Vinnytsia National Technical University

STRUCTURAL OPTIMIZATION OF REGRESSION MODELS BASED ON A GENETIC ALGORITHMS

1

Speakers:

Roman Kvyetnyy, DrSc in Engineering, Professor, Corresponding Member of NAES of Ukraine, Professor of the Department of Automation and Intellectual Information Technologies

Yaroslav Ivanchuk, DrSc in Engineering, Professor, Professor of the Department of Computer Science

Oleksii Kozlovskyi, Postgraduate student, Department of Computer Science

PROBLEM STATEMENT

$$\hat{y} = f\left(x; \theta, s\right)$$

$D = \left\{\left(x_i, y_i\right)\right\}_{i=1}^{N}$ - training sample;

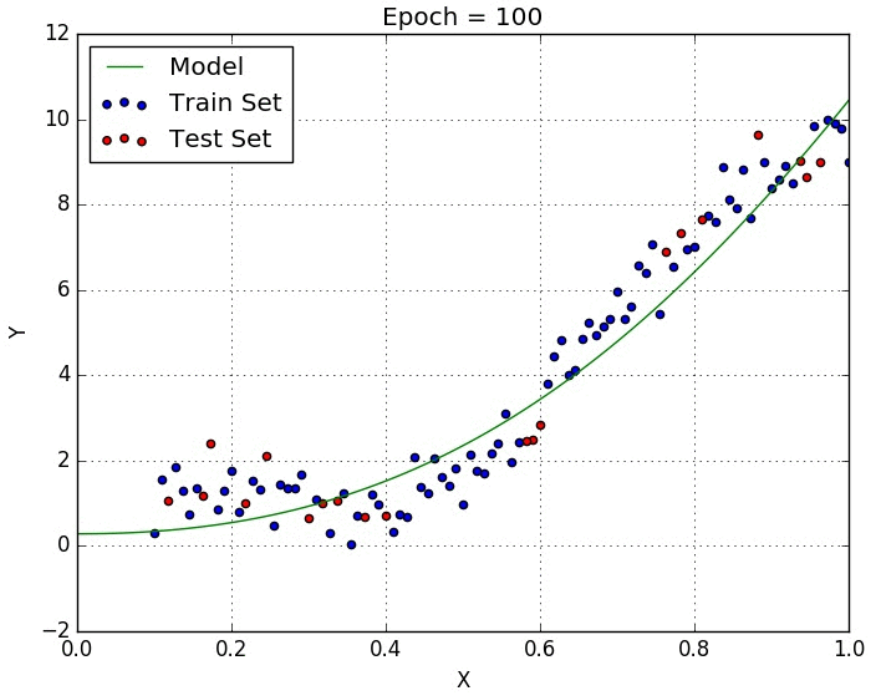
$x_i \in \mathbb{R}^p$ – vector of input features;

$y_i \in \mathbb{R}$ – observed response value;

$f \in F_s$ – class of admissible regression functions;

$\theta \in \Theta(s)$ – vector of model tuning parameters;

$s \in S$ – vector of structural variables that determine the structure of regression model;



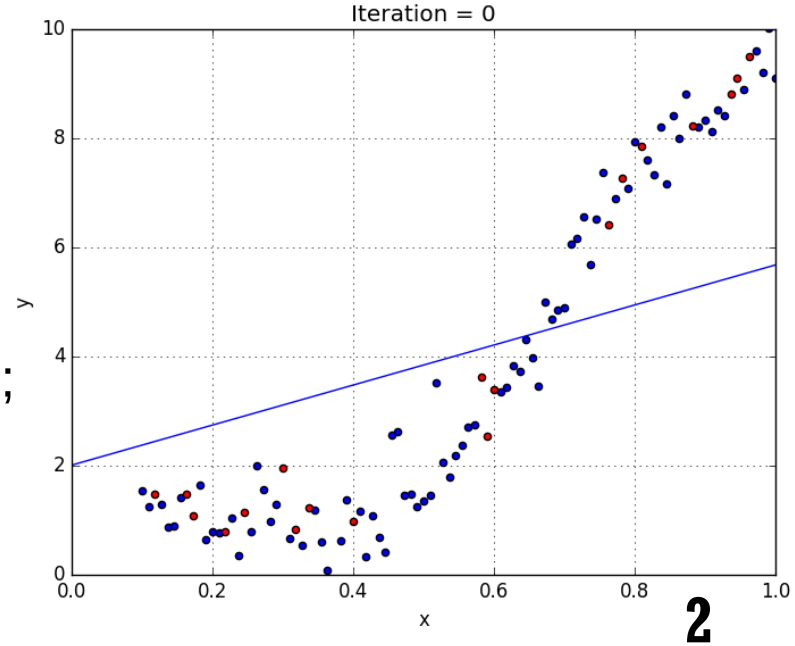
Regression model training task:

$$\frac{1}{N}\sum_{i=1}^{N} L\left(y_i, f\left(x_i; \theta, s\right)\right) + \lambda\Omega\left(\theta, s\right) \to \min_{s \in S} \min_{\theta \in \Theta}$$

$L\left(y_i, f\left(x_i; \theta, s\right)\right) = \left(y_i - f\left(x_i; \theta, s\right)\right)^2$ - mean squared error (MSE) loss;

$\Omega\left(\theta, s\right)$ – regulatory functionality;

λ – regularisation coefficient.



AIM AND OBJECTIVES

- *The aim of this work* is to improve the accuracy of predicting target variable values based on observed data by developing a structural optimization method using evolutionary computation.

Research tasks:

– to develop a genetic algorithm-based optimization method that determines the optimal architecture and type of regression model in high-dimensional feature spaces by sequentially selecting, combining, and varying candidate solution parameters using evolutionary mechanisms.

– to develop an approach for identifying the operational parameters of a genetic algorithm using a benchmark nonlinear function, which will enable the effective detection of the optimal structure of a regression model;

– to assess the effectiveness of established regression model types by applying the developed genetic algorithm-based structural optimization method to benchmark dataset models that represent the descriptive characteristics of real-world systems.

GENETIC ALGORITHM-BASED STRUCTURAL OPTIMIZATION METHOD FOR REGRESSION MODELS

$$c = (s, \theta) \quad (1)$$

s – chromosomes encoding the model structure;

θ – genes directly encoding the regression model parameters.

The chromosome value as a solution for searching the optimal number of the most important features:

$$c \equiv s = (s_1\ s_2\ \ldots\ s_p) \in B^p,\ B = \{0,1\} \quad (2)$$

$$s_j = \begin{cases} 1, \text{if the } j-\text{th feature is selected;} \\ 0, \text{if the } j-\text{th feature is not selected,} \end{cases}$$

Fitness function: $$F(c) = -J(c) = -\left[\frac{1}{N}\sum_{i=1}^{N} L\left(y_i,\ f\left(x_i^{(c)}; \theta^*(c)\right)\right) + \lambda \sum_{j=1}^{p} c_j\right]. \quad (3)$$

$\sum_{j=1}^{p} c_j$ – the number of selected features;

λ – the penalty value for an excessive number of features.

Structural optimization of a regression model is the search for a solution to another fitness function optimization problem:

$$c^* := \arg\max_{c \in \{0,1\}^p} F(c) \quad (4)$$

Population: $$P^{(m)} = \left(c_1^{(m)}, c_2^{(m)}, \ldots, c_M^{(m)}\right) \in C^M = \left(\{0,1\}^P\right)^M \quad (5)$$

m – generation number, M – population size.

GENETIC ALGORITHM FOR STRUCTURAL OPTIMIZATION OF A REGRESSION MODEL

$$P^{(m+1)} = \mathrm{E}\left(P^{(m)}\right) \cup \mathrm{M}\left(\aleph\left(S\left(P^{(m)}\right)\right)\right) \quad (6)$$

$$\mathrm{E}\left(\boldsymbol{P}^{(m)}\right) = \arg\max_{\varepsilon}\left\{F\left(\boldsymbol{c}\right) \mid \boldsymbol{c} \in \boldsymbol{P}^{(m)}\right\} \quad (7)$$

$\varepsilon \in \{1, 2, \ldots, M\}$ - number of elite individuals;

- $\boldsymbol{P}_{sl}^{(m)} = \boldsymbol{S}\left(\boldsymbol{P}^{(m)}\right) = \left\{\boldsymbol{c}_1^{sl}, \boldsymbol{c}_2^{sl}, \ldots, \boldsymbol{c}_M^{sl}\right\}$ - reproduction (selection) operator; (8)

  $\boldsymbol{c}_j^{sl} = \arg\max_{\boldsymbol{c} \in \boldsymbol{T}_j} F(\boldsymbol{c})$ - independent tournament of size τ;

- $\boldsymbol{P}_{cr}^{(m)} = \aleph\left(\boldsymbol{P}_{sl}^{(m)}\right) = \bigcup_{j=1}^{M/2} \aleph\left(\boldsymbol{c}_{2\delta-1}, \boldsymbol{c}_{2\delta}; r_\delta\right)$ - genetic crossover operator; (9)

  $r_\delta \in \{1, 2, \ldots, p-1\}$ - randomly selected crossover point for the δ-th pair of parent chromosomes;

  $\aleph\left(\boldsymbol{c}_{2\delta-1}, \boldsymbol{c}_{2\delta}; r_\delta\right) = \boldsymbol{c}'_{2\delta-1}, \boldsymbol{c}'_{2\delta}$ - two offspring chromosomes are generated according to a given rule.

- $\boldsymbol{P}_{mut}^{(m)} = \mathrm{M}\left(\boldsymbol{P}_{sl}^{(m)}\right) = \left\{\mathrm{M}\left(\boldsymbol{c}_j\right) \mid \boldsymbol{c}_j \in \boldsymbol{P}^{(m)}\right\}$ - genetic mutation operator; (10)

$$\mathrm{M}\left(\boldsymbol{c}_j\right) = \boldsymbol{c}'_j = \left(\boldsymbol{c}'_{1j}, \boldsymbol{c}'_{2j}, \ldots, \boldsymbol{c}'_{ij}\right) \quad \boldsymbol{c}'_{ij} = \begin{cases} 1 - \boldsymbol{c}_{ij}, \text{ with probability } p_{mut}; \\ \boldsymbol{c}_{ij}, \text{ with probability } 1 - p_{mut}. \end{cases}$$

REGRESSION METHODS

LINEAR REGRESSION

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + \boldsymbol{b}$$

$\boldsymbol{w}$, $\boldsymbol{b}$ – model parameter vectors

GRADIENT BOOSTING REGRESSION (GBR)

$$f(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t h_t(\boldsymbol{x})$$

$\alpha_t(\boldsymbol{x})$ – weight coefficient of the t-th decision tree

k-NEAREST NEIGHBORS (kNN)

$$f(\boldsymbol{x}) = \frac{1}{k} \sum_{j \in N_k(\boldsymbol{x})} y_i$$

k – number of nearest points,

$N_k(\boldsymbol{x})$ – k nearest points

RANDOM FOREST (RF)

$$f(\boldsymbol{x}) = \frac{1}{T} \sum_{t=1}^{T} h_t(\boldsymbol{x})$$

T - number of decision trees,

$h_t(x)$ - predicted value from the t-th decision tree

MULTILAYER PERCEPTRON (MLP)

$$f(\boldsymbol{x}) = f_L(\ldots f_2(f_1(\boldsymbol{x})))$$

L - Number of network layers

GENETIC ALGORITHM TUNING

$$y_i = 10\sin\left(\pi x_{i1} x_{i2}\right) + 20\left(x_{i3} - 0,5\right)^2 + 10x_{i4} + 5x_{i5} + \varepsilon_i; \qquad \boldsymbol{X} = \left[x_{ij}\right] \in \mathbb{R}^{N\times p} \qquad (11)$$

$x_{i1}, \ldots, x_{i5}$ - relevant features; $x_{i6}, \ldots, x_{ip}$ - noisy features; ε_i - additive noise;

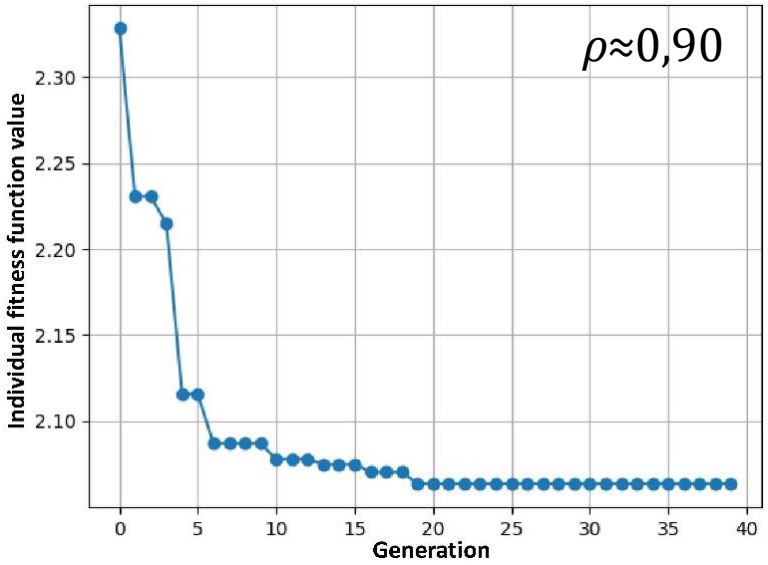
Table 1 - The target variable value based on relevant and noisy features

№	x_1	x_2	x_3	x_4	x_5	x_6	x_{17}	x_{18}	x_{19}	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	y
0	0,542	0,601	0,558	0,735	0,953	0,854	0,623	0,875	0,320	0,275	0,032	0,242	0,845	0,119	0,774	20,72
1	0,537	0,377	0,939	0,269	0,289	0,597	0,638	0,047	0,152	0,904	0,602	0,677	0,083	0,489	0,697	13,93
2	0,217	0,488	0,488	0,068	0,565	0,702	0,546	0,534	0,401	0,312	0,971	0,608	0,653	0,592	0,048	6,770

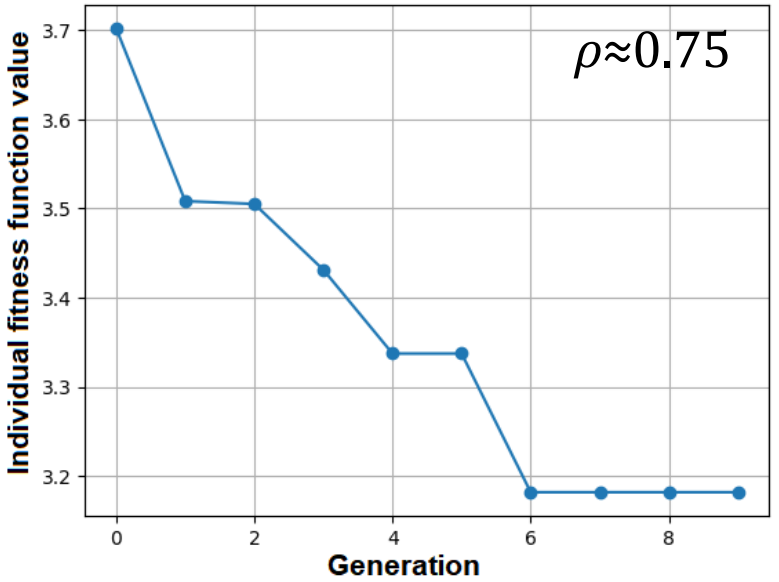
Table 2 - Genetic algorithm parameters for the Friedman-1 dataset

Type of regression model	Population size, M	Crossover/mutation probability, p_{cr}/p_{mut}	Penalty value, λ
kNN	14-50	0,7/0,2	0,01
RF	16-60	0,9/0,1	0,01
GBR	18-40	0,8/0,1	0,04
LR	16-40	0,8/0,2	0,02
MLP	18-40	0,9/0,2	0,01

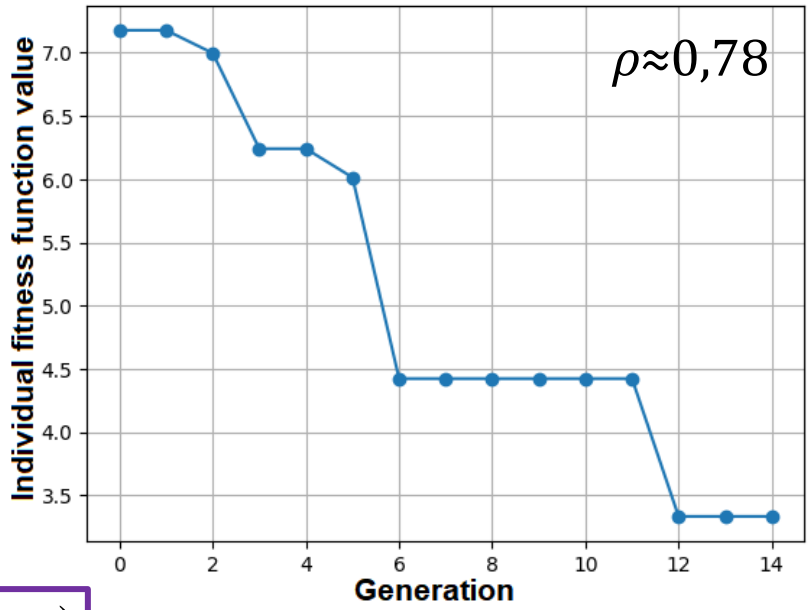
Gradient Boosting Regression



Random Forest

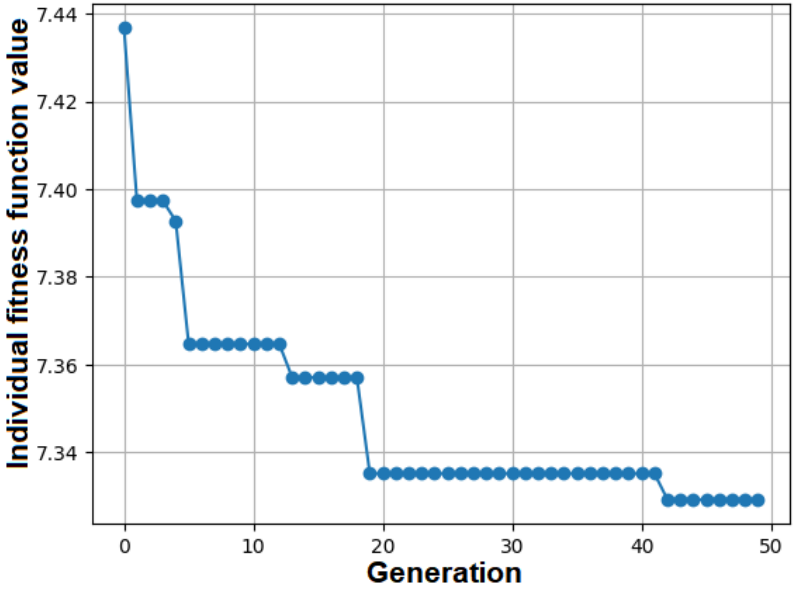


kNN

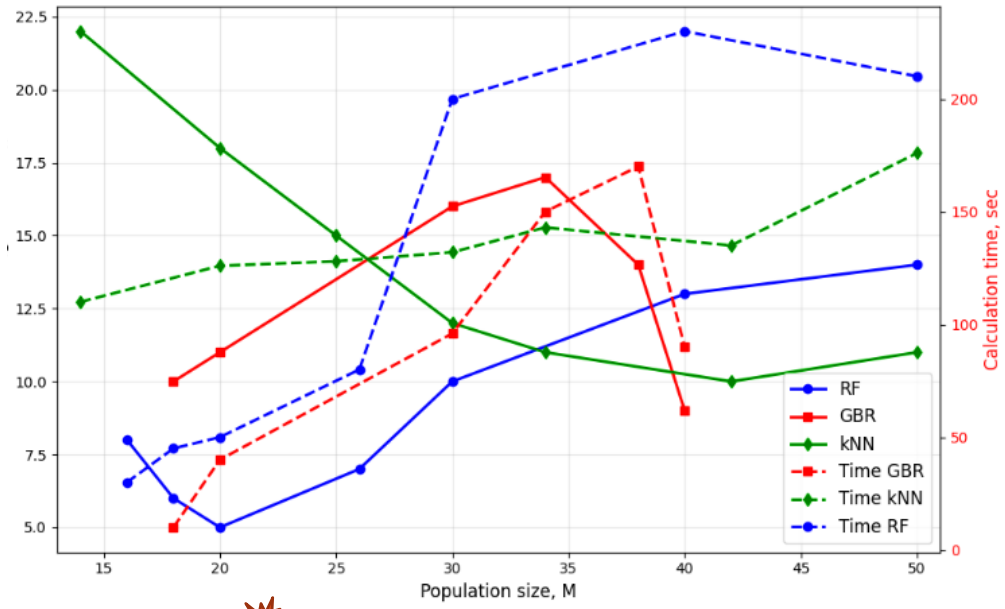


$$s = (x_1\ x_2\ \dots\ x_{15}) = (111110000000000)$$

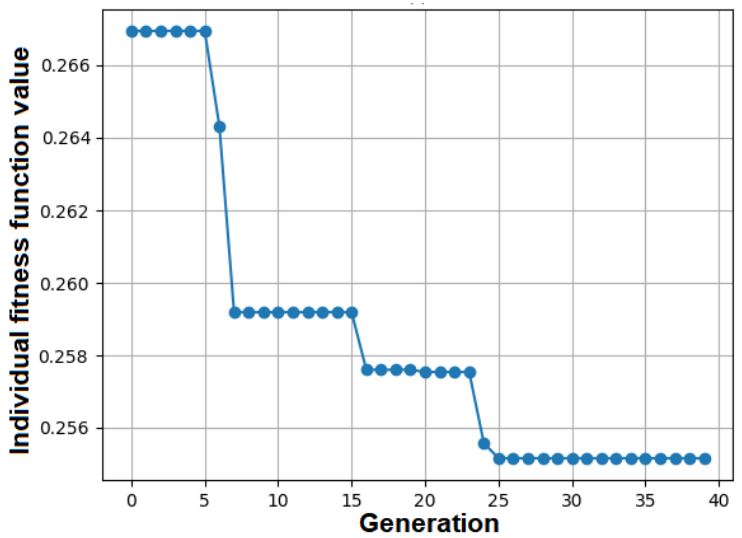
Linear regression



(111110000100100)



MLP



(111111000100000)

EXPERIMENTAL RESULTS

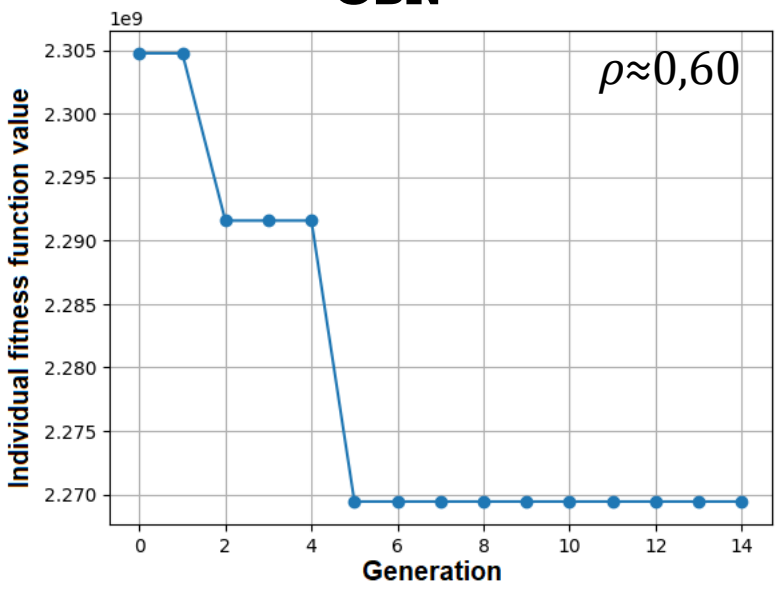
$\rho \approx 0{,}90$

Table 3. Response variable of the California Housing dataset

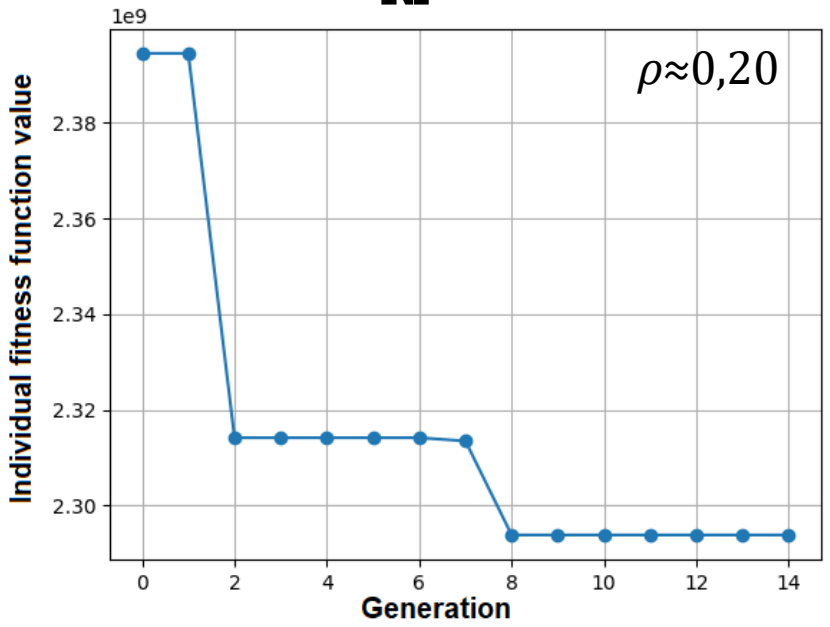
№	Longitude	Latitude	Housing_median_age	Total_ rooms	Total_ bedrooms	Population	Households	Median_income	Median_house_ value
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Y
0	-114,31	34,19	15,0	5612,0	1283,0	1015,0	472,0	1,4936	66900,0
1	-114,47	34,40	19,0	7650,0	1901,0	1129,0	463,0	1,8200	80100,0
2	-114,56	33,69	17,0	720,0	174,0	33,0	117,0	1,6509	85700,0

$$\boldsymbol{s} = (x_1\ x_2\ \ldots\ x_8) \equiv (s_1\ s_2\ \ldots\ s_8)$$

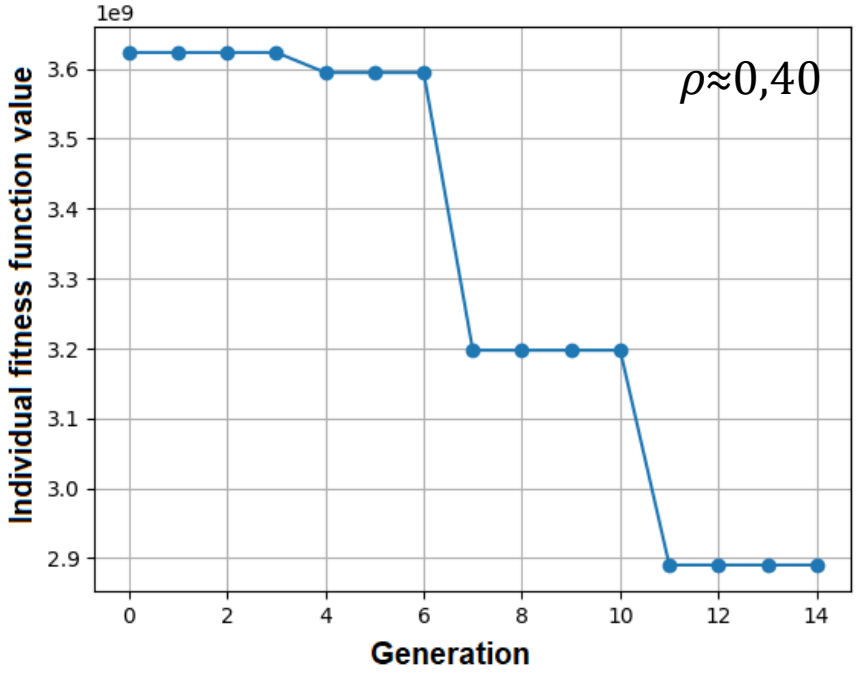
GBR



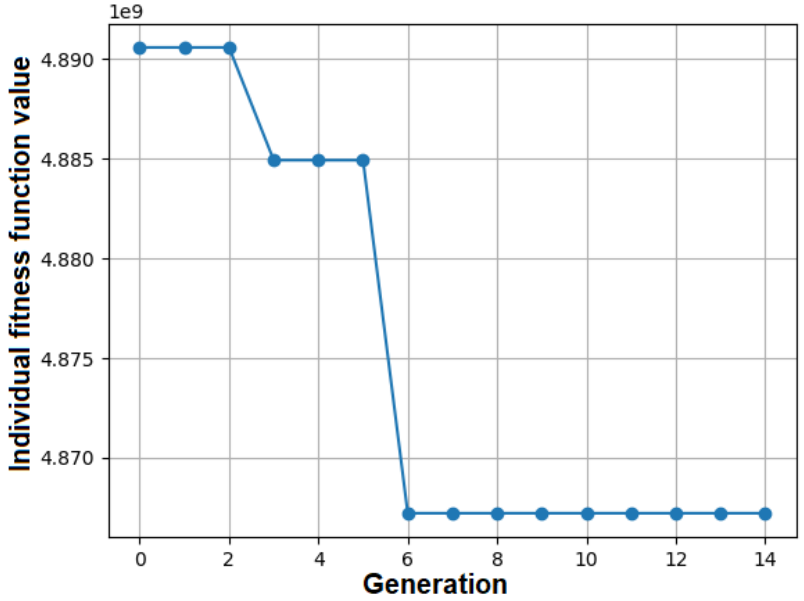
RF



kNN



LR



MLP

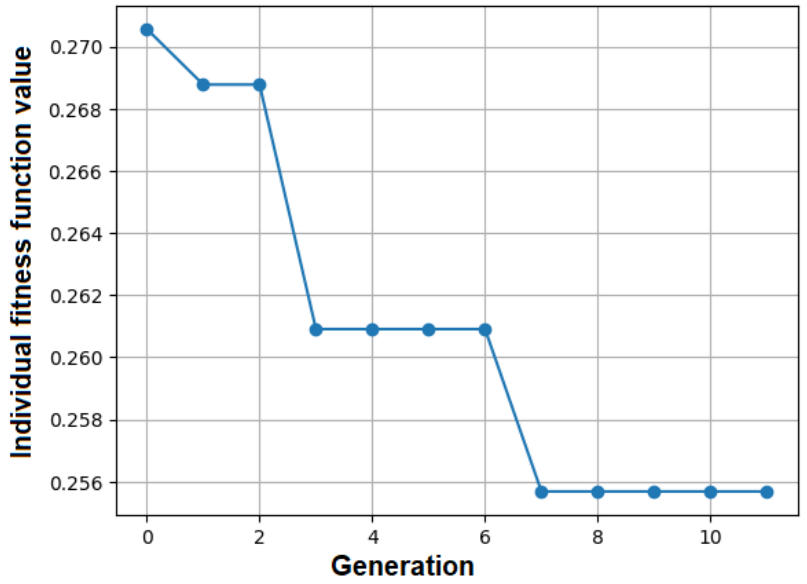
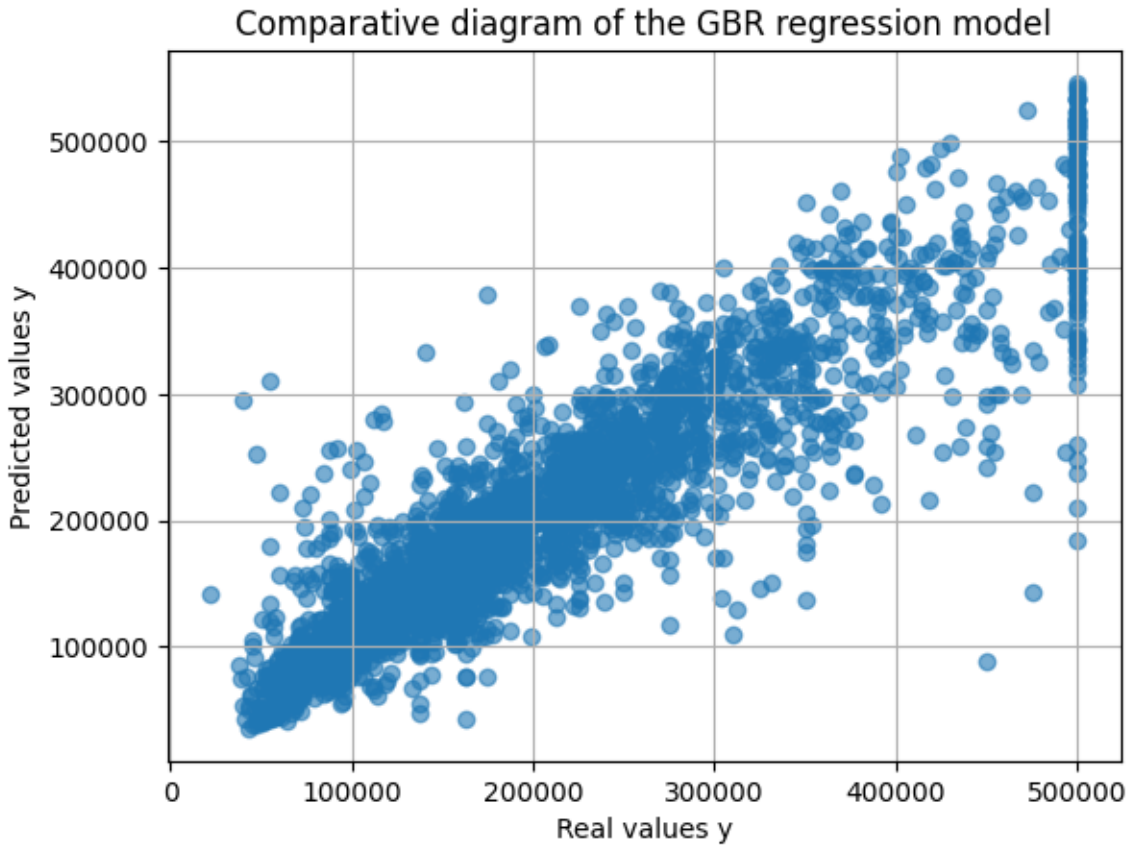


Table 4. Results of regression model structural optimization using a genetic algorithm

Regression Model	Vector of optimal features			Coefficient of determination of the baseline regression model, R^2	Coefficient of determination after structural optimization, R^2		
	Genetic Algorithm	Lasso	Recursive Feature Elimination (RFE)		Genetic Algorithm	Lasso	Recursive Feature Elimination (RFE)
kNN	[1,1,0,0,0,0,0,0]	[1,1,1,1,1,1,1,1]	[1,1,0,0,1,1,0,1]	0,70	0,77	0,70	0,70
RF	[1,1,0,1,0,1,0,1]			0,80	0,82	0,80	0,81
GBR	[1,1,1,1,0,1,0,1]			0,81	0,83	0,81	0,81
LR	[1,1,1,1,1,1,1,1]			0,62	0,62	0,62	0,60
MLP	[1,1,1,1,1,1,1,1]			0,75	0,75	0,75	0,74

REFERENCES

- The paper addresses the solution of an important scientific and practical problem of structural optimization of regression models (SORM) using a genetic algorithm (GA). The application of the GA increased the prediction accuracy of regression models by an average of 5% compared to the use of regularization methods.
- A conditional multi-objective SORM problem is formulated using a two-level minimization of the empirical risk functional on a training dataset for ensemble, local, and globally parametric regression methods. The problem is solved using a genetic algorithm with a fitness function based on a quadratic loss and a feature-number penalty.



- A parameter identification approach for the genetic algorithm based on the nonlinear Friedman-1 function was developed to determine the optimal population size, crossover and mutation probabilities, and the fitness penalty value.
- The testing results demonstrated an average increase in the coefficient of determination of the regression models by $\Delta R^2 = 0.036$, while the number of irrelevant features was reduced by an average of 46% for ensemble and local regression methods based on the genetic algorithm. In addition, the results of the GA iterative computations in SORM exhibited linear-type asymptotic convergence with an empirical coefficient of $\rho \approx 0.90$.

Thank you for your attention!

Vinnytsia National Technical University

Roman Kvyetnyy,
DrSc in Engineering,
Professor,
Corresponding Member
of NAES of Ukraine,
Professor of the
Department of
Automation and
Intellectual Information
Technologies

Yaroslav Ivanchuk,
DrSc in Engineering,
Professor, Professor of
the Department of
Computer Science

Oleksii Kozlovskyi,
Postgraduate student,
Department of Computer
Science