Improving evaluability of machine learning challenges by applying deterministic protocols

Manuel Huber
Institute of Health Economics and Health Care Management
Helmholtz Zentrum Muenchen
Munich, BY 85764
manuel.huber@helmholtz-muenchen.de

Abstract

Machine learning competitions lead to the development of very sophisticated prediction models. To evaluate the influence of data cleaning, feature engineering, model selection, hyper-parameter 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 machine learning competitions by dividing challenges into separated and researchable tasks. This proposal outlines deterministic alternatives of current challenge protocols.

1 Introduction

The design of machine learning challenges largely remains the same. Providers supply data, participants clean data, engineer features, select machine learning models, tune hyper-parameters, combine models via ensemble learners and make predictions. There are no standard operating procedures and each participant has a sheer unlimited amount of possibilities [1]. Challenge tasks are interconnected and it's difficult to deceiver what task is responsible for what outcome. Therefore and due to the lack of approach transparency it's very difficult for researchers to translate user work and challenge outcomes into scientific findings. Moreover, since there is no free lunch in optimization [2], overall applicability as well as practicability of winning models remains unknown [3]. The aim of this proposal is to offer alternative competition designs that are based on the principles of divide and conquer as well as deterministic sensitivity analysis to improve transparency, reduce uncertainty and increase scientific evaluability of machine learning challenges.

2 Methods

The divide and conquer principle is applied in several study fields including molecular biology [4] and algorithm design [5]. The main goal of this approach is to divide very complex problems - in this case deciphering the influence of single tasks on the overall outcome of machine learning challenges (figure 1) - into less complex problems that can be solved or "conquered". Like done in deterministic sensitivity analysis, variating one element and keeping remaining elements fixed, helps to assess single element influence on outcome.

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

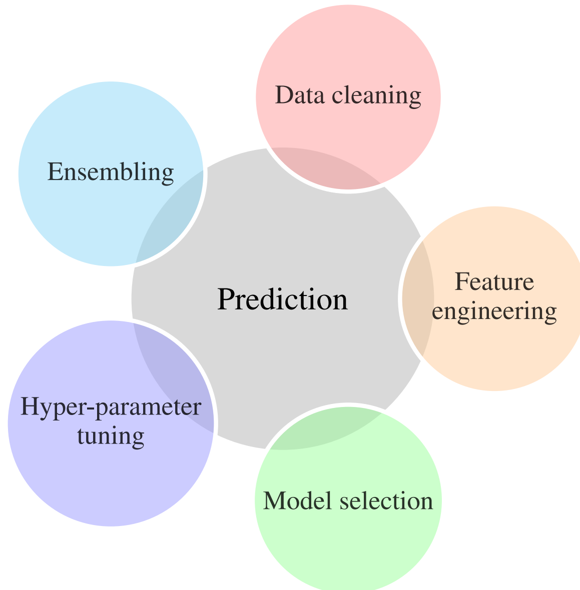

Figure 1: Simplified cycle of Interaction, prevailing challenge protocols

3 Results and discussion

Table 1 shows a simple overview of deterministic protocol alternatives. To reduce uncertainty associated with data manipulation, 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. Hyper-parameters could be pre-defined 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 pre-defined 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.

Table 1: Deterministic challenge task alternatives

Challenge task	Standard design	Deterministic alternative
Data cleaning	Free to choose	Already cleaned data set
Feature engineering	Free to choose	Already feature engineered data set
Model selection	Free to choose	Pre-defined model(s)
Hyper-parameter tuning	Free to choose	• Pre-defined hyper-parameters • Pre-defined grid-search
Ensembling	Free to choose	• No ensembling • Phased-in ensembling • Ensembling based on pre-defined models

4 Conclusion

Applying deterministic challenge protocols increases evaluability of machine learning challenges. Finding the right balance between task uncertainty and participant freedom should be based on the expectations of the challenge provider.

References

[1] L'Heureux, A., Grolinger, K., Elyamany, H.F. & Capretz, M. A. M. (2017) Machine Learning With Big 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 Evolutionary Computation* **1**(1): 67-82.

[3] Bell, R.B. & Koren, Y. (2007) Lessons from the Netflix prize challenge, *Acm Sigkdd Explorations Newsletter* **9**(2): 75-79.

[4] Rivera, M., Juan, C. & David, M.G. (2016) Replicating large genomes: Divide and conquer. *Molecular cell* **62**(5): 756-765.

[5] Smith, D. R. (1985) Top-down synthesis of divide-and-conquer algorithms. *Artificial Intelligence* **27**(1): 43-96.