UDC 004.85 10.23947/2587-8999-2020-1-2-114-119

ALGORITHM FOR A FORMATION OF A SMALL TRAINING SET USING A MULTILAYER PERCEPTRON FOR A PRIORI ESTIMATES*

M. Seleznov

✉ mihailselezniov@yandex.ru

Moscow Institute of Physics and Technology, Dolgoprudny, Russian Federation

The paper proposes an algorithm for the formation of a small training set, which ensures a reasonable quality of a surrogate machine learning model trained using this set. The algorithm uses multilayer perceptron to estimate heuristics and select the best next sample for the inclusion in a set. The paper tests the algorithm proposed applying it to the problem of deformation and breaking of a thin thread under the action of a transverse load pulse on it. The possibility to generalize the approach and apply it to building surrogate machine learning models for other physical problems is discussed.

Keywords: machine learning, train set, numerical modeling, surrogate model, multilayer perceptron.

Introduction. Solving various optimization problems, as well as inverse problems, traditionally requires enumerating a large number of combinations of parameters and solving a direct problem for each set of parameters.

For example, to optimize the parameters of a composite protective shield oriented to shock loads, it will be necessary to sort out the possible parameters of the shield, such as elastic and strength properties of the material, as well as the thickness and geometry of the shield, while for each set of shield parameters, perform calculations of the consequences of impacts by particles of different mass and shape acting at different speeds and at different angles.

Such problems are traditionally solved by methods of mathematical modeling, since carrying out a large number of full-scale experiments is an extremely time consuming and expensive process.

However, even numerical modeling can be a quite resource-intensive process - for a dynamic three-dimensional problem, the formulation can be extremely difficult, the time for calculating one direct problem can take many hours or even days [1]. If solving optimization problem requires thousands of direct problems to be solved, it becomes quite slow and costly process.

To speed up the solution of such problems, one can use surrogate models built using machine learning (ML) methods [2]. In this approach, a traditional solver for a direct problem is used to generate the data on which the ML model is trained. After the training, the ML model gives answers in a fraction of a second, compared with the hours of direct calculation. Further, it is the ML model that is used to solve the problem of multi-parameter optimization - an "instant" ML prediction is used to obtain an answer for a specific set of parameters required by the optimization problem solver. Of course, due to the nature of machine learning, the ML model's answer may be incorrect in some cases. However, if in most cases it gives a reasonable answer, then its application allows one to estimate

* The research is done within the frame of the independent R&D.

promising ranges of parameters orders of magnitude faster than using a direct solver. If necessary, these ranges of parameters can be further checked and refined by direct solvers.

A separate problem in this case is the size of the training sample, which will be sufficient to build a surrogate model of reasonable quality for describing a dynamic physical problem. Obviously, for practical problems it is highly desirable to be able to use not too large samples, since obtaining each sample for training still requires the calculation of a complete physical problem, which can require many hours of machine time.

Models and methods. This paper considers the deformation and breakage of a thin thread under the action of a transverse load pulse on it. The thread is described by the model from [3], implemented programmatically in [4].

The complete formulation of the direct problem is described by eight parameters: thread length, thread diameter, thread Young modulus, thread material density, thread material deformation limit, pressure pulse duration, pressure pulse radius, pressure pulse amplitude. The result of the calculation within the framework of this paper is a binary answer - if the thread is broken under the action of the pulse or not. The parameters and results of calculations are input data for training the surrogate model. The trained model should, according to the given parameters, give an answer - this thread remains intact when exposed to this load, or it breaks. In fact, this task comes down to assigning a set of eight input parameters of the calculation to one of two classes (possible outcomes).

The following standard methods and algorithms were used in this work to construct the ML model:

1. Logistic Regression [5];
2. Support Vector Classification [6];
3. Multilayer Perceptron with 1 hidden layer of 100 neurons [7];
4. Multilayer Perceptron with 3 hidden layers, each layer of 90 neurons.

The software implementation uses the Scikit-Learn library.

This work uses 2 data sets, described below. The direct calculations for all samples in both data sets were performed using the solver from [4].

The data set #1 was created by taking 10 values for each of the 8 parameters of the problem, and all their combinations were considered (in total, 100 million direct problem statements). Parameter values in the data set #1 vary in the following ranges: thread length (10 - 100 centimeters), thread diameter (0.01 - 0.5 millimeters), thread Young modulus (60 - 300 GPa), thread material density (1000 - 2000 kg / m3), thread material deformation limit (0.2 - 10.0 percents), pressure time (0 - 100 microseconds), pressure radius (0 - 5 centimeters), pressure amplitude (0 - 200 MPa). The selected 10 parameter values are evenly distributed over the specified ranges. For example, the values for length are 10, 20, 30, 40, 50, 60, 70, 80, 90, 100.

The data set #2 contains 9 values of each parameter (a total of 43 million problem statements). The parameter values in the data set #2 are located between the values in the first sample. For example, for length, the values in the second sample are: 15, 25, 35, 45, 55, 65, 75, 85, 95.

Results. The goal of this work was to propose the algorithm to create a relatively small data set #3, that will be used to train ML models. The data set #3 should be small to address the problem of training process being too resource intensive and time consuming. The models, trained using the data set #3, were validated using large data sets #1 and #2.

The algorithm proposed in this work uses multilayer perceptron to estimate heuristics and select the best next sample for the inclusion in a set. The algorithm works as follows:

1) Take 2 initial samples to be included into the set #3. The first sample is a thread with large values of length, diameter, Young modulus, density, strength and small values of pressure time, pressure radius, pressure amplitude. There is no thread breakage when using these parameters. The second sample is a thread with small values of length, diameter, young, density, strength and large values of pressure time, pressure radius, pressure amplitude. There is a break in this case. These two samples differ as much as possible in parameters.

2) The multilayer perceptron is trained on all current samples of the set #3. Of course, at the initial stages this training on few samples does not provide any quality, but is formally possible.

3) This helper multilayer perceptron performs prediction for 100 million samples of the set #1. Actual responses from the direct solver are not used at this stage.

4) Those samples are selected on which the helper multilayer perceptron gave the probability of finding a specimen in the class of strong threads with a probability of 60% - 80%. This threshold is significantly above the boundary value between two classes of possible outcomes.

5) For each sample from stage 4 the Euclidean distance in the parameter space is calculated to all samples without a break in the current set #3. A sample with the maximum given distance is selected (or several samples if their distances coincide). In other words, one tries to select the sample without a break (based on the perceptron prediction) which is the farthest from existing samples in the set.

6) The distances from the selected samples from stage 5 to samples with a break from the set #3 are calculated in the same way. The sample with the maximum distance is selected. If there are 2 or more samples with equal distance, then one of them is selected at random. This sample is considered the best candidate for inclusion in the training sample.

7) For the sample from stage 6, the calculation is performed by the direct solver (potentially slow resource-intensive operation) and the actual (rather than predicted) response is determined. This sample is added to the set #3.

8) Stages 2-7 are repeated until set #3 has a required number of samples (from 100 to 500 in this work).

The main indicator of the quality of prediction in this problem is the F-score metric for the class of strong threads, since other metrics can be overly optimistic based on the results of the prediction of the dominant class. All graphs below show only those algorithms that showed good results and whose results were stable when the size of the training sample was changed.

Figure 1 shows the results when using a small training set. The set #3 was used to train the models. The size of the training set varied from 100 to 500 samples. The set #2 was used for testing, which contains completely different values of parameters (compared with the set #3). Figure 2 shows the results when using random samples for training.

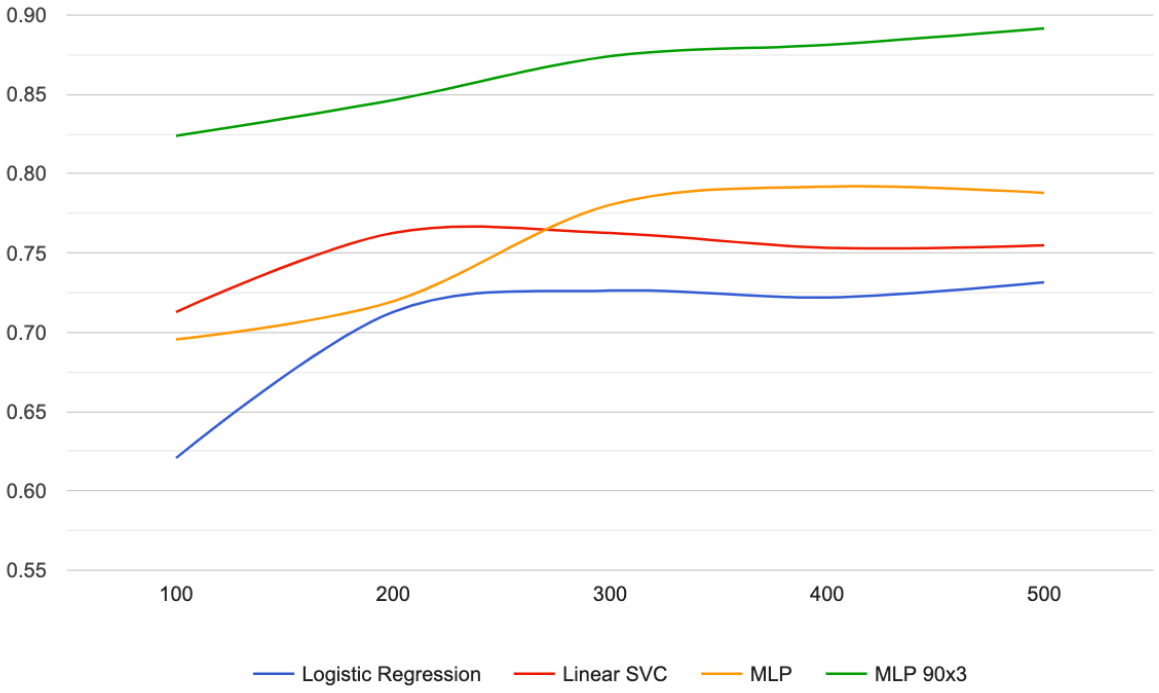

Fig. 1. F-score dependency from a train set size when using the set #3 for training and the set #2 for testing

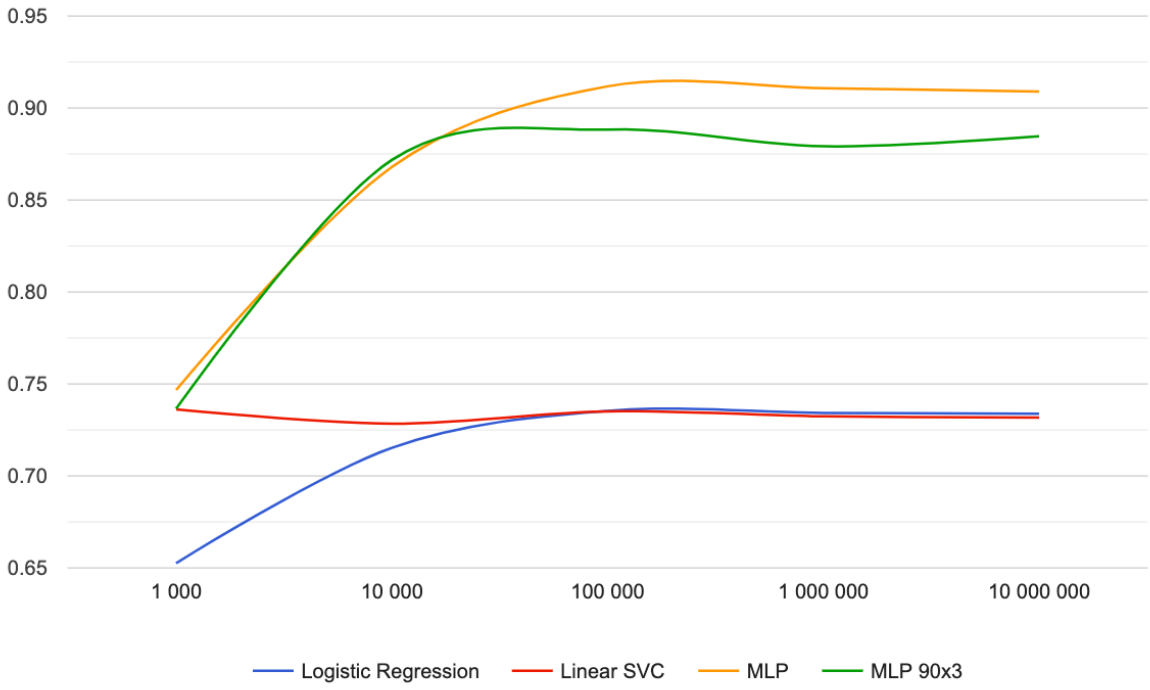

Fig. 2. F-score dependency from a train set size when using random samples for training and the set #2 for testing.

Discussion and conclusions. This paper demonstrates that ML models can provide a reasonable prediction quality for a dynamic physical problem even with a small training set. The algorithm proposed was tested on the problem of a thin thread deformation and breakage under shock transverse load, and surrogate models demonstrated F-score above 0.87 when using just 300 samples in the test set.

For the Support Vector Classification algorithm, it can be seen that the training set# 3 gives better results on 200 samples (0.76) when compared with a large set of random samples. The F-score freezes at 0.73 when using a set of 100 thousand random samples and does not grow further.

For Logistic Regression, the value of the F-score metric freezes at around 0.73 with a large training set of 100 thousand samples. At the same time, we see the same value with just 300 samples of the training set #3.

For the 3-layer perceptron, the F-score values are equal (around 0.87) when using 300 samples from the set #3 and when using 100 thousand random samples. Also, the F-score reaches 0.89 with 500 samples from the set #3, and this result cannot be achieved even with 10 million random samples in the training set.

The same approach can be applied to different physical problems, since the algorithm is formulated for an arbitrary space of parameters of a direct problem. However, the quantitative results will vary depending on the problem, and should be measured separately for each new statement.

References

1. Beklemysheva, K.A., Petrov, I.B. Damage modeling in hybrid composites subject to low-speed impact // Mathematical Models and Computer Simulations2019. – Vol. 11. – P. 469-478.
2. Kim, S.H., Boukouvala, F. Machine learning-based surrogate modeling for data-driven optimization: a comparison of subset selection for regression techniques // Optimization Letters. – 2020. – Vol. 14. – P. 989-1010.
3. Rakhmatulin Kh.A., Demianov Yu.A. Strength under high transient loads. – New York : Daniel Davey, 1966. — 348 p.
4. Vasyukov A.V., Elovenkova M.A., Petrov I.B. Modeling of thin fiber deformation and destruction under dynamic load // Matem. Mod. – 2020. – Vol. 32, iss. 5. – P. 95-102.
5. Hsiang-Fu Y., Fang-Lan H., Chih-Jen L. Dual coordinate descent methods for logistic regression and maximum entropy models // Machine Learning. – 2011. – Vol. 85. – P. 41-75.
6. Platt J. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: Advances in Large Margin Classifiers // Cambridge : MIT Press, 1999. – P. 61-74.
7. Rumelhart D.E., Hinton G.E., Williams R.J. Learning internal representations by error propagation. In: Parallel distributed processing: Explorations in the microstructure of cognition // Cambridge : MIT Press, 1986. – P. 318-362.

Author:

Mykhailo Seleznov, Researcher, Moscow Institute of Physics and Technology (Institutskiy lane, 9, Dolgoprudny, Russian Federation)

УДК 004.85 10.23947/2587-8999-2020-1-2-114-119

АЛГОРИТМ ФОРМИРОВАНИЯ МАЛОЙ ОБУЧАЮЩЕЙ ВЫБОРКИ С ИСПОЛЬЗОВАНИЕМ МНОГОСЛОЙНОГО ПЕРСЕПТРОНА ДЛЯ АПРИОРНЫХ ОЦЕНОК *

М. Селезнёв

✉ yuliadanik@gmail.com, mdmitriev@mail.ru, olga.proncheva@gmail.com

Moscow Institute of Physics and Technology, Dolgoprudny, Russian Federation

В статье предлагается алгоритм формирования небольшой обучающей выборки, обеспечивающей приемлемое качество суррогатной модели машинного обучения, обученной с использованием этой выборки. Алгоритм использует многослойный персептрон для выполнения предварительной оценки и выбора следующего лучшего образца для включения в выборки. В статье тестируется предложенный алгоритм применительно к задаче о деформации и разрыве тонкой нити под действием на нее импульса поперечной нагрузки. Обсуждается возможность обобщения подхода и его применения для построения суррогатных моделей машинного обучения для других физических задач.

Ключевые слова: машинное обучение, обучающая выборка, численное моделирование, суррогатная модель, многослойный перцептрон.

Автор:

Селезнёв Михаил, Научный сотрудник, Московский физико-технический институт (РФ, г. Долгопрудный, Институтский пер., д. 9)

* Работа выполнена в рамках инициативной НИР.