technische universität dortmund

computer science 12

Master Thesis

Verification for the Concept Drift of Hyper-Parameter when Dataset Shift

Junjie Shi
Prof. Dr. Jian-Jia Chen
Otto-Hahn Str. 16
Technische Universität Dortmund
15.Sept.2021

The predictive performance of a machine learning model highly depends on the corresponding hyper-parameter setting. Hence, hyper-parameter tuning is often indispensable. Normally such tuning requires the dedicated machine learning model to be trained and evaluated on centralized data to obtain a performance estimate. Model-Based Optimization (MBO) also known as Bayesian optimization [2], is one state-of-the-art method for tuning hyper-parameters.

During the tuning process of hyper-parameters, we also assume the distribution of the training data set and test data set are consistent. However, in the real-world applications, the trained machine learning model can suffer from *dataset shift*, which means the data used to train a machine learning model is different from where the model operates. The *dataset shift* problem is also known as *concept drift* problem or dynamic optimization problem. Such a problem can significantly downgrade the performance of the pre-trained machine learning model.

To overcome the aforementioned dataset shift problem, the pre-trained machine learning model has to be re-trained using the latest collected data to obtain the stable performance. During the re-training process of the machine learning model, the pre-tuned hyper-parameter may also need to be re-trained. In [4], we assume the *dataset shift* can also result in the drift of the optimal hyper-parameter. Therefore, the corresponding hyper-parameter of the machine learning model has to be re-tuned and two approaches are proposed by assuming the tuning process is a black box optimization problem. However, the assumption that the optimal hyper-parameter also drifts by dataset shift has not been verified.

In this thesis, the student is supposed to figure out if the aforementioned assumption is correct or not, by using synthetic data set with artificially injected data shift(s). The details of the work flow are as follows:

1. Generate the data sets with shift by using the approaches presented in [3];
2. Tune two sets of optimal hyper-parameter for the dedicated data set, i.e., before and after the shift, by using BoTorch [1];
3. Cross-validate the two sets of tuned hyper-parameter for the entire data set to evaluate the variance of performance of the trained machine learning. That is using these two sets of tuned hyper-parameter to train two different machine learning models using the same training data set, and the performance is evaluated by using the same data set (with shift).

To make a more comprehensive conclusion, the student is supposed to evaluate different data sets, different shift methods, and different machine learning models.

Required Skills:

- Knowledgeable of Python programming
- Knowledgeable of basic machine learning model, e.g., RF, MLP, and CNN
- Knowledgeable of model based optimization or Bayesian optimization

Acquired Skills after the work:

- Knowledge of machine learning
- Knowledge of hyper-parameter tuning
- Knowledge of Open-Source Software

References

[1] M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems 33*, 2020.

[2] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[3] S. Maggio and L. Dreyfus-Schmidt. Ensembling shift detectors: an extensive empirical evaluation. *CoRR*, abs/2106.14608, 2021.

[4] J. Richter, J. Shi, J. Chen, J. Rahnenführer, and M. Lang. Model-based optimization with concept drifts. In C. A. C. Coello, editor, *GECCO '20: Genetic and Evolutionary Computation Conference, Cancún Mexico, July 8-12, 2020*, pages 877–885. ACM, 2020.