Springfield Publishers

Open Access Journal of Engineering Sciences

Open access

Mini Review

Submission: March 05, 2020 | **Published:** March 11, 2020

COMPOSITES: A Pragmatic Knowledge-Based Engineering Data Boosting Process

Sunil Chandrakant Joshi*

School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore

***Corresponding author:** Sunil Chandrakant Joshi, School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore Nanyang Avenue, Singapore, 639798 Phone: +65-67904725/ 67905170.

Abstract

Machine Learning (ML) is growing in engineering. It is often a challenge to gather sufficient and representative data for ML. Experiments are possible only in small numbers due to the constraints on materials, manufacturing, or testing. Simulation and models need skills and appropriate validation. If none of these meets the requirements for the quality and the size of the dataset required to train ML, professionals tend to create synthetic data, which may not address the quality aspect of the new data. A pragmatic 10-step knowledge-based data boosting process is proposed to systematically address data sparsity for ML in engineering without compromising the data quality.

Keywords: Knowledge-based Data Boosting, Machine Learning, Engineering Data, Data quality

Introduction

The fields of Artificial Intelligence and Machine learning (ML) are fairly well developed and are recognized as promising tools in engineering and manufacturing [1], especially for materials design [2], manufacturing [3], optimization [4] and prognosis [5]. ML allows a machine to learn from heritage without being explicitly programmed. ML depends greatly on algorithms and data. Many algorithms and programming frameworks are now available [6]. However, it is impossible for a machine to learn without data. The dataset/s shall be accurate and sufficient for any sound ML exercise. In many cases, synthetic data [7] or data augmentation process [8] are adopted. Use of synthetic data, fully or partially, is not advisable for engineering applications without understanding, as the unreliability involved is high. Data augmentation certainly works well with imaging. However, boosting numerical data is not possible using those techniques. Statistical methods can certainly help in generating numbers, but there is no guarantee that the generated data is accurate and representative of the property or the parameter in question. These cannot replace the engineering judgment and acumen required in making reliable predictions about an engineering parameter.

Experiments may be conducted to produce more data, provided they are affordable. Simulation and numerical models also need skills and proper validation. If either one or both of these cannot meet the requirements for the quality and the size of the dataset to train ML, a holistic approach is necessary to address the quality aspect of the new data while generating it to boost the existing dataset. In this paper, a knowledge-based data boosting process is proposed to address an engineering data sparsity systematically and reliably without compromising the data quality.

Knowledge-Based Engineering Data Boosting (KEDB) Process

Generation of any engineering data requires basic understanding of the parameters and phenomena involved. Adding new data points to an available dataset requires the same attention and rigor in order to maintain accuracy and reliability. Below are TEN generic steps that sum-up the proposed KEDB process, which also form an acrostic 'COMPOSITES'.

a) Collect (Gather authentic raw data.)

b) Organize (Sort and select data points based on the certain parameter or criterion.)

c) Mathematics (Tabulate and/or plot the data points. Check scatter, trends.)

d) Physics (Examine the underlying reasons.)

e) Oddities (Remove outliers, extreme points.)

f) Space (Mark the space domain within which data may be boosted.)

g) Infer (Build guidelines to be observed while boosting the data.)

h) Translate (Form mathematical propositions based on the inferred guidelines.)

i) Employ (Apply and adopt those propositions to augment the data and visualize.)

j) Scrutinize (Examine the new dataset. Look at every detail. Build suggestions.)

Once an engineering parameter requiring prognostic model to be built is identified, the related raw data may be collected. Sources can be the literature, own laboratory testing, own finite element simulation, other analyses, or even old databases. Organize the data gathered in a form where it can be looked at as a dataset and/or as subsets. Tabulation and graphical tools may be used such that the data at hand may be viewed as a group. Further, use mathematical measures and statistical functions to get minima, mean, maxima, regression lines, standard deviation, R-squared (or, Coefficient of Determination), for any two-dimensional presentation of the data. These provide necessary elements to understand the behavior of the dataset at hand. Conduct further investigation to understand the physics behind those data trends and the behavior. Only domain experts can provide correct insights into those and help identify the oddities including outliers, boundaries and the zones of the lean data. If the raw data has been already sanitized, this task becomes relatively easy. Otherwise, expert may need to spend time in weeding out suspicious data points. Once that is done, space for data boosting can be defined and marked. This is to avoid generation of inappropriate data points or avoid addition of data points that may violate the physics behind the original data. The next step requires drawing inferences from the fully sanitized original dataset and defining guidelines for data boosting. These guidelines can be, for example, to affect zonal densification, to expand the data band while balancing the upper and lower bounds, to add points in the vicinity within the error margins, or to populate the defined space. The inferences can be in the form of limiting percentages, error margins, extrapolation, interpolation, limits, etc., based on the likely variations in the material, processes or testing scenarios that can occur in real life. Once finalized, employ those guidelines using possible means to generate newer data. The tools can be as simple as Boolean operations, parametric manipulation in regression lines, or use of numerical methods [9]. The generated data points shall be scrutinized individually and as part of the new dataset. Regression lines may be redrawn, and their coefficient of determination may be studied to evaluate the impact of the new data points. This shall ensure that the synthetic data added to the original dataset is as authentic as any data that can be generated experimentally or realistically. Thus, formed dataset, may be termed later as a hybrid dataset, will certainly be reliable for ML.

Conclusion

A knowledge-based engineering data boosting (KEDB) method is devised for reliable data augmentation in engineering. The proposed, 10-step COMPOSITES process provides a sound, systematic and pragmatic approach.

References

1. Chen CT, Gu G (2019) Machine learning for composite materials. MRS Communications 9(2): 556- 566.
2. Sherman S, Simmons J, Przybyla C (2019) Mesoscale characterization of continuous fiber reinforced composites through machine learning: Fiber chirality. Acta Materialia 181: 447-459.
3. Sharp M, Ak R, Hedberg T (2018) A survey of the advancing use and development of machine learning in smart manufacturing. J Manuf Syst 48: 170-179.
4. Fernandes H, Zhang H, Figueiredo A, Malheiros FIL, Sfarra S, et al. (2018) Machine Learning and Infrared Thermography for Fiber Orientation Assessment on Randomly Oriented Strands Parts. Sensors 18(1): 288.
5. Pathan MV, Ponnusami SA, Pathan J, Pitisongsawat R, Erice B, et al. (2019) Predictions of the mechanical properties of unidirectional fibre composites by supervised machine learning. Sci Rep 9(1): 13964
6. Sengupta S (2019) Machine Learning Algorithms in 7 Days. (1st Edn), Safari an O'Reilly Media Company.
7. Bisong E, Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners. Chapter 14 Principles of Learning pp: 176.
8. Surendra H, Mohan HS (2017) A Review of Synthetic Data Generation Methods for Privacy Preserving Data Publishing. International Journal of Scientific & Technology 6(3).
9. Babu Ram (2010) Numerical Methods. India, ISBN: 9788131732212.

-- *** --