Data-Driven Delta Machine Learning Models for Improved Extrapolation

Adam BIRCHALL [a*], Isaac CHANG [a], Zidong WANG and Carla BARBATTI, [b,c]

[a*] Brunel University London, BCAST, Uxbridge, United Kingdom

[b] Constellium University Technology Centre, Brunel University London, Uxbridge, United Kingdom

[c] Constellium Technology Center, Parc Economique Centr'alp, Voreppe, France

ABSTRACT

This work demonstrates a method of testing a machine learning model's extrapolation accuracy, a capability that is significant to efficiently aid with discovery of improved alloys and presents the application of a pure data-driven method to make steps to reduce this extrapolation error. By using linear models to capture general trends in the data and then the subsequent application of more complex machine learning methods, extrapolation capabilities can be reduced. Being purely data-driven, this type of model can be coupled with other Delta-Machine Learning techniques such as those that utilize physics-domain knowledge, and coupled with active learning methods, with better extrapolation capabilities reducing the number of iterations needed to outperform existing alloys.

1. Introduction

When using machine learning techniques to attempt to improve upon an existing dataset, the extrapolation, and consequentially generalization ability, of a model needs to be a priority for a model to accurately predict beyond what has come before. The prediction of mechanical properties of an alloy given its composition and processing is often a goal of many models, with the aim of using such a model to find inputs which result in improved properties.

Testing of such models generally demonstrates high accuracies when using standard cross validation techniques, but few fail to demonstrate their poor performance when attempting the aim of accurate predictions of properties beyond the dataset, with models generally being phenomenological, generalizing poorly [1].

Linear regression models generalize a great deal and as such often result in poor accuracies, but capture the underlying, linear, relationships in a dataset. However, in combination with a higher bias models, linear models can provide the general trend which other models then predict the difference of, resulting in higher accuracies [2].

Delta machine learning models are such models where a primary model maps a generalized view of the domain, with a secondary model providing predictions of the errors, or deltas, of the first.

2. Experimental procedures

Using data collected from the online database, MatMatch [3], the composition, processing, and yield strength of aluminum alloys was recorded into a database. The composition of the alloys and the temper that each entry received was used to form the input to models with yield strength predictions as the target output.

To test extrapolation capabilities, the top 20% of alloys based on yield strength values were set aside to be used as an extrapolation test set, with the remaining 80% used as an interpolation set. Of this interpolation set, a subsequent split was made of 80% for training data, and 20% for the judgement of interpolation accuracy. Models were then trained upon this interpolation training set and evaluated on both interpolation and extrapolation capabilities using respective datasets.

Delta models consisted of an initial, high-generalization model, linear regression, used to make general predictions, the errors of which a subsequent higher bias model would attempt to predict. Through the combination of both general predictions from the first model, and error predictions of the second, a final prediction from the Delta model is made.

Linear models, Random Forests, SVRs, Neural Networks, and Delta models were all tested with standard scaling techniques used where appropriate, and all implemented within python utilizing Scikit-learn and modules made to conform to its API [4].

3. Results and discussion

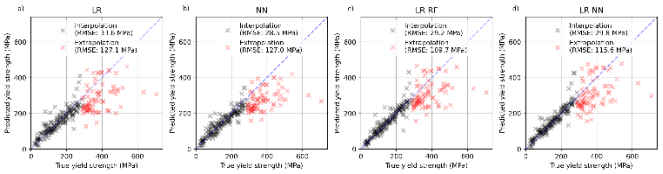

Figure 1. The prediction of yield strength of the dataset using a a) Linear regression model, b) Neural network, c) Delta Model composed of linear regression and Random Forest, d) Delta model composed of Linear regression and Neural Network. While the Neural network achieves the best interpolation results, it is surpassed in extrapolation accuracy by the Delta model composed of linear regression and random forest.

The results shown in Figure 1 demonstrate the improved extrapolation capability of this data-driven delta model.

Delta models, under different names, have been used for materials problems before [5], but are often based upon a physics model as the initial generalization model, these can result in good accuracies, but a physics-based domain model is often not available. This data driven delta model can be applied quickly and separately or in combination with such methods to achieve further improvements.

These models also have active learning use cases, as a substitute model with better extrapolation capabilities, aiming to reduce the number of iterations needed to outperform datasets which the extension to this work will attempt to show.

Acknowledgments

This work was completed as part of an industrial CASE studentship funded by both the Engineering and Physical Sciences Research Council (EPSRC) and industrial partner Constellium.

Reference

[1] N. Fujinuma, B. DeCost, J. Hattrick-Simpers, and S. E. Lofland, 'Why big data and compute are not necessarily the path to big materials science', *Commun Mater*, vol. 3, no. 1, Art. no. 1, Aug. 2022, doi: 10.1038/s43246-022-00283-x.

[2] H. Zhang, D. Nettleton, and Z. Zhu, 'Regression-Enhanced Random Forests'. arXiv, Apr. 23, 2019. Accessed: Oct. 17, 2022. [Online]. Available: http://arxiv.org/abs/1904.10416

[3] MatMatch, 'Make better material decisions'. [Online]. Available: https://matmatch.com/

[4] F. Pedregosa *et al.*, 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[5] L. Lu, R. Pestourie, S. G. Johnson, and G. Romano, 'Multifidelity deep neural operators for efficient learning of partial differential equations with application to fast inverse design of nanoscale heat transport', *Phys. Rev. Res.*, vol. 4, no. 2, p. 023210, Jun. 2022, doi: 10.1103/PhysRevResearch.4.023210.