SYMBOLIC REGRESSION FOR FORMULATION OF SHEAR RESISTANCE OF BEARING-TYPE BOLTED CONNECTIONS

Y. Cui [(1)], M. Zhang [(2)], Q. Tang [(3)]

[(1)] *Associate Professor, State Key Laboratory of Costal and Offshore Engineering, Dalian University of Technology, Dalian, China, cuiyao@dlut.edu.cn*
[(2)] *Undergraduate Student, State Key Laboratory of Costal and Offshore Engineering, Dalian University of Technology, Dalian, China, zmz11@mail.dlut.edu.cn*
[(3)] *Ph.D. Candidate, State Key Laboratory of Costal and Offshore Engineering, Dalian University of Technology, Dalian, China, tangqi@mail.dlut.edu.cn*

Abstract

Discovering knowledge from data is a development trend of modern science. In this era of exponential data growth, symbolic regression (SR) is playing a greater role in the discovery of knowledge from data. Unlike other machine learning "black-box" algorithms, SRs can discover mathematical formulas from data quickly and accurately. When combined with domain knowledge, the explicit interpretable formulas may be scientifically meaningful.

This paper presents an application of SR in the field of civil engineering. *Eureqa*, a state-of-art SR tool, was used for uncovering the underlying design formula of shear resistance of bearing-type bolted connections. In the preliminary research phase, the dataset consisted of shear resistance of single bolted connections calculated per Specification for Structural Steel Buildings (ANSI/AISC 360-16). Since the yield to strength ratio varies from different steel grades, only the case of f_y equals 345MPa was considered here. Four independent influencing parameters, the core plate (or intermedia plate) thickness t, the core plate width W, the bolt diameter d and the clear distance between bolt hole and the edge of core plate l_c were selected as the input parameters. And the output parameter is the shear resistance F_R. The input parameters were preprocessed by nondimensionalization and normalization before training and thus the input data are all dimensionless and mapped into the range [0,1]. A 70:30 split of the data in terms of training set and validation set were selected. MAE was employed and *Eureqa* attempts to maximize this quantity in its fits. And CPU-hour was taken as the measure of algorithm efficiency.

Preliminary results show that the candidate formulas obtained are in good accordance with the formula in the design code and SR can be used as reliable algorithms for prediction of shear resistance of bearing-type bolted connections, pointing to a potential means for both the discovery of physical mechanism from experimental data, as well as the use of further developed model for artificial-intelligent-based rapid structural design. The selected four input parameters could fully describe the underlying design formula of the considered problem. In addition, some suggestion on the selection of dataset scale and operators were provided.

Keywords: high-strength bolt, bearing-type connection, shear resistance, design formula, symbolic regression

1. Introduction

Bolted connections are critical elements in overall structures. And bolt connections are quite commonly used in steel structures. To evaluate the resistance, codes of different countries, such as AISC 360-16 [1], Recommendation for Design of Connections in Steel Structures (AIJ) [2], Eurocode 3 (EN 1993-1-8:2005) [3], provide certain simplified equations by considering the safety coefficients for the evaluation. Generally, both numerical simulation and experimental study are conducted to investigate the resistance mechanism of the bolted connections. Compared with the expensive cost of experimental studies, detailed numerical modeling such as FEM is favorable. However, the main drawback of the FEM techniques is their high computation costs.

As an alternative to FE models, some authors [4-6] began to use artificial intelligence (AI) and machine learning (ML) techniques such as artificial neural networks (ANNs), which use results from experimental tests for training purposes and obtain prediction models capable of providing reliable results almost in real time. AI and ML-based models can be either explicit or inexplicit. An inexplicit AI and ML-based model function is a "black-box", which usually involves many flexible mathematical functions containing a number of parameters to fit data. Prediction accuracy is the main goal for such "black-box" models, and the interpretation of the models is secondary. Although such "back-box" models are widely applied, they are not the best choice if our goal is to purse the understanding of underlying mechanisms. Data to knowledge necessitates models with simple, explicit mathematical expressions. Symbolic regression is one of the most popular AI and ML methods to obtain explicit models from data for (exactly or approximately) describing target phenomena or mechanisms, and thus can be widely used by materials scientists and engineers to gain knowledge from data.

Symbolic regression, namely symbolic function identification, is a function discovery approach for analysis and modeling of numeric multivariate datasets. Unlike traditional linear and nonlinear regression methods that fit parameters to an equation of a given form, symbolic regression tries to form mathematical equations by searching the parameters and the form of equations [7,8]. In other words, symbolic regression method searches nonlinear equation forms and its parameters simultaneously for an addressed modeling problem. It attempts to derive a mathematical function to describe the relationship between dependent and independent variables [8,9].

The present paper proposes the use of symbolic regression (SR), based on genetic programming (GP) from evolutionary algorithms, to describe the maximum strength of a bolted connection under pure axial load. As a pre-study of the physical relationship between parameters using experimental data and finite element model data in the future study, the data prepared using the evaluations of the AISC code (AISC 360-16) [1] is used to verify the feasibility of the SR procedure. Here, the effects of the size of the training data and the selection of algorithm on the training results are discussed. The paper provides a guide to rediscover the explicit models for various types of bolted connections.

2. Problem Description

The case study presented involves a bolted connection subject to shear, as shown in Fig. 1. In this paper, to simplify the calculation case, the single bolt connection with two shear planes was considered. ASTM A529 steel plate (f_y = 345MPa, f_u = 500 MPa) and F3043 bolts ($f_{nv,bolt}$ = 620MPa) were considered for the connected plates and bolts.

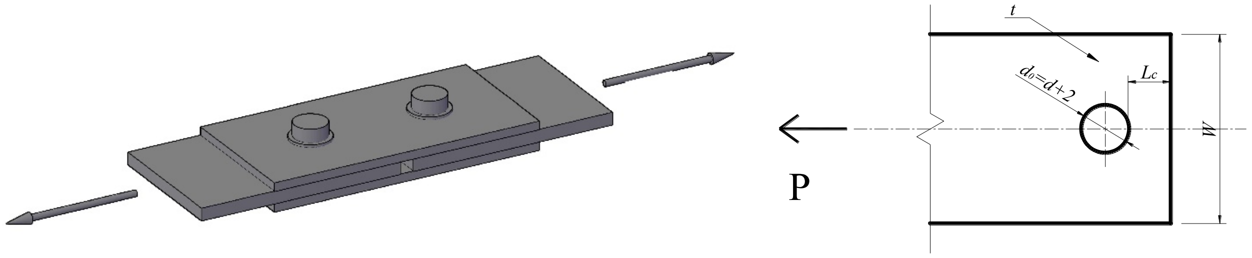


Fig. 1 – Bolted connection subjected to shear load

Three basic limit states govern the response of bolts in bolted connections: shear through the shank or threads of the bolt, bearing on the elements being connected, and tension in the bolt. The most common application of bolts in connections is to resist shear. Shear through the shank of the bolt is the means whereby the load is transferred from one plate to the other. Fig. 2 lists the basic failure modes observed for the bolted connection subjected to shear load.

According to the AISC provision (AISC 360-16) [1], the evaluations of these failure modes are listed as follows.

For the limit state of bolt shear, the nominal strength is based on the tensile strength of the bolt and the bolt size.

$$R_{\text{bolt}} = A_b f_{nv,b} \tag{1}$$

where, $f_{nv,b}$ =0.563$f_{u,b}$ is the nominal shear stress of bolt in bearing-type connections, A_b is the area of the bolt shank, $f_{u,b}$ is the nominal tensile stress of bolt.

If the bolt shank is stronger than the connected plate, the failure will be concentrated in the plate. Therefore, the failure mode of the connected plate should be considered in the design, too.

The net section failure of the connected plate could be the net section fracture or gross section yield, as described by Eq. (2).

$$R_{\text{net}} = f_u(W\text{-}d\text{-}2)t \tag{2}$$

$$R_{\text{gr}} = f_y W t \tag{3}$$

where, A_g is the gross section of the connected plate, A_n is the net section of the connected plate, f_u is the tensile strength of the connected plate, and f_y is the yield stress of the connected plate.

The Specification provision considers two limit states for bearing strength at bolt holes: the limit state based on shear in the material being connected, as shown in Fig. 2(c), and the limit state of material crushing, as shown in Fig. 2(d).

When the clear distance from the edge of the hole to the edge of the part or next hole is less than twice the hole diameter, the limit state of shear in the plate material, also referred to as tearout, will control. In this case, failure occurs by a piece of material tearing out of the end of the connection as shown in Fig. 2(c). The resistance strength of this mode is as follows:

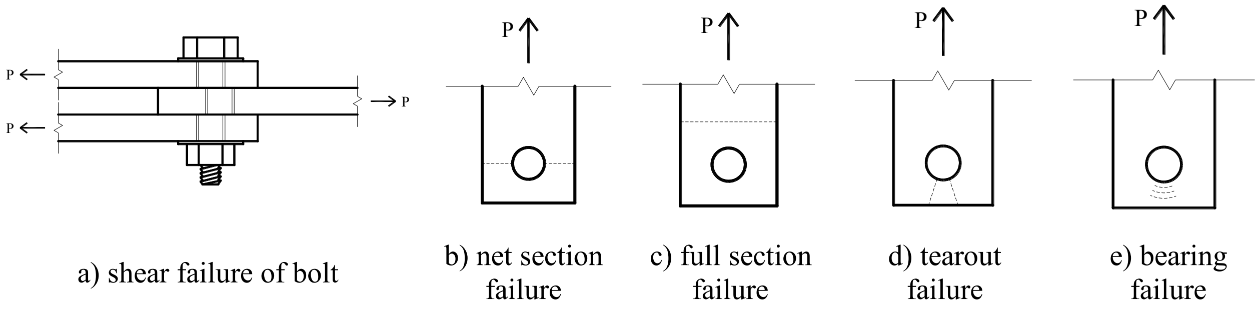
$$R_{\text{tearout}} = 0.6 f_u (2 l_c) t = 1.2 l_c t f_u \tag{4}$$

where $0.6f_u$ is the ultimate shear strength of the connected plate, t is the thickness of the connected plate, l_c is the clear distance from the edge of the hole to the edge of the connected plate.

If the clear distance exceeds $2d$, bearing on the connected material will be controlling limit states, as shown in Fig. 2(d). In this case, the limit state is that of hole distortion and the calculated bolt strength will be

$$R_{\text{bearing}} = 2.4dtf_{\text{u}} \tag{5}$$

where, d is the nominal bolt diameter, t is the connected plate thickness, f_u is the tensile strength of the connected plate.



a) shear failure of bolt

b) net section failure

c) full section failure

d) tearout failure

e) bearing failure

Fig. 2 – Failure modes of bolted connection

3. Symbolic Regression

Symbolic regression, an evolutionary function discovery method based on genetic programming, was first proposed by John Koza in 1992 [10]. It is able to extract freeform equations that correlated with data from a given experimental dataset. Different from traditional regression methods, symbolic regression is able to determine both parameters and structures of the regression models simultaneously [11]. In traditional numerical regression, the functional form is pre-defined to be linear, polynomial, or nonlinear, and the task is to determine the coefficients in the functional form. In symbolic regression, the task is to automatically find a suitable functional form in the complex data, either linear or nonlinear, and simultaneously determine the coefficients of the functions. Schmidt and Lipson [12] distilled freeform natural laws from experimental data without any prior knowledge about physics, kinematics, or geometry.

The fundamental idea of symbolic regression is rooted in Darwin's evolution theory, where the competitive mechanism ensures that the promising individuals will have more chances to survive and the individuals with poor performance will be gradually removed. There are three genetic operators commonly adopted to implement the mechanism, including selection, crossover, and mutation, as shown in Fig. 3. Based on the genetic operators, once a superior gene appears in some individuals, it will be selected, duplicated and spread across the population of individuals. Whether a gene remains in an individual during the competitive evolutionary process is determined by its contribution to the fitness of the model. In other words, only the important genes will be selected to form the models gradually, just like the survival mechanism in Darwin's theory of evolution. The emergence of a superior gene could help us to identify which factor contributes significantly to the functions found by the symbolic regression. That is, the occurrence of each factor shows its ability to describe the data, and higher frequency indicates more importance.

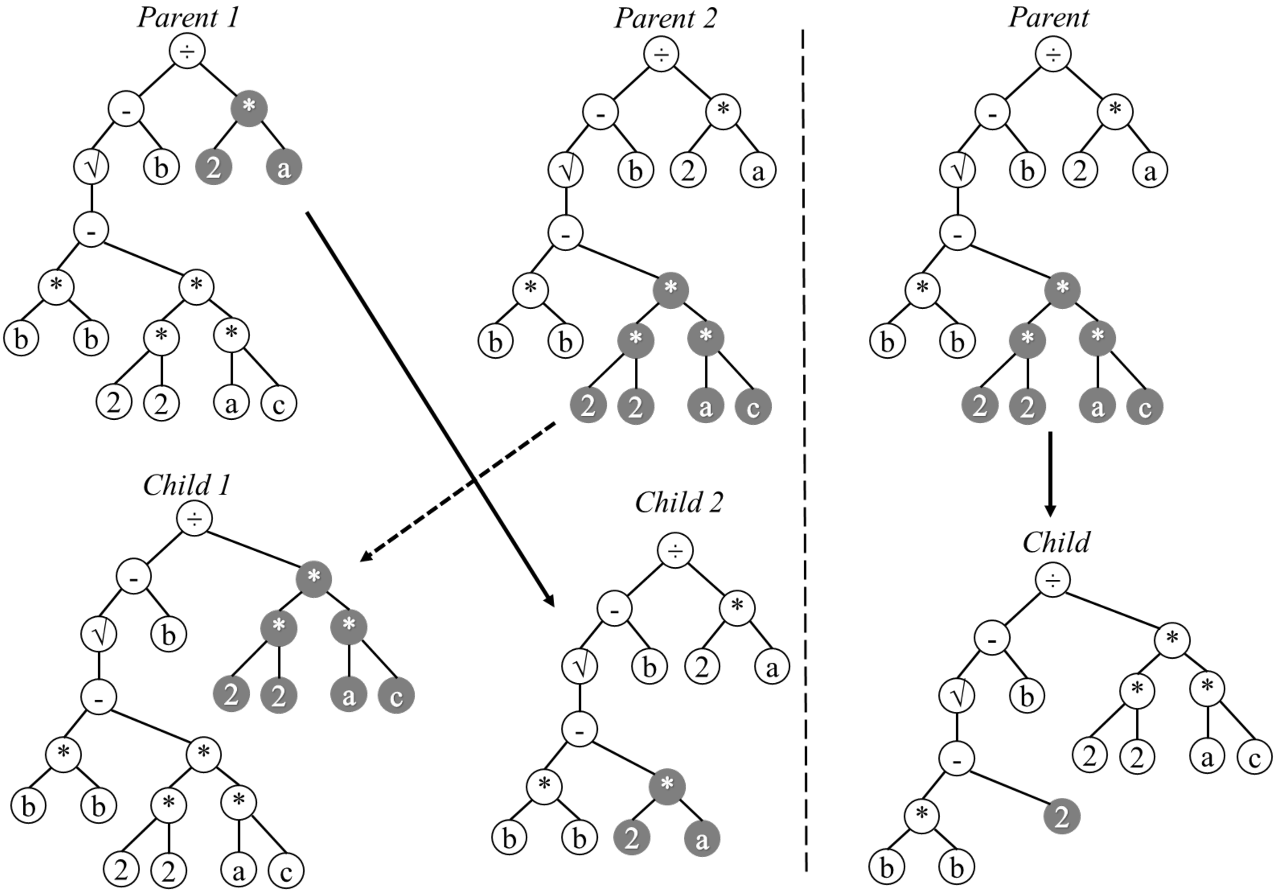


Fig. 3 – (a) Crossover, (b) mutation genetic operations in the symbolic regression

The original idea for symbolic regression was presented in [10], and it was subsequently popularized by the commercial tool *Eureqa* software [13]. In this program, candidate solutions are encoded as trees, with terminal nodes corresponding to constants and variables of the problem, while intermediate nodes encode mathematical functions such as {+, -, *, /, …}. All the nodes are collectively termed building blocks and are user-defined. The fitness function is usually proportional to the absolute or squared error between experimental data and values predicted by a candidate solution, with parsimony corrections to favor more compact equations.

The advantage of symbolic regression is that it could discover models automatically, and at the same time, such advantage will meet a huge searching space of potential solutions and a huge computing resource. Additional advantage of SR technique is the delivery of human-readable models. While ANN and SVM models are usually hard to make sense of, information can always be extracted from equations, even when they are extremely complex: a human expert could use automatically generated results to infer properties of the target phenomenon, and eventually use them as a base to build a better model.

4. Preparation of Dataset

Comparing with the dimension variations of bolted connections, the variations of the steel plate and bolt properties could be neglected. Therefore, the steel grade and bolt grade are considered as a constant value. The input parameters of the regression include the nominal diameter of bolt *d,* and width *W*, clear edge distance l_c, and thickness of connected plate *t*. The target parameter is the ultimate resistance of the bolted connection F_R. Here, the selection of these parameters is explained.

According to the AISC Specification (A360-16) [1], the common sizes of bolts are M16, M20, M22, M24, M27, M30. The plate thickness t varies from 6 mm to 16 mm, by considering the common application of plate thickness. The minimum and maximum edge distance from the center of a standard hole to an edge of a connected part in any direction are also given in the Specification (AISC 360-16) [1]. The maximum distance from the center of any bolt to the nearest edge of the connected part shall be 12 times the thickness of the connected part under consideration, but shall not exceed 150 mm. Take M16 bolt as an example, the diameter of bolt is 16 mm. The variation range of the edge distance is then 22 mm to 150 mm. Then the clear edge distance l_c varies from 13 mm to 141 mm, by considering the bolt hole size is 2mm larger than the bolt size. The variation range of the width of the connected plate is 44 mm to 300 mm.

The selected parameters and the range of the parameters are listed in Table 1. The value of parameters related to the connected plate is continuous with the incremental of 1mm. The value of bolt diameter is discrete, with the total five numbers of 16, 20, 22, 24, 27, 30.

As explained in previous, the minimum value of the Eq. (1)-(5) determine the ultimate resistance of the bolted connection, as shown in Fig. 1.

$$F_R = \min\{R_{\text{bolt}}, R_{\text{net}}, R_{\text{gr}}, R_{\text{tearout}}, R_{\text{bearing}}\} \tag{6}$$

The calculated ultimate resistance of the bolted connection used for the training is calculated using the aforementioned value, which are listed in Table 1.

Table 1 – Input and target parameters for SR

Parameter	**Input Parameter**				**Target parameter**
	d	t	l_c	W	F_R
Unit	mm	mm	mm	mm	kN
Varying range	16~30	6~16	13~141	44~300	51~634

This dataset consists of four characteristic input variables (d, t, l_c, W) and a target variable (F_R). It should be noted that in order to minimize the likelihood of numerical instabilities and/or low convergence rates, the values of the input and target parameters have been normalized in the range [0,1]. In order to ensure the design equations (Eq. (1)-(5)) of the five failure modes fairly treated in the training, the proportion of data size for each design equation should be the same. There 1000 data were selected from each design equation dataset. Therefore, there 5000 data, in which 3500 data were used for training and 1500 data were used for testing, were adopted for the following symbolic regression.

5. Training and Evaluation of SR models

5.1 Training of SR models

For symbolic regression, all candidate solutions are represented by regular functions, whose structure is determined from "building blocks" defined by sets of input variables, constants and function symbols. For the present regression, it is considered as function symbol set containing addition, subtraction, multiplication, division, power, minimum, constant, integer constant and input variable operations ("+", "-", "*", "/", "^", "min", "c", "n" and "x" respectively).

The fitness function that associates a numerical fitness value to each candidate solution and defines the problem to be solved by the genetic programming is the mean absolute error (MAE) of the normalized ultimate strength approximation against the normalized calculated ultimate strength using Eq. (6). This function is described by averaging the absolute difference of the n-observations for the i-model during the evaluation as shown by the Eq. (7).

$$\mathrm{MAE}=\frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}}\left|F'_{\mathrm{R,cal}}[i]-F'_{\mathrm{R,\,SR}}[i]\right| \tag{7}$$

where n is the total number of evaluated models before the evolution.

The generation of function's candidates was constructed by implementing the Monte Carlo's method. Random expressions from given operations and variables/constants were generated, and the fitness function was also evaluated through *Eureqa* software [13]. The best models (with the minimum fitness value) in each tournament were allocated in a new data-frame dismissing the remaining models. Then, the new data-frame was populated again evolving the best models by reproduction, mutation or crossover operations. Considering the function's complexity, the number of necessary operations in the candidates is limited filtering the candidates by i) the fitness once MAE<5%, and ii) the number of operations. With this, there is not only a better solution but a family of possible solutions from which it is selected the one that best fits the simulation responses.

When starting a symbolic regression run, the first generation is unaware that there is any target function that needs to be optimized. The initial programs encompass a totally random mix of the available functions and variables, generated from the initial population, where a random maximum depth is chosen for each individual, and the program is grown. Consequently, by generating a population of programs, the decision regarding the selection of the programs allowed to evolve into the next generation was carried out through tournaments where the fittest individuals in the tournament subsets are selected to move on to the next generation after genetic operations are performed on them.

5.2 Results of Symbolic regression

The best solution was sought considering the complexity (number of parameters and operation quantity) of the obtained expression. In addition, the dimensional consistency of physical quantities must be maintained in SR for constructing meaningful equations.

There selected candidates Eq. (8)-(10) with the validation errors listed in Table 2. It is noted that the all these three equations give quite low errors with the target parameters. It is most due to the clean of the prepared data. As it may be noted that although three equations could predict the ultimate strength accurately, the three equations exhibited different expression. Considering the failure modes as described in the previous part, it could conclude that Eq. (8) represents the physical meaning in a reasonable way.

Table 2 – Validation errors for each candidate

Name	R^2	ME	MAE	Complexity	CPU-Time
F_{SR} (1)	1.000000	0.036	0.002	31	2h 30m 16s
F_{SR} (2)	0.999999	0.165	0.010	33	11h 54m 6s
F_{SR} (3)	1.000000	0.045	0.001	38	2h 30m 22s

$$F_{SR}' = \min\,(1.46{*}d'^2,\ 1.09{*}t'{*}d',\ 1.09{*}t'{*}l_c',\ 1.77{*}t'{*}W',\ t'{*}(3.00{*}W'\text{-}0.26)) \tag{8}$$

$$F_{SR}' = 2.98{*}\min\,(0.50{*}d'^2,\ 0.36{*}t'{*}d',\ 0.36{*}t'{*}l_c',\ 0.59{*}t'{*}W',\ 1.00{*}t'{*}(W'\text{-}0.09)) \tag{9}$$

$$F_{SR}' = \min\,(1.46{*}d'^2,\ 1.08{*}t'{*}d',\ 1.08{*}t'{*}l_c',\ 1.76{*}t'{*}W',\ t'{*}(3.19{*}W'\text{-}0.29),\ t'{*}(3.00{*}W'\text{-}0.27)) \tag{10}$$

Since the material properties of the connected plate and bolt are considered as the constant value, and both the input and target parameters are normalized, the design and predicted equations are expressed in a readable way in Table 3. It is noted that only the equation for net section fracture failure mode is not well predicted. It is speculated that the variation range of the diameter is relatively limited comparing with the other parameters in the data set.

Table 3 – Comparison between the predicted equations with the design equations

Failure mode	**Design Equation**	**Predicted Equation**
Bolt shear (Fig. 2(a))	*$R_{\mathrm{cal,bolt}}$ =0.487*2*d^2	$R_{\mathrm{SR,bolt}}$ = 0.502*2*d^2
Connected plate net section fracture (Fig. 2(b))	$R_{\mathrm{cal,net}}$ =1.00*t *(W-d-2) *f_{u}	$R_{\mathrm{SR,net}}$ =1.00*t *(W-17.96) *f_u
Connected plate gross section yield (Fig. 2(c))	$R_{\mathrm{cal,gr}}$ =1.00*t *W*f_y	$R_{\mathrm{SR,gr}}$ =1.01*t *W*f_y
Connected plate tearout (Fig. 2(d))	$R_{\mathrm{cal,tearout}}$ =1.20*t*l_c*f_u	$R_{\mathrm{SR,tearout}}$ =1.21*t*l_c*f_u
Connected plate bearing (Fig. 2(e))	$R_{\mathrm{cal,bearing}}$ =2.40*d*t*f_u	$R_{\mathrm{SR,bearing}}$ =2.42*d*t*f_u

*Note that in the equation for bolt shear failure mode. The value of 0.487 represents $\frac{\pi f_{nv,b}}{4}$×0.001, where $f_{nv,b}$ = 620 MPa.

5.3 SR-model's verification with experimental results

For verification of the selected predict equation based on symbolic regression, the values of testing data set were used. Fig. 4 shows that the symbolic regression model predictions tend to the calculated results using Eq. (6), with a mean absolute error MAE=0.002 and a maximum absolute error ME=0.036. It further proved the feasibility of the selected predict equation.

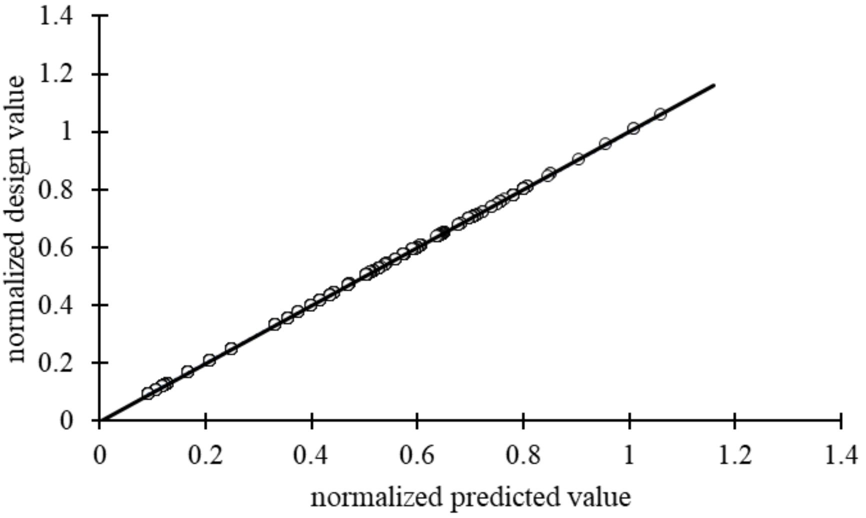


Fig. 4 – Comparsion between results from design equations and SR predict equations

6. Conclusion

Discovering knowledge from data is a development trend of modern science. In this era of exponential data growth, symbolic regression (SR) is playing a greater role in the discovery of knowledge from data. Unlike other machine learning "black-box" algorithms, SRs can discover mathematical formulas from data quickly and accurately. When combined with domain knowledge, the explicit interpretable formulas may be scientifically meaningful.

Preliminary results show that the candidate formulas obtained are in good accordance with the formula in the design code and SR can be used as reliable algorithms for prediction of shear resistance of bearing-type bolted connections, pointing to a potential means for both the discovery of physical mechanism from experimental data, as well as the use of further developed model for artificial-intelligent-based rapid structural design. The selected four input parameters could fully describe the underlying design formula of the considered problem.

In the present study, only the calculated results using design equations was considered. In the future study, the data set composed of experimental data and finite element model data will be used for training, and it is expected to quantitatively study the relationship between different failure modes and their corresponding resistance and characteristic variables through symbolic regression.

7. Reference

[1] American Institute of Steel Construction (2016): ANSI/AISC 360-16, *Specification for Structural Steel Buildings*. USA.

[2] Architectural Institute of Japan (2012): *Recommendation for Design of Connection in Steel Structure*. Japan. (in Japanese)

[3] European Committee for Standardization (2005): BS EN1993-1-8, *Design of Steel Structures-Part 1-8: Design of Joints*. Belgium.

[4] Anderson D, Hines EL, Arthur SJ, Eiap EL (1997): Application of artificial neural networks to the prediction of minor axis steel connections. *Computers & Structures*, **63** (4), 685-92.

[5] Lima LRO de, Vellasco PCG da S, Andrade SAL de, Silva JGS da, Vellasco NMBR (2005): Neural networks assessment of beam-to-column joints. *Journal of The Brazilian Society of Mechanical Sciences and Engineering*, **27** (3), 314-324.

[6] Lima LRO de, Vellasco PCG da S, Silva JGS da, Borges LAC, Silva LAPS da (2005): Post-limit stiffness prediction of semi-rigid joints using genetic algorithms. *Latin American Journal of Solids and Structures*, **2** (4), 305-320.

[7] Schmidt MD, Lipson H (2007): Genetic Programming Theory and Practice IV: Genetic and Evolutionary Computation, *Chapter 9 in Co-evolving fitness predictors for accelerating and reducing evaluations*. Springer, 2007th edition.

[8] Karaboga D, Ozturk C, Karaboga N, Gorkemli B (2012): Artificial bee colony programming for symbolic regression. *Information Sciences,* **209** (209), 1-15.

[9] Kleijnen J, Sargent R (2000): A methodology for fitting and validating meta models in simulation. *European Journal of Operational Research,* **120** (1), 14-29.

[10] Koza RJ (1992): *Genetic Programming: On the programming of computers by means of natural selection*. MIT press.

[11] Vladislavleva EJ, Smits GF, Hertog D (2009): Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming. *IEEE Transactions on Evolutionary Computation*, **13**(2), 333–349.

[12] Schmidt M, Lipson H (2009): Distilling free-form natural laws from experimental data. *Science*, **324** (5923), 81-85.

[13] Schmidt M, Lipson H (2014): *Eureqa* (version 0.98 Beta) [software], Boston, MA ⟨www.nutonian.com⟩.