Faculty of Science and Engineering

Materials design through ensemble learning: The impact of very small data sets

Danny E.P. Vanpoucke [A], Onno S.J. Van Knippenberg [B], Ko Hermans [B], Siamak Mehrkanoon [C], Katrien V. Bernaerts [A]

[A] Maastricht University, Aachen-Maastricht Institute for Biobased Materials, Brightlands Chemelot campus, Urmonderbaan 22, 6167 RD Geleen, The Netherlands
[B] CCL Olympic B.V., Keizersveld 30, 5803 AN Venray, The Netherlands
[C] Maastricht University, Department of Data Science and Knowledge Engineering, Paul Henri Spaaklaan 1, 6226 GS Maastricht The Netherlands

Background: Artificial Intelligence in materials design

Over the last few decades, computational modeling and simulations have become an integral part of modern materials design. As highly specialized materials properties are rooted in the atomic scale behavior of materials, quantum chemical modeling has taken on a central role in the design of novel materials by providing fundamental insights in the underlying mechanisms.

More recently also **Artificial Intelligence (AI)** and **Machine Learning (ML)** are starting to play a more important role in the field of materials science. This is made possible through the access to big data sets of both theoretical and experimental origin. Although **large data sets** are becoming more common-place, they do not represent the **typical data sets most materials chemists work with** or generate in their day-to-day labor. In general, those data sets **are much smaller**.

Given the successes of Machine Learning with large data sets, there is a growing interest in their application on small data sets, in hope of reaping similar successes here.

Small, smaller … trouble

Standard ML approaches deal with at least thousands and often millions of data samples. But what would happen if you only have a handful?

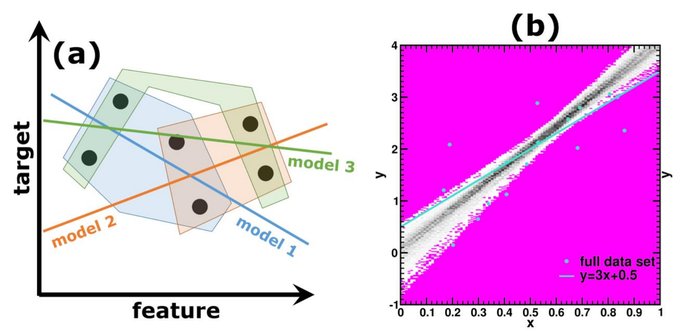

Figure 1: (a) schematic representation of linear model instances, trained on different subsets. (b) Heat map of 1000 model instances trained on a 20 point data set. **[1]**

⇒ Broad distribution of possible model instances

⇒ Model instances appear spread around a central optimum

AMADEUS framework: An ensemble approach [2]

The AMADEUS framework is build around the **scikit-learn library**, and provides additional functionality for handling small data sets:

- Error-handling & sanity checks
- Parallelization over model instances
- Automation of hyper-parameter tuning
- Construction of *Average* model

The core aspect is an ensemble approach to modeling the small data set. A large number of random train-test splits are generated, and on each a model instance is trained.

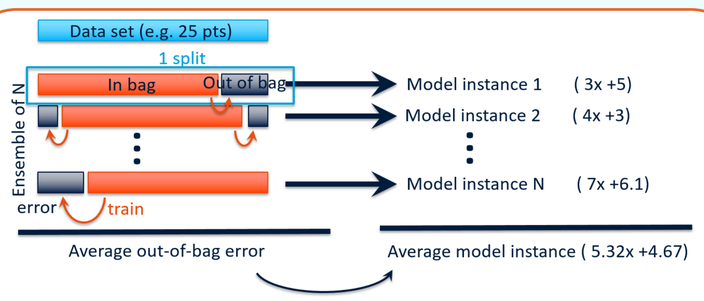

Figure 2: Ensemble approach at the core of AMADEUS

Modeling small experimental data sets

Experimental data set: Coating formulation [3]

Features: 3

1. 2-ethylhexyl acrylate (2EHA)
2. Acrylic acid (AA)
3. N-vinyl caprolactam (VCL)

Targets: 2

1. peel strength on steel (PSS)
2. elongation at break (EB).

Repeated computer experiments

- 100 repetition of ensembles of 1000 model instances.
- Best & Worst model instance are selected. Average is calculated (*cf.*, box below)

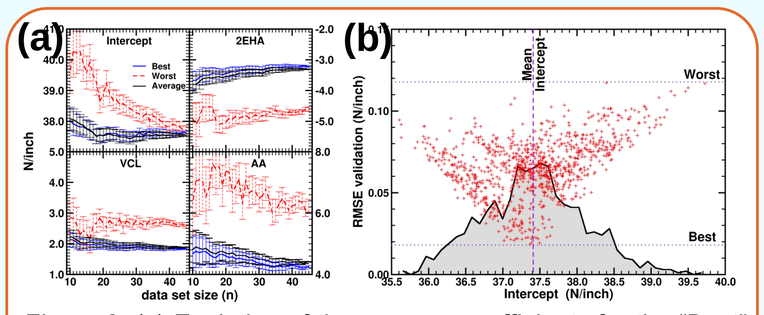

Figure 3: (a) Evolution of the average coefficients for the "Best", "Worst", and "Average" model instance. (b) Volcano-plot relating model quality to coefficient value. **[1]**

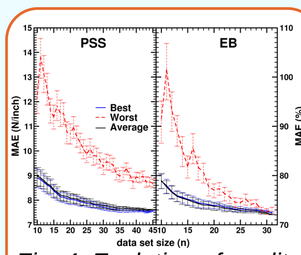

Fig. 4: Evolution of quality measure with data set size.

Average instance is not so average

1. Coefficients coincide with those of the best model instances.
2. Quality measure equal to that of the best model instances.
3. Average does not rely on lucky circumstance, so it might outperform best instance in small ensembles.

Two routes to prediction

Ensemble averages are well known to present improved accuracy over individual weak learners. In case of simple polynomial models, it can analytically be shown that the prediction by an average model is identical to the ensemble average.

$$\begin{aligned}\bar{y} &= \frac{1}{N}\sum_i^N y_i \\ &= \frac{1}{N}\left(\sum_i^N\left[b_i + \sum_j^h a_{i,j}x_j\right]\right) \\ &= \frac{1}{N}\sum_i^N b_i + \sum_j^h x_j\left(\frac{1}{N}\sum_i^N a_{i,j}\right) \\ &= \bar{b} + \sum_j^h \bar{a}_j x_j = y_{avg}\end{aligned}$$

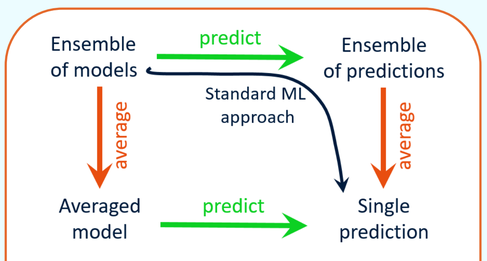

Figure 5: Two routes to prediction using ensembles of model instances.

⇒ **Advantages of average model**

- Only 1 evaluation/prediction
- Only 1 model instance to store
- 1 model instance = simple analytical model

Conclusions

- Small data sets can successfully be modelled using Ensemble based Machine Learning approaches.
- Ensemble averages can be replaced without loss of quality by an average model instance
- Average model instances are fast, minimal in storage and give access to simple analytical models, improving the interpretability.

Moral:

Do you feel lucky?

Acknowledgement

AMIBM has been made possible with the support of the Dutch Province of Limburg

References

[1] Danny E. P. Vanpoucke *et al*., J. Appl. Phys. 128, 054901 (2020), DOI: https://doi.org/10.1063/5.0012285
[2] AMADEUS software, https://github.com/DannyVanpoucke/Amadeus
[3] Patent WO2018/002055A1

Correspondence to:
Danny E.P. Vanpoucke
DannyVanpoucke@gmail.com
www.dannyvanpoucke.be
@DelocalizedD

Katrien V. Bernaerts
Katrien.Bernaerts@maastrichtuniversity.nl
T +3143 3882636

Aachen Maastricht Institute for Biobased Materials (AMIBM)
www.amibm.org

Maastricht University
An-Institut der RWTH Aachen University

Brightlands Chemelot Campus
Urmonderbaan 22
6167 RD Geleen, The Netherlands