an industrial case about defect prediction, the experimental results show that performance of the proposed approach is superior or same to other models in terms of accuracy. In terms of complexity, performance of the proposed approach is satisfactory.**

***Index Terms*—Genetic Programming, symbolic regression, Multi-Objective, Non-dominated Sorting, fluid dispensing systems.**

I. Introduction

GENETIC Programming [1] is an effective approach for Symbolic Regression. It can produce some solutions whose fitness can be accuracy. However, the GP tends to generate models with redundancy. This phenomenon is known as bloat, making solutions uninterpretable [2], [3].

In order to deal with the bloat problem, multi-objective GP (MOGP) is universally utilized to eliminate the bloat [1], [4].Furthermore , MOGP , whose objectives are accuracy (MSE) and complexity (size) , can produce a set of non-dominated solutions with trade-off between accuracy and complexity.The second version of the non-dominated sorting genetic algorithm(NSGA-II) [5] is a widely applied kind of MOGP framework.As some researches show, NSGA-II, where the first and second objective are MSE and individual's size separately, is extremely inefficient. The reason is the over-replication of low complexity models [6]. The complex and accurate models cannot be found. This results in the evolutionary process getting stuck within the local optimum. A similar concept is called *objective selection bias issue* [7].

We propose an algorithm called size diversity (SD) to deal with the *the population collapse* [8], [9] problem in NSGA-II. Specifically, the second objective (size) of MOGP is replaced by SD. MOGP with SD makes the entire population occupied with individuals of different sizes. The proposed algorithm adjusts the selection pressure of MOGP on individuals of different sizes, so that MOGP does not excessively tend to select low complexity individuals.

Eventually, we propose a practicable workflow of NSGA-II with SD. In addition, the algorithm proposed in this paper is

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/CEC65147.2025.11042993, 2025 IEEE Congress on Evolutionary Computation (CEC)

used to solve the actual engineering case, the defect prediction of dispensing machine. In addition, the proposed algorithm is compared with other evolutionary algorithms in terms of the complexity and accuracy of the output model.

II. Related Work

MOGP effectively addresses the bloat problem by setting accuracy and complexity as two objectives [10]. Accuracy is measured by MSE or RMSE, while complexity can be defined by tree depth [11], number of nodes [12], or semantic measures [13]. By considering complexity during selection process, MOGP can reduce model bloat. In this paper, we define complexity as the number of nodes (i.e., GP tree size).

A. NSGA-II

NSGA-II [5] is a widely used multi-objective evolutionary algorithm (MOEA). Therefore, the MOGP algorithm we mainly research is NSGA-II.The framework of NSGA-II is shown in the figure 1.

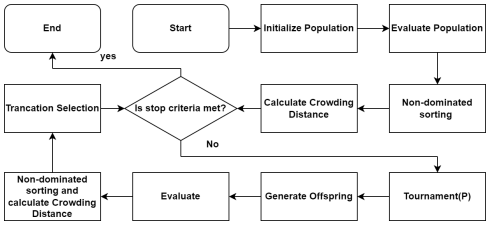

Fig. 1. Workflow of NSGA-II

However, some studies have shown a problem that the entire population converges to copies of single solution in NSGA-II with the objectives,accuracy and size. This phenomenon is called population collapse [8], [9] or evolvability degeneration [6].In addition, a concept similar to population collapse is mentioned, which refers to a phenomenon that ineffective (usually small) solutions instead of effective ones(usually large) are more likely to be selected. This is called *the objective selection bias issue* [7].

III. Size Diversity

Size Diversity (SD) is a strategy to advance the evolutionary process by adapting selection pressures for individuals of various sizes. The goal of adopting SD is to address the issue of over-replication of small individuals in NSGA-II. The reason given rise to this issue is the over-replication of small individuals in NSGA-II [6]. Furthermore,modification to selection pressure is a way to solve evolutionary stagnant or popluation collpase. SD determine selection pressure depending on proportion of size in the population.

A. Implementation of Size Diversity

Implementation of SD is obtained by counting the number of individuals of the size and the number of individuals of a nearby size (determined by l).The calculated SD is assigned to all individuals of that size as a second objective (fitness).As the figure 2 and eqution 1 shows, the blue block is the size to be calculated, and the red colour is the individuals in the vicinity of the blue. If l = 1 and the centre size=24, only the 1 block(red block) around the centre size is considered. l is a hyperparameter.

$$size\_diversity = \frac{N(23) + N(24) + N(25)}{2l + 1} \tag{1}$$

Where $2l + 1 = 3$. $N(23)$ is the number of individuals of size 23.The aim of SD is to control the selection pressure depending on proportion of size in the population. If the sum of number of centre size and the vicinity is too large, it gives rise to a decline in the selection pressure on all individuals of this size, which can overcome an issue that population convergence to copies of small size individuals. The detailed calculation method of SD is shown as follows.

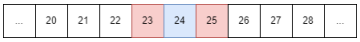

Fig. 2. SD Calculation for l=1,s=24

As the expression 2 and 3 shows,

$$size\_diversity(s, l) = \frac{\sum_{i=-l}^{l} N(s + i)}{A} \tag{2}$$

Where l is a hyperparameter and $N(s)$ means the number of individuals of size s in the population. A is number of considered sizes ,including centre size and the vicinity. For example, $l = 1, s = 25$, the SD for individuals of size 25 is $\frac{N(25)+N(24)+N(26)}{3}$, $A = 2l + 1$.

$$size\_diversity(s, l) = \frac{\sum_{i=-l}^{l} N(s + i)}{2l + 1} \tag{3}$$

B. Implementation of MOGP with Size Diversity

The figure 3 shows a practical workflow of NSGA-II with SD, after evaluating all the offspring, the second objective (size) for all individuals are modified to SD. Taking NSGA-II as an example of a multi-objective algorithm framework, an adjustment module (SD module) is added to the original multi-objective algorithm framework as is shown in figure 3. MOGP with SD utilizes proposed SD in workflow of NSGA-II.Complex and accurate models can be maintained due to the use of SD algorithm, which deal with evolutionary stagnation.In this way, the models produced by MOGP with SD achieves the performance of low complexity and high accuracy.

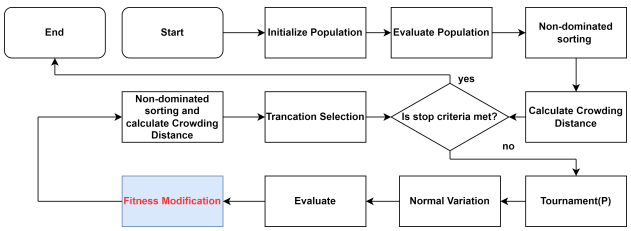

Fig. 3. The workflow of MOGP with Size Diversity

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/CEC65147.2025.11042993, 2025 IEEE Congress on Evolutionary Computation (CEC)

TABLE I
EXPERIMENTAL SETUP FOR GENERALIZED DATASET TESTING

General Parameter	Values
Population size	500
Tournament size	2
Generation	1000
Crossover-mutation proportion	0.9-0.1
Initialization	Ramped half-&-half (2–6)
Maximum individual size	100
Function set	$\{+, -, \times, \div^*, a^2\}$
Terminal set	$\{Constant \cup Features\}$
l	2

C. Experimental Setup and Performance Evaluation

In this paper, symbolic regression datasets includes Airfoil, Boston, Concrete, Dow Chemical, Energy (cooling/heating), and Yacht. Each dataset is randomly split into 80% training and 20% testing, and normalized to zero mean and unit variance.We record the most accurate individuals in the last generation of the population and their size. In addition, NSGA-II, NSGA-II with adaptive alpha dominance (α-dom.) [7], GP and SPEA-II are considered for comparison.

The specific experimental parameters for each algorithm are shown in Table I, where the random constants are in the range of (-5,5). $\div^*$ means that the division is a protected division, i.e. the function returns 1 if the denominator is 0. MSE of the most accurate individual was averaged for evaluating performance of these algorithms (for 30 runs).

1) Performance of Population Distribution: To verify the effectiveness of the SD strategy for solving objective selection bias issue [7], we run 30 experiments on these datasets to evaluate the performance of NSGA-II with SD and conventional NSGA-II.The figure 4, 5 reveals the distributions of size in population of NSGA-II with SD and original NSGA-II. In population of NSGA-II,the distributions show that there exists population collapse problem. However,in population of NSGA-II with SD, there is no population collapse or population convergence to copies of several specific individuals,which means that the population collapse is solved by SD algorithm.

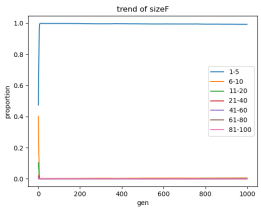

Fig. 4. Distribution of NSGA-II

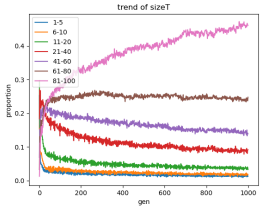

Fig. 5. Distribution of NSGA-II with SD

2) Performance of Accuracy and Complexity: We utilize the average mean of MSE as the performance (accuracy) measure. If some solution produced by some run is great different from the one from other runs, we should remove these outliers. Results where MSE is larger than 1 are treated as outliers. The table II reveals accuracy(MSE) performance of these algorithms, including NSGA-II with SD, Adaptive α-dominance, NSGA-II , and SPEA-II and Standard (Single Objective) GP. The table III conveys mean of complexity(size).

TABLE II
ACCURACY FOR TEST SET

Dataset	SD	α-dom.	NSGA-II	SPEA-II	GP
Airfoil	**0.280**	0.301	0.574	0.430	0.363
Boston	**0.225**	0.252	0.451	0.321	0.234
Concrete	**0.206**	0.225	0.508	0.389	0.272
Dow chemical	**0.449**	**0.445**	0.871	0.794	0.458
Energy:Cooling	**0.106**	0.176	0.175	0.143	0.244
Energy:Heating	**0.064**	0.153	0.194	0.129	0.228
Yacht	**0.052**	0.0814	0.347	0.216	0.107

TABLE III
COMPLEXITY FOR TEST SET

Dataset	SD	α-dom.	NSGA-II	SPEA-II	GP
Airfoil	90.3	97.5	8.7	20.0	98.2
Boston	85.1	91.4	5.0	13.1	82.2
Concrete	92.5	97.2	8.3	21.3	96.7
Dow chemical	73.3	85.0	3.1	5.8	95.8
Energy: cooling	61.7	88.8	7.2	13.4	90.3
Energy: heating	61.6	96.6	6.0	13.9	88.4
Yacht	75.8	94.5	6.0	9.9	94.7

As is shown in tables II and III, in terms of accuracy (MSE), the model generated by NSGA-II with SD is the best model, except for the Dow Chemical dataset. In the Dow Chemical dataset, it is almost equal to the adaptive α-dominance model in accuracy. Except for the proposed algorithm NSGA-II with SD, performance of the adaptive α-dominance is best,compared to Standard GP, NSGA-II, and SPEA-II. Due to population collapse, the evolutionary process of NSGA-II almost stall at the end stage of evolution, which causes that performance of NSGA-II is worst.From the complexity (size) point of view, the proposed algorithm produces solutions of lower complexity than Adaptive α-dominance and Standard GP. Both NSGA-II and SPEA-II are able to produce solutions of lower complexity.For example, in terms of accuracy, MSE of the proposed algorithm and the adaptive α-dominance are seperately 0.052 and 0.0814 in Yacht dataset. In terms of complexity, size of the proposed algorithm is 75.8 better than that of the the adaptive α-dominance ,94.5. Therefore, in Yacht dataset, the proposed algorithm is better than the adaptive α-dominance in both of accuracy and complexity. Overall, the algorithm proposed in this paper, NSGA-II with SD, generates a model with the best accuracy, and its complexity is better than other algorithms but NSGA-II and SPEA-II.

IV. AN INDUSTRIAL CASE STUDY

Silicon chips are attached to circuit boards using conductive adhesive dispensed precisely by a system. The bonding head places the die into the adhesive to a controlled height. Two

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/CEC65147.2025.11042993, 2025 IEEE Congress on Evolutionary Computation (CEC)

common dispensing failures(insufficient or excessive glue) can damage components and lead to rejection. Poor control over glue shape and volume affects component functionality.

A study [14] identifies several factors influencing dispensing machine defects, including glue temperature and level, glue and air pressure, glue fillet threshold, room temperature, machine calibration decay, and needle age. These factors affect both the sensitivity of glue output and the positioning of the glue droplet center.

The proposed MOGP with SD algorithm is used to predict the glue volume in order to generate a low-complexity model while ensuring model accuracy. Then the model generated by the algorithm is used to predict whether samples in datasets are defects or not. In order to obtain reliable results, we run 10 experiments. The generation of algorithms is 100, and the rest of the experimental parameters are the same as Section III-C. Table IV shows the performance of evolutionary algorithms in an industrial case on predicting defects.

In terms of accuracy, our proposed NSGA-II algorithm with SD achieves 0 error detection which outperforms compared algorithms. In addition, the accuracy (MSE) of the symbolic regression model produced by our proposed algorithm outperforms the other compared algorithms. Adap. α-dom. performs better than NSGA-II and Standard GP.

In terms of the average runtime performance of the algorithm, the average runtime of the algorithm proposed is close to that of Adap. α-dom.

In terms of complexity (Size), the NSGA-II algorithm produces the lowest complexity model due to population collapse.

In Section III-C, we take the pareto front generated by the last generation of populations from the last experiment in the Airfoil dataset and compute the Hypervolume (HV) in order to judge the algorithm's ability to produce a high-quality Pareto front, where the reference point is set to (1.1, 110). NSGA-II with SD outperforms the comparison algorithm in this respect.

Overall, our proposed algorithm is a promising approach to provide low-complexity, high-accuracy symbolic regression models that have achieved satisfactory performance in real-world engineering cases.

TABLE IV
DEFECT PREDICTION ACCURACY AND MODEL COMPLEXITY

Algorithms	Errors	Size	Runtime	MSE	HV(Airfoil)
NSGA-II (SD)	**0**(+)	81.2	314.73s	**3.88e-06**	**745.40**
Adap.α-dom.	1(-)	77.7	289.36s	6.07e-06	7.54
NSGA-II	5(-)	**8.7**	**102.64s**	1.64e-05	29.80
GP	5(-)	78.9	226.41s	1.61e-05	0.00

V. CONCLUSION AND FUTURE WORKS

The major objective of this work is to find accurate and low complexity models for symbolic regression. We propose SD to deal with evolutionary stagnation caused by the inefficiency of NSGA-II. To apply SD strategy, we propose a algoirthm framework for MOGP with SD.In terms of accuracy,we found that the model of the proposed algorithm that performs the best. In terms of model complexity, the performance of the model was just inferior to NSGA-II and SPEA-II.Overall, the proposed algorithm is more promising. In the future, we will continue to explore SD, such as application of SD in Gene Expression Programming.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 12411530119 and U21A20447.

REFERENCES

[1] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press, 1992.

[2] S. Luke and L. Panait,"A comparison of bloat control methods for genetic programming,"*Evolutionary Computation*,vol. 14, no. 3, pp. 309–344, 2006.

[3] E. D. de Jong, R. A. Watson, and J. B. Pollack, "Reducing bloat and promoting diversity using multi-objective methods,"in *Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO'01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, p. 11–18.

[4] S. Bleuler, M. Brack, L. Thiele, and E. Zitzler, "Multiobjective genetic programming: reducing bloat using SPEA2," in *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546)*, vol. 1, 2001, pp. 536–543 vol. 1.

[5] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[6] D. Liu, M. Virgolin, T. Alderliesten, and P. A. N. Bosman, "Evolvability degeneration in multi-objective genetic programming for symbolic regression," in *Proceedings of the Genetic and Evolutionary Computation Conference*, ser. GECCO '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 973–981. [Online]. Available: https://doi.org/10.1145/3512290.3528787

[7] S. Wang, Y. Mei, and M. Zhang, "A multi-objective genetic programming algorithm with α dominance and archive for uncertain capacitated arc routing problem," *IEEE Transactions on Evolutionary Computation*, pp. 1–1, 2022.

[8] K. M. S. Badran and P. I. Rockett, "The roles of diversity preservation and mutation in preventing population collapse in multiobjective genetic programming," in *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 1551–1558. [Online]. Available: https://doi.org/10.1145/1276958.1277272

[9] E. D. De Jong and J. B. Pollack, "Multi-objective methods for tree size control," *Genetic Programming and Evolvable Machines*, vol. 4, pp. 211–233, 2003.

[10] M. Kommenda, G. Kronberger, M. Affenzeller, S. M. Winkler, and B. Burlacu, "Evolving simple symbolic regression models by multi-objective genetic programming," *Genetic Programming Theory and Practice XIII*, pp. 1–19, 2016.

[11] A. Rafiq, E. Naredo, M. Kshirsagar, and C. Ryan, "On the effect of embedding hierarchy within multi-objective optimization for evolving symbolic regression models," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 2022, pp. 594–597.

[12] M. Virgolin, A. De Lorenzo, E. Medvet, and F. Randone, "Learning a formula of interpretability to learn interpretable formulas," in *Parallel Problem Solving from Nature–PPSN XVI: 16th International Conference, PPSN 2020, Leiden, The Netherlands, September 5-9, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 79–93.

[13] M. Kommenda, A. Beham, M. Affenzeller, and G. Kronberger, "Complexity measures for multi-objective symbolic regression," in *Computer Aided Systems Theory – EUROCAST 2015*, R. Moreno-Díaz, F. Pichler, and A. Quesada-Arencibia, Eds. Cham: Springer International Publishing, 2015, pp. 409–416.

[14] Danishvar, M., Mousavi, A. & Danishvar, S. The Genomics of Industrial Process Through the Qualia of Markovian Behavior. *IEEE Transactions On Systems, Man, And Cybernetics: Systems*. **52**, 7173-7184 (2022)