HelmholtzZentrum münchen
German Research Center for Environmental Health

Improving evaluability of machine learning challenges by applying deterministic protocols

Manuel B. Huber

Institute of Health Economics and Health Care Management, Helmholtz Zentrum München, Germany

Background

Machine learning challenges (MLCs) lead to the development of very sophisticated prediction models. To evaluate the influence of data cleaning, feature engineering, model selection, hyperparameter tuning and ensemble learning on the outcome metric, is difficult because these tasks are interconnected. Moreover, data heterogeneity, user interaction and algorithm diversity cause inherent complexity. By applying principles of divide and conquer as well as approaches from deterministic sensitivity analysis, alternative competition protocols could improve transparency as well as evaluability of MLCs by dividing challenges into separated and researchable tasks. This poster outlines deterministic alternatives of current challenge designs and provides examples for implementation.

Methods

The "divide et impera" principle is mainly associated with politics but found its way into other study fields including molecular biology and algorithm design. The main goal of divide and conquer strategies is to divide very complex problems—in this case deciphering the influence of single tasks on the overall outcome of MLCs (figure 1)—into less complex problems that can be solved or "conquered". Additionally applying the approach of deterministic sensitivity analysis—variating one element and keeping remaining elements fixed—helps to assess single element influence on the respective outcome. Applying these methods may require specific environments that differ from prevailing MLC grounds.

Results

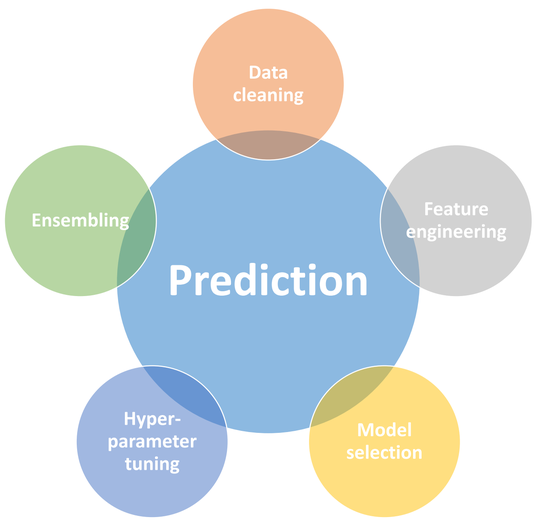

Figure 1 Simplified cycle of Interaction, prevailing challenge protocols

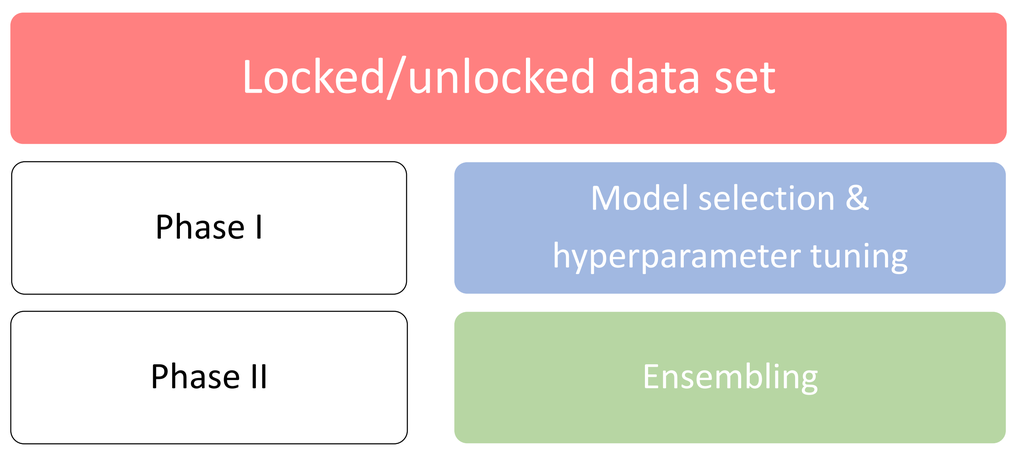

Figure 2 Exemplary deterministic protocol

To reduce uncertainty associated with data manipulation (figure 2), challenge providers could supply a clean data set that does not allow feature engineering. Single models or a list of models could be provided to guide participants. Hyperparameters could be pre-selected or a range of grid search parameters could be offered. Ensembling could be saved for a later phase of the challenge or it could be based on supplied models. Removing all uncertainty would undermine the purpose of respective challenges, since all participants would apply the same methods and would get the same results. Therefore, deterministic uncertainty reduction should be treated as a trade-off between exploring certain aspects of machine learning and finding the best model. Controlled environments (AWS, Azure, Google Cloud etc.) could facilitate the implementation of deterministic designs and are already used by MLC providers as well as participants.

Table 1 Deterministic alternatives for prevailing MLC tasks

Challenge task	Standard design	Deterministic alternative
Data cleaning	Free to choose	Not to be cleaned data set
Feature engineering	Free to choose	Not to be feature engineered data set
Model selection	Free to choose	Pre-selected model(s)
Hyperparameter tuning	Free to choose	• Pre-selected hyperparameters • Pre-defined grid-search range
Ensembling	Free to choose	• No ensembling • Phased ensembling • Ensembling based on pre-selected models

Conclusion

- Deterministic protocols increase evaluability of MLCs by separating interconnected tasks and by reducing uncertainty
- Task separation represents a trade-off between exploring aspects of machine learning and using participant freedom/creativity to find the best model
- Preferences of the challenge provider and respective consequences should be taken into account before restrictive designs are applied
- A stronger move towards controlled environments may be needed to foster the implementation of deterministic designs

Contact manuel.huber@helmholtz-muenchen.de

References
[1] L'Heureux, A., Grolinger, K., Elyamany, H.F. & Capretz, M. A. M. (2017) Machine Learning With Big 49 Data: Challenges and Approaches, IEEE Access 5: 7776-7797.
[2] Wolpert, D.H. & Macready, W.G. (1997) No free lunch theorems for optimization, IEEE Transactions on 51 Evolutionary Computation 1(1): 67-82.
[3] Bell, R.B. & Koren, Y. (2007) Lessons from the Netflix prize challenge, Acm Sigkdd Explorations Newsletter 53 9(2): 75-79.
[4] Rivera, M., Juan, C. & David, M.G. (2016) Replicating large genomes: Divide and conquer. Molecular cell 55 62(5): 756-765.
[5] Smith, D. R. (1985) Top-down synthesis of divide-and-conquer algorithms. Artificial Intelligence 27(1): 57 43-96.

HELMHOLTZ ASSOCIATION