InfoTech conference

2021 IEEE International Conference on Information Technologies

Genetic Programming using Cooperative Coevolution and Problem Decomposition for Solving Large-scale Symbolic Regression Problems

Evgenii Sopov

Department of System Analysis and Operations Research, Reshetnev Siberian State University of Science and Technology

Krasnoyarsk, Russia

email: evgenysopov@gmail.com

Mariia Semenkina

Research Center Hagenberg, School of Informatics, Communications and Media, University of Applied Sciences Upper Austria

Hagenberg Austria

email: mariia.semenkina@fh-hagenberg.at

The reported study was funded by RFBR and FWF according to the research project №18-01-00001.

InfoTech 2021, 16-17 September 2021

Motivation

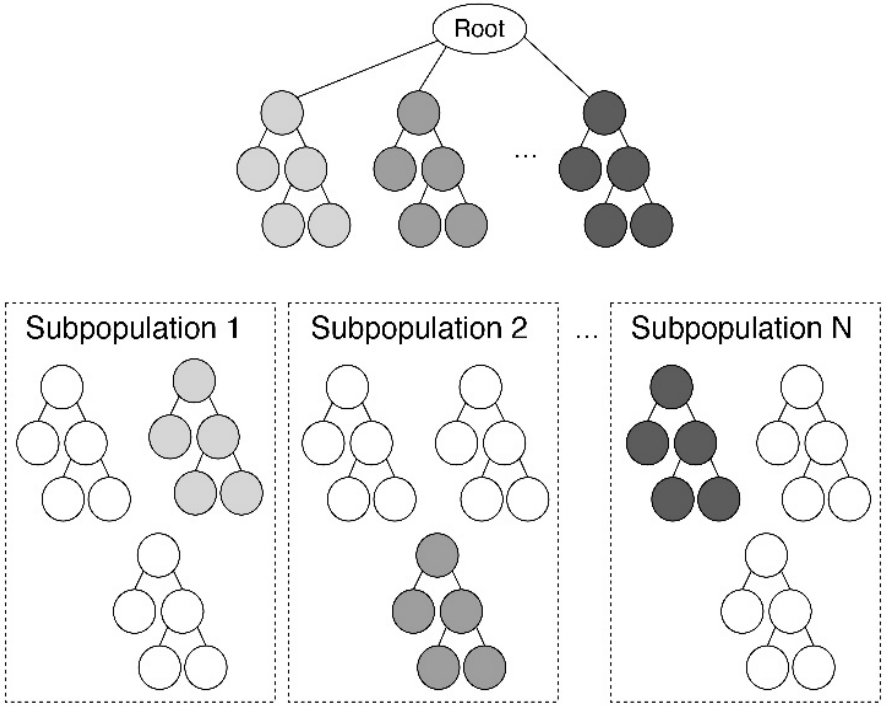
- Symbolic regression using genetic programming (SRGP) usually fails in solving high-dimensional problems.
- Many popular benchmarks mainly use one- and two-dimensional test problems and do not use more than 5 variables.
- Large-scale problems lead to rapid bloating of trees, when the size of a solution grows faster than its fitness.
- The problems require special techniques for preventing always-destructive genetic operations and the loss of variables in trees.
- In the field of evolutionary numerical optimization, there exist efficient approaches for dealing with large-scale “black-box” problems and the majority of efficient approaches are based on the conception of cooperative coevolution (CC) with problem decomposition.

The Proposed Approach. The Conception

- We will use the standard Koza-like GP algorithm with the tree representation as the core SR solver.
- For performing problem decomposition, we will evolve many sub-populations, which process pairwise disjoint subsets of objective variables.
- Each sub-population can be assigned to its own GP algorithm with specific settings, which evolve independently.
- We will apply the dynamic variables grouping strategy (namely, random adaptive grouping (RAG)).

The Proposed Approach. The problem decomposition

- Initial decomposition is performed at random.
- After a predefined number of generations (adaptation period), we will estimate the average fitness improvement for each subcomponent.
- We will keep a portion of subcomponents with the best improvements and will perform random grouping again for the rest.
- When estimating the performance of subcomponents, we will build the complete solution with all variables by merging all the best-found solutions from other sub-populations (see Figure).



- We will use a convex linear combination as the root node in the complete solution (1)-(2):

$$f(\bar{x}) = \sum_i \alpha_i \cdot f_i(\bar{x_i}) \qquad (1)$$

$$\forall i: \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1 \qquad (2)$$

where $\bar{x}$ is a vector of objective variables, $f_i(\bar{x_i})$ is a solution from the i-th sub-population, which uses a subset of objective variables, $\bar{x_i}$, $\bigcup_i \bar{x_i} = \bar{x}$ and $\forall i \neq j$: $\bar{x_i} \cap \bar{x_j} = \emptyset$.

Benchmark Problems and Experimental Setups

- We will investigate two variants of the proposed algorithm (denoted as pdccGP1 and pdccGP2) and two versions of the standard GP (sGP1 and sGP2).

$$F_1(\bar{x}) = \frac{1}{n}\sum_{i\bar{\pi}1}^{n}\left[\frac{x_i^2}{5^2}\right], \forall i:\ x_i \in [-5,5]$$

$$F_2(\bar{x}) = \sum_{i=1}^{}\left[\frac{1}{x_i}\right], \forall i:\ x_i \in [-5,5]$$

$$F_3(\bar{x}) = \frac{10}{5+\sum_{i=1}^{n}(x_i-3)^2}, \quad \forall i:\ x_i \in [0.05, 6.05]$$

$$F_4(\bar{x}) = \sum_{i=1}^{n}\left[0.05x_i^2 - \cos(1.5x_i)\right] + n, \quad \forall i:\ x_i \in [-5,5]$$

$$F_5(\bar{x}) = \sum_{i=1}^{\frac{n}{2}}[x_{2i-1}x_{2i}], \forall i:\ x_i \in [-5,5], \quad n: n \bmod 2 = 0$$

Title	Notation	Description
Forest Fires Data Set	FF	Sample volume – 517. 12 input and 1 output variables.
Wine Quality Data Set	WQ	Red wine sample is used. Sample volume – 1599. 10 input and 1 output variables (the original dataset contains 11 inputs, 1 input have been removed based on the correlation analysis).
Energy efficiency Data Set	EE	Sample volume – 768. 8 input and 1 output variables. The heating load is used as the output.
Student Performance Data Set	SP	Sample volume – 649. 30 input and 1 output variables. The final grade is used as the output.
Residential Building Data Set	RB	Sample volume – 649. 100 input and 1 output variables (the original dataset contains 103 inputs, 3 inputs have been removed based on the correlation analysis). The sale price is used as the output.

The Results

MSE for Artificial SR Problems

	pdccGP1		pdccGP2		sGP1		sGP2	
Problem (dim)	Median	Mean/Std	Median	Mean/Std	Median	Mean/Std	Median	Mean/Std
1 (10)	0.0055	0.0063/0.0025	0.0031	0.0032/0.0006	0.0088	0.0090/0.0020	0.0109	0.0099/0.0021
1 (50)	0.0309	0.0298/0.0091	0.0220	0.0210/0.0072	0.1035	0.1023/0.0162	0.1333	0.1347/0.0242
2 (10)	6.5652	6.4065/1.1275	3.9981	4.0962/0.5341	6.2478	6.5417/1.9246	6.7958	6.8421/1.2099
2 (50)	16.390	16.317/1.931	12.904	12.690/1.759	28.199	28.0960/5.040	28.164	28.671/3.416
3 (10)	0.1234	0.1304/0.0353	0.1036	0.1047/0.0180	0.1598	0.1563/0.0373	0.1329	0.1325/0.0427
3 (50)	0.2366	0.2348/0.0552	0.2072	0.2081/0.0434	0.2703	0.2641/0.0516	0.2312	0.2261/0.0633
4 (10)	2.6316	2.6477/0.3337	2.0789	2.0335/0.2233	2.9576	2.8684/0.4066	3.0164	3.0124/0.3485
4 (50)	13.088	13.192/1.0335	11.690	11.811/1.0433	18.900	18.580/1.7506	20.222	20.412/3.5080
5 (10)	19.554	19.915/1.9333	18.045	18.164/1.3045	24.094	23.903/2.2753	21.735	21.896/2.6091
5 (50)	80.345	80.663/5.0132	74.879	75.287/3.7421	89.043	89.166/5.9373	88.628	89.095/6.9822

The Number of Variables in the Best-found Solutions

Problem (Dim)	pdccGP1	pdccGP2	sGP1	sGP2
1 (10)	10	10	9	10
1 (50)	50	50	36	41
2 (10)	10	10	8	10
2 (50)	50	50	30	32
3 (10)	10	10	8	10
3 (50)	50	50	29	33
4 (10)	10	10	8	10
4 (50)	50	50	28	32
5 (10)	10	10	9	10
5 (50)	50	50	29	31

MSE for Real-world SR Problems

	pdccGP1		pdccGP2		sGP1		sGP2	
Problem (dim)	Median	Mean/Std	Median	Mean/Std	Median	Mean/Std	Median	Mean/Std
FF (12)	19.248	20.568/2.5530	18.894	19.110/1.3070	23.845	24.523/2.2969	25.570	25.024/2.5471
WQ (10)	0.6108	0.6407/0.0568	0.5016	0.5376/0.0570	0.5574	0.5964/0.0716	0.5910	0.6167/0.0623
EE (8)	8.5246	8.5362/0.0643	8.4553	8.4845/0.0646	8.5362	8.5767/0.0738	8.5389	8.5620/0.0623
SP (30)	3.3576	3.3894/0.0830	3.3011	3.3322/0.0784	4.4637	4.4380/0.1633	4.4722	4.4288/0.1729
RB (100)	202.75	202.75/9.0750	203.07	203.60/10.4885	305.00	301.57/22.2900	306.89	306.39/29.7105

The Number of Variables in the Best-found Solutions

Problem (Dim)	pdccGP1	pdccGP2	sGP1	sGP2
FF (12)	12	12	7	10
WQ (10)	10	10	9	10
EE (8)	8	8	8	8
SP (30)	30	30	16	26
RB (100)	100	100	33	57

Conclusions and Further Works

- We have combined conceptions from the field of evolutionary LSGO and have proposed a novel approach, which efficiently performs random adaptive decomposition of large-scale SR problems using cooperative coevolution.
- The experimental results have demonstrated that the proposed approach always outperforms the standard GP algorithm and builds solutions with the complete set of input variables.
- The proposed decomposition-based approach has some parameters for more deep analysis such as the length of the adaptation period and the size (sizes) of subcomponents (or the number of subcomponents).
- In further works, we will also considered the problem of interpretability of solutions.