Online Explainable Forecasting using Regions of Competence

Amal Saadallah[†], Matthias Jakobs[†]

[1] Lamarr Institute for Machine Learning and Artificial Intelligence, TU Dortmund University, Dortmund, Germany

Abstract

Several machine learning models have been applied to time series forecasting. However, it is generally acknowledged that none of these models is universally valid for every application and over time. This is due to the complex and changing nature of time series data that may involve non-stationary processes but can also be explained by the fact that different models have varying expected regions of expertise or so-called Regions of Competence (RoCs) over the time series. Therefore, adequate and adaptive online model selection is often required. In this work, we review online model selection works that exploit the notion of RoCs by summarizing their methods and highlighting their strengths as well as limitations. In particular, we present a taxonomy of these methods and show how they can be exploited for both single model selection and ensemble of models pruning. We additionally discuss how the RoCs can promote explainability. Finally, we suggest future directions to provide useful research guidance.

Keywords

Time Series Forecasting, Explainability, Model Selection, Ensembling

1. Introduction

Time series forecasting is considered a key step in making informed decisions in a wide range of applications [1, 2, 3]. Several Machine Learning (ML) models have been applied to the time series forecasting task [4, 5, 6]. While certain guidelines, such as task complexity or data size, can help select a suitable family of ML models for forecasting [7], it is generally acknowledged that no existing ML model is universally applicable to all forecasting problems. This observation aligns with Wolpert's No Free Lunch theorem [8], suggesting that no learning algorithm can be optimal for all learning tasks. In addition, ML models may demonstrate a time-dependent performance [9, 10]. This means their accuracy may not remain consistent over time due to the dynamic nature of time series data, which may involve non-stationary processes and, as a result, be subject to the so-called concept drift phenomenon [11]. Thus, it is clear that different forecasting models have different expected areas of expertise placed over different parts in the input time series. The part where a specific model outperforms certain candidate models is referred to as a Region of Competence (RoC) of that model [5, 12, 9].

Several works in the ML literature have used this notion of RoCs either implicitly [5, 13, 12] or explicitly [9, 10, 14, 15, 16] to perform online model selection for forecasting that copes with the time-evolving nature of time series and the fact that models have a certain expected level of competence in predicting a particular region in the time series. While some of these works have focused on the online selection of a single model [9, 15, 16], others have extended on the assumption that no single model is an expert all the time, and have proposed to adaptively select and combine multiple forecasting models into a single model using an ensemble technique [5, 17, 12, 10, 14]. The selection of individual models that make up the ensemble is referred to as ensemble pruning [18].

TempXAI@ECML-PKDD'24: Explainable AI for Time Series and Data Streams Tutorial-Workshop, Sep. 9th, 2024, Vilnius, Lithuania

[†] These authors contributed equally.

amal.saadallah@cs.tu-dortmund.de (A. Saadallah); matthias.jakobs@tu-dortmund.de (M. Jakobs)
0000-0003-2976-7574 (A. Saadallah); 0000-0003-4607-8957 (M. Jakobs)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

2. Regions of Competence Computation

Let X be a univariate time-series and let X_t be the value of that time-series at time t. Additionally, we denote with $X_{a:b}$ the subsequence entailing values from $t = a$ to $t = b$ with $a < b$. A Region of Competence (RoC) for a forecasting model $f \in \mathbb{P}$, where $\mathbb{P}$ is a set of given models, refers to a set of subsequence $X_{a:b}$ where f exhibits superior performance compared to all other models in $\mathbb{P}$.However, pinpointing these subsequences and determining their optimal length through empirical evaluations is computationally infeasible. Recent literature has leveraged machine learning to approximate these RoCs. These approaches fall into two main categories. The first category includes model-agnostic methods that compute RoCs independent of a family of forecasting models. Within this category, two primary machine learning paradigms have been used, namely pattern matching and meta-learning. The second category comprises model-specific methods tailored to a particular class of forecasting models, such as tree-based models or DNNs.

The main idea of pattern matching methods is to, using the current subsequence, find the closest subsequence in the training or validation data. The model exhibiting the best performance on these patterns is considered as the expert and this recent subsequence is treated as one of its RoCs. In [12], 1-Nearest Neighbors is used to determine the closest pattern. In [16], clustering is used to cluster the known datapoints first. Then, each model in $\mathbb{P}$ is evaluated on all subsequences from each cluster. The Mean Squared Error (MSE) for each model is computed and each cluster is assigned as Regions of Competence (RoC) of the model that minimizes the computed MSE. This ensures that each cluster is associated with an expert model that excels in predicting values based on the dominant pattern in the corresponding cluster.

In [17, 5], meta-learning is used to build models capable of modeling the competence of each candidate model across the input time series. Their meta-learning approach is composed of an arbitrating architecture [19] and a mixture of experts [20]. A meta-learner is created for each candidate model. While each model f_i is trained to model the future values of the time series, its meta-learning associate g_i is trained to model the error of f_i. The arbiter g_i then can make predictions regarding the error that will incur when predicting the future values of the time series. At test time, the candidate models f_i are weighted according to their expected degree of competence in $X_{t-p:t-1}$, estimated by the predictions of the meta-learners. These methods can be viewed as an implicit exploitation of the RoC concept as they do not result in some identified subsequences in the input time series but rather a modeling of the competence of the models using error estimates.

The works in this category mainly focus on CNN-based models, comprising 1D-convolutional layers with varied filter and kernel sizes [21, 10]. More recently, in [14], a hybrid architecture is proposed by concatenating the convolutional layers to a set of heterogeneous models that can be trained jointly using a gradient-based approach. However, these works use the same principle to compute the RoCs. They suggest a modified version of Grad-CAM[22], which utilizes spatial information preserved in convolutional layers to identify important regions of the input. To do so, they start by evaluating the Mean Squared Error on a specific validation time window X_{val}, denoted as ζ_{f_i}, for model f_i. The objective is to determine the significance of each time point in this window with respect to the computed error. The last feature maps layer of the CNN is utilized for this purpose. Importance weights α_{f_i} associated with ζ_{f_i} are computed for each activation unit u in each generic feature map A by calculating the gradient of ζ_{f_i} relative to A. Finally, a global average is computed over all units in A: $\alpha_{f_i} = \frac{1}{U} \sum_u \frac{\partial \zeta_{f_i}}{\partial A_u}$, where U is the total number of units in A. We use α_{f_i} to compute a weighted combination between all the feature maps for a given measured value of the error ζ_{f_i}. Since we are mainly interested in highlighting temporal features contributing most to ζ_{f_i}, ReLU is used to remove all the negative contributions by: $L_{f_i} = ReLU(\sum \alpha_{f_i} A)$. To identify the RoC in X_{val} that primarily contributed to ζ_{f_i}, $L_{f_i} \in \mathbb{R}^U$ is used. Note that U is chosen such that U is smaller than X_{val} size. Note that multiple time windows are created from X_{val} using a time-sliding approach to evaluate the performance of the same model on different windows and increase the number of computed RoCs. While the methods presented in the previous section proved successful in providing an explanation for the model selection and ensembling processes, the problem of opaque, non-interpretable base models still persists. Recent work [15] utilized tree-based models to make the model decision process more transparent in addition to having an interpretable model selection algorithm. The authors create a pool of Decision Trees, Random Forests, and Gradient Boosting Trees that are subsequently trained on X_{train}. Shapley values, a well-established Explainability method for feature attribution, are used to generate the explanations necessary for the creation of RoCs. Since computing Shapley values is in general NP-hard [23], the authors utilize TreeSHAP [24], which is an estimation method designed for tree-based models that is able to estimate Shapley values in polynomial time. Similar to [9], the loss of the prediction for each model is explained rather than the prediction itself.

3. Online Model Selection

At test time t, after computing the RoCs of the candidate models for selection, an online decision on model selection for forecasting the value of X at t has to be made. Note that the following can be applied to any future time instant $t + h, h \geq 0$. For simplicity of notations, assume $h = 0$. In this part, we focus on the works that explicitly compute the RoCs, i.e., result in identified subsequences within the input time series.

The models in $\mathbb{P}$ are devised such that they use the same p-lagged values of the time series as input. As a result, at time t, the input subsequence is $X_{t-p:t-1}, t \geq p$. To perform the selection, the distance of the input subsequence $X_{t-p:t-1}$ to the RoCs of each candidate model is measured [9, 12, 16, 15]. Since the length of each RoC can be different from p (i.e. length of $X_{t-p:t-1}$), Dynamic Time Wrapping (DTW) [25] is generally used to measure the similarity between $X_{t-p:t-1}$ and each RoC member. The model with a RoC member of smallest DTW distance to $X_{t-p:t-1}$ is selected for forecasting. In [16], the RoCs are smoothed using a moving-average process and filtered to keep only the RoCs with length p, and DTW is replaced by Euclidean Distance.

Recent works have extended the RoC concept for single model selection to ensemble pruning [16, 14, 10]. They based their reasoning on the expected error of the ensemble that can be decomposed at each time point into a weighted average error term of the individual composing models and an ambiguity term, which is simply the variance of the ensemble around the weighted mean of its composing models. The weighted error term reflects the ensemble *accuracy*, while the ambiguity term reflects the ensemble *diversity* [26, 27]. Since one model can have multiple RoCs, the first step is to select one representative RoC for each model based on the topological closeness to the current input subsequence $X_{t-p:t-1}$. In this manner, we avoid selecting the same model multiple times. This contributes to the ensemble *diversity* [16, 14]. In [10], *diversity* is promoted further by clustering the RoCs representatives so that selecting models with similar competencies or expertise is avoided. The second step consists of ranking the models using the distance of their representative RoCs to the current input subsequence $X_{t-p:t-1}$ to promote *accuracy*. The top-M closest models are selected to make up the ensemble. The top-M is arbitrarily set to a fixed number in [16]. The authors in [10] derived bounds for both error types to set the top-M number automatically.

4. Explainability Aspects

Most studies mentioned earlier emphasized the importance of transparency in models and learning decisions, including the selection of models [9, 10, 16, 14, 15]. They demonstrated that the computed Regions of Competence serve as a valuable tool for explaining why a specific forecast value is generated at a given time and for choosing a particular individual model or ensemble member within a specific time frame or interval. In essence, these RoCs act as an explanatory mechanism, shedding light on the decision-making process behind forecast outputs and model selections. More specific, they are part of the so-called exemplar-based or prototype-based explainability approaches [28, 29]. Thus, local explanations, i.e., why a certain model was chosen at time t can be given by visualy inspecting the

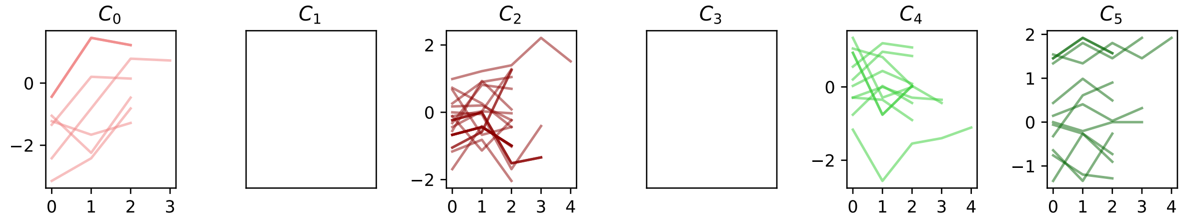

Figure 1: Visualization of RoCs for AbnormalHeartbeat data using the DNN-base approach for RoCs computation [9]. Notice that some RoCs can be empty, reflecting that this model did not outperform the anothers at any point during validation.

closest RoC member for each model. Global insights can be gained by visualizing the (clustered) RoCs for all models, as seen in Figure 1.

5. Conclusion

This work provides a unified view of several methods for online adaptive time series forecasting methods that are based on Regions of Competence. These regions not only enable state-of-the-art selection and ensemble performance but can also be used to gain insights into the selection and ensembling process on a local, as well as a global level.

One avenue of further research could be concerned with utilizing model-agnostic explainability methods such as KernelSHAP to enable heterogeneous model selection and ensembling. However, such a method may be limited by runtime, which would need to be improved to allow for concept drift adaption. Another direction of future work might be to improve overall explainability by limiting the model pool to small, transparent models. An open question remains if it is possible to select or ensemble these small models using Regions of Competence in a way that reaches state-of-the-art performance compared to incomprehensible methods such selecting from pools of Deep Neural Networks.

Acknowledgments

This research has been funded by the Federal Ministry of Education and Research of Germany and the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence.

References

[1] J. G. De Gooijer, R. J. Hyndman, 25 years of time series forecasting, International journal of forecasting 22 (2006) 443–473.

[2] A. Saadallah, L. Moreira-Matias, R. Sousa, J. Khiari, E. Jenelius, J. Gama, Bright-drift-aware demand predictions for taxi networks, IEEE Transactions on Knowledge and Data Engineering (2018).

[3] R. Godahewa, C. Bergmeir, G. I. Webb, R. J. Hyndman, P. Montero-Manso, Monash time series forecasting archive, in: Neural Information Processing Systems Track on Datasets and Benchmarks, 2021. Forthcoming.

[4] N. K. Ahmed, A. F. Atiya, N. E. Gayar, H. El-Shishiny, An empirical comparison of machine learning models for time series forecasting, Econometric reviews 29 (2010) 594–621.

[5] V. Cerqueira, L. Torgo, F. Pinto, C. Soares, Arbitrated ensemble for time series forecasting, in: Joint European conference on machine learning and knowledge discovery in databases, Springer, 2017, pp. 478–494.

[6] B. Lim, S. Zohren, Time-series forecasting with deep learning: a survey, Philosophical Transactions of the Royal Society A 379 (2021) 20200209.

[7] V. Cerqueira, L. Torgo, C. Soares, Machine learning vs statistical methods for time series forecasting: Size matters, arXiv preprint arXiv:1909.13316 (2019).

[8] D. H. Wolpert, The lack of a priori distinctions between learning algorithms, Neural computation 8 (1996) 1341–1390.

[9] A. Saadallah, M. Jakobs, K. Morik, Explainable online deep neural network selection using adaptive saliency maps for time series forecasting, in: N. Oliver, F. Pérez-Cruz, S. Kramer, J. Read, J. A. Lozano (Eds.), Machine Learning and Knowledge Discovery in Databases. Research Track, Springer International Publishing, Cham, 2021, pp. 404–420.

[10] A. Saadallah, M. Jakobs, K. Morik, Explainable online ensemble of deep neural network pruning for time series forecasting, Machine Learning 111 (2022).

[11] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, A. Bouchachia, A survey on concept drift adaptation, ACM computing surveys (CSUR) 46 (2014) 1–37.

[12] F. Priebe, Dynamic model selection for automated machine learning in time series (2019).

[13] V. Cerqueira, F. Pinto, L. Torgo, C. Soares, N. Moniz, Constructive aggregation and its application to forecasting with dynamic ensembles, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2018, pp. 620–636.

[14] A. Saadallah, M. Jakobs, Online deep hybrid ensemble learning for time series forecasting, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2023, pp. 156–171.

[15] M. Jakobs, A. Saadallah, Explainable Adaptive Tree-based Model Selection for Time-Series Forecasting, in: 2023 IEEE International Conference on Data Mining (ICDM), 2023, pp. 180–189. doi:10.1109/ICDM58522.2023.00027.

[16] A. Saadallah, Online explainable model selection for time series forecasting, in: 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2023, pp. 1–10.

[17] V. Cerqueira, L. Torgo, F. Pinto, C. Soares, Arbitrage of forecasting experts, Machine Learning (2018).

[18] G. Tsoumakas, I. Partalas, I. Vlahavas, An ensemble pruning primer, in: Applications of supervised and unsupervised ensemble methods, Springer, 2009, pp. 1–13.

[19] J. Ortega, M. Koppel, S. Argamon, Arbitrating among competing classifiers using learned referees, Knowledge and Information Systems 3 (2001) 470–490.

[20] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, Adaptive mixtures of local experts, Neural computation 3 (1991) 79–87.

[21] A. Saadallah, M. Tavakol, K. Morik, An actor-critic ensemble aggregation model for time-series forecasting, in: IEEE ICDE, 2021.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017.

[23] X. Deng, C. H. Papadimitriou, On the complexity of cooperative solution concepts, Mathematics of Operations Research 19 (1994) 257–266. arXiv:3690220.

[24] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence 2 (2020) 56–67. doi:10.1038/s42256-019-0138-9.

[25] D. J. Berndt, J. Clifford, Using dynamic time warping to find patterns in time series., in: KDD workshop, volume 10, 1994, pp. 359–370.

[26] A. Krogh, J. Vedelsby, Neural network ensembles, cross validation, and active learning, in: Advances in Neural Information Processing Systems, volume 7, MIT Press, 1995.

[27] G. Brown, J. L. Wyatt, P. Tiňo, Managing Diversity in Regression Ensembles, Journal of Machine Learning Research 6 (2005) 1621–1650.

[28] A. Patil, M. Patil, A Comprehensive Review on Explainable AI Techniques, Challenges, and Future Scope, in: V. E. Balas, V. B. Semwal, A. Khandare (Eds.), Intelligent Computing and Networking, Springer Nature, Singapore, 2023, pp. 517–529. doi:10.1007/978-981-99-3177-4_39.

[29] C. Molnar, Interpretable Machine Learning, 2020.