Prediction quality, domain adaptation and robustness of machine learning methods: a comparison

Payman Goodarzi, Andreas Schütze, and Tizian Schneider
Saarland University, Lab for Measurement Technology, 66123 Saarbrücken, Germany,
p.goodarzi, schuetze, and t.schneider @lmt.uni-saarland.de

Abstract

Domain or database shift causes performance degradation in machine learning models encountering real-life scenarios. However, it is not clear how and to what extent this degradation can be prevented, and which methods are more robust against that. In this paper, we compare a workflow based on conventional machine learning methods and deep neural networks for condition monitoring with emphasis on domain shift. It is shown that possible domain shifts can be detected using visualization techniques at feature level. Also, the conventional method shows superior results in the domain shift scenario compared with the deep learning model. Finally, domain adaptation is used to improve the models' performance.

1 Introduction

One of the important applications of machine learning (ML) methods is condition monitoring (CM) [1]. Industrial sensors and signals, e.g., pressure, vibration, and temperature measurements, are used to predict possible faults and upcoming failures. However, one very important issue in this field is the domain (or dataset) shift (DS) problem [2]. The performance of ML methods is highly dependent on the basic assumption that all data samples are drawn from the same distribution. However, in many real-life scenarios the mentioned assumption cannot be fulfilled, e.g., a model is trained in the lab and applied in the field, causing an out-of-distribution (OOD) problem. A reason for the DS problem is changes in the working conditions, e.g., for a ball bearing, the temperature or load variations may influence the recorded signals. A proper design-of-experiment would try to cover the possible variations of the working conditions, however due to practical limitations (time and cost of experiments) it is generally not possible to cover all possible variations and a model must generalize to all relevant conditions using a subset of the full data distribution. In this study we compare different ML algorithms in a scenario that suffers from the DS problem.

2 Dataset

A hydraulic system (HS) dataset from the Center for Mechatronics and Automation Technology (ZeMA gGmbH) is used in this study [3]. The ZeMA dataset contains recordings of 17 sensors and comprises various common faults of an HS. Four types of faults are simulated in this dataset, the main valve switching performance, internal pump leakage, accumulator pre-charge pressure reduction and cooler performance degradation. Visualizing extracted features using Principal Component Analysis (PCA) demonstrates that the cooler performance causes dominant shifts in the data distribution, **Figure 1**. The selected target in this example is detecting the valve switching state from 100% (fully functional) to 72% (barely working). To show that the system is robust against the cooler performance only data from two cooler states (20% and 100%) are used for training and 3% cooler state (near failure) is used as test data.

3 Algorithms

FESR: Conventional ML methods can be formulated as a stack of feature extraction (FE), feature selection (FS), and classification or regression methods. In this study, we used an open-source MATLAB toolbox [1] that performs a search to find the best combination of FE, FS, and classification or regression methods for a target task. One of the limitations of conventional ML is explicit feature engineering [4], but by using automatic hyperparameter (HP) tuning this framework resolved this drawback. Based on the results of the toolbox search a combination of statistical signal features [1], feature selection by Pearson correlation and partial least squares regression (PLSR) [5] is chosen.

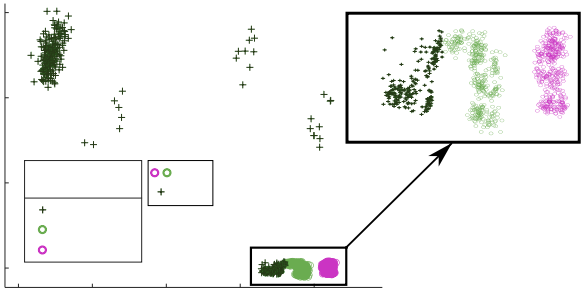

Figure 1 PCA plot of features from the ZeMA HS dataset. All data samples are colored by the cooler performance.

Deep learning methods: To find the best architecture for the task, a neural architecture search (NAS) that showed superior results and outperforms human-designed networks [6] was performed. Convolutional neural networks (CNN) are widely used in condition monitoring applications [7], therefore CNN is the network architecture that is

ISBN 978-3-8007-5835-7

selected for this study. The final model chosen for the defined scenario after training and validating about 500 different networks is a 9-layer CNN; the detailed parameters are described in [8].

Domain adaptation: Domain adaptation is one of the methods that is developed to remedy the DS problem. The main idea is to use unlabeled test data or a small subset of the labeled version, to adapt a model to a new data distribution (test data) for the same task. In this paper an offset calibration with the valve working at 100% and the cooler working at 3% is used to compensate the effect of the change of temperature introduced by the low cooler performance at 3% on the valve switching prediction.

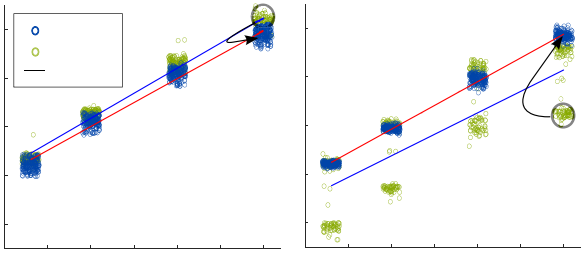

Figure 2 Prediction results for valve switching performance. For better visibility a jitter in x-direction is applied.

4 Experiments and results

The training, validation, and test results of the FESR model are reported in **Table 1**. The test error (RMSE) of the model is 2.5% which is also visible in **Figure 2** as a positive offset in the predicted values for the test data. In other words, changing the working condition of the HS results in an offset error of about 2.5%. This situation is similar to common cross-sensitivity problems in sensor systems.
The results of the CNN model are also reported in Table 1, the validation error (random cross-validation) is as low as 1.15% and does not show signs of overfitting towards the training set. However, the test error is 9.75% which is about five times larger than for the FESR model. The reason for this shift is that the model is unaware of the test data distribution and just fits on the training and validation data which have a different distribution. Two distinct groups are apparent in the prediction of the test data. These two groups are also evident at the feature level in Figure 1. The group which is closer to the training data in Figure 1, i.e., has smaller shifts, results in the predictions with smaller errors.

Model	Validation RMSE	Test RMSE	Test RMSE after offset calib.
FESR	1.53	2.45	1.58
CNN	1.15	9.74	3.34

Table 1 Error rates of the FESR and CNN models before and after offset calibration.

The described offset calibration is used to reduce the offset between the training and test data. Using this technique test errors decreased for both FESR and CNN models to 1.58% and 3.34%, respectively. Again, the FESR method shows a lower error rate. In fact, the resulting error is close to its validation RMSE, indicating that the DS problem is almost suppressed completely.

5 Conclusion

DS is a very common problem in real-life applications, especially in CM scenarios, and often is ignored even in widely used datasets [8]. By visualizing the data at different levels, it was shown how a DS in a dataset can affect the final prediction results degrading the model performance, but also how DS might be recognized using simple data visualization. In the examined scenario the CNN model achieved a lower validation error while the FESR method achieved much better results on the test data with DS, probably due to its lower complexity and therefore higher generalization capability. To remedy the DS problem recalibration was used as a simple domain adaptation technique. The recalibration approach improved the results for both models, however, the FESR model still achieved superior results.

6 References

[1] Schneider et al., “Industrial condition monitoring with smart sensors using automated feature extraction and selection,” IOP Meas. Sci. Technol., 2018, doi: 10.1088/1361-6501/aad1d4.

[2] Moreno-Torres et al., “A unifying view on dataset shift in classification,” Pattern Rec ognition, 2012, doi: 10.1016/j.patcog.2011.06.019.

[3] Schneider et al., “Condition monitoring of hydraulic systems Data Set at ZeMA,” Zenodo, 2018 , doi: 10.5281/zenodo.1323611.

[4] He et al., “AutoML: A survey of the state-of-the-art,” Knowledge-Based Systems 212, 2021, doi: 10.1016/j.knosys.2020.106622.

[5] Wold et al., “PLS-regression: a basic tool of chemometrics,” Chemometrics and Intelligent Laboratory Systems, 2001, doi: 10.1016/S0169-7439(01)00155-1.

[6] Zoph & Le, “Neural Architecture Search with Reinforcement Learning,” arXiv:1611.01578, 2016.

[7] Jiao et al. “A comprehensive review on convolutional neural network in machine fault diagnosis,” Neurocomputing 417, 2020, doi: 10.1016/j.neucom.2020.07.088.

[8] Goodarzi et al., “Comparison of diff erent ML methods concerning prediction quality, domain adaptation and robustness,” tm Technisches Messen, 2022, doi: 10.1515/teme-2021-0129.