Meta-Learning with Neural Networks and Landmarking for Forecasting Model Selection

An Empirical Evaluation of Different Feature Sets Applied to Industry Data

Mirko Kück
BIBA – Bremer Institut für Produktion und Logistik GmbH
University of Bremen
Bremen, Germany
e-mail: kue@biba.uni-bremen.de

Sven F. Crone
Research Centre for Forecasting
Lancaster University Management School
Lancaster, United Kingdom
e-mail: s.crone@lancaster.ac.uk

Michael Freitag
University of Bremen, Faculty of Production Engineering
BIBA – Bremer Institut für Produktion und Logistik GmbH, University of Bremen
Bremen, Germany
e-mail: fre@biba.uni-bremen.de

***Abstract*— Although artificial neural networks are occasionally used in forecasting future sales for manufacturing in industry, the majority of algorithms applied today are univariate statistical time series methods for level, seasonal, trend or trend-seasonal patterns. With different statistical methods created for different time series patterns, large scale applications on 10,000s of times series require automatic method selection, often done manually by human experts based on various time series characteristics, or automatically using error metrics of past performance. However, the task of selecting adequate forecasting methods can also be viewed as a supervised learning problem. For instance, a neural network can be trained as a meta-learner relating characteristic time series features to the ex post accuracy of forecasting methods for each time series. Past research has proposed different sets of time series features for meta-learning including simple statistical or information-theoretic as well as model-based features, but have neglected the use of past forecast errors. This paper studies the predictive accuracy of using different feature sets for a neural network meta-learner selecting between four statistical forecasting models, introducing error-based features (landmarkers) and statistical tests as time series meta-features. A large-scale empirical study on NN3 industry data shows promising results of including error-based feature sets in meta-learning for selecting time series forecasting models.**

Keywords—meta-learning; time series forecasting, meta-features, industry data

I. INTRODUCTION

Demand planning is an important task for manufacturing companies because it is the main basis for all following steps of production planning [1]. Typically, the future demand per stock-keeping-unit has to be forecasted for a large number of items based on univariate time series of past customer orders. For this purpose, various forecasting methods can be applied, which all have different strengths and weaknesses. While statistical methods like exponential smoothing variants performed best in some studies [2], methods of computational intelligence (CI), such as neural networks, have shown promising performance in more recent studies [3]. Unfortunately, the no-free-lunch theorem [4] states that there is no universally best method outperforming all others for a broad problem domain. This theoretical statement was confirmed in several empirical studies [2], [3], [5], [6]. Hence, a suitable model selection approach is needed in order to find appropriate forecasting models for each case of different time series evolution. This problem was formulated as the algorithm-selection problem [7]. In this context, meta-learning is a promising approach for model selection [8], [9], [10], [11] describing how to deploy knowledge from past tasks in order to select predictive models for new tasks with certain characteristics. While meta-learning is most often applied for selecting classification algorithms [12], [13], [14], [15], fewer applications exist for selecting models for univariate [16], [17], [18], [19] or multivariate time series forecasting [20], [21]. Moreover, in the area of time series forecasting, all different studies deployed different meta-feature sets. In particular, the meta-learning approaches neglected error-based features as normally used in model selection for time series forecasting. This raises the questions whether error-based features are appropriate for meta-learning and which features are most applicable for the selection of time series forecasting models.

The paper at hand studies the impact of different feature sets for a meta-learning approach conducting a neural network as meta-learner to select one of four forecasting models often used in industry. For this purpose, the impact of seven different feature sets on the performance of model selection is studied. These are the feature sets used in a meta-learning approach by Wang et al. [17], features of recurrent quantification analysis [22] used in a meta-learning approach by Scholz-Reiter et al. [19], several statistical measures as well as four different sets

based on training or validation errors of the four forecasting models. The impact of the different feature sets is evaluated within an empirical study conducted on the univariate industry time series of the NN3-competition [3]. The remainder of this paper is structured as follows. Section II describes related work in the context of meta-learning for selecting time series forecasting models. Section III illustrates the process of meta-learning for model selection as well as the different sets of meta-features considered in this paper. Sections IV and V describe the design and the results of an empirical study comparing the impact of different meta-feature sets on model selection performance. The paper closes with a conclusion and an outlook on future research directions.

II. Related Work

In order to obtain sophisticated forecasts of time series, a suitable forecasting model has to be selected. Traditionally, experts inspect given time series and select appropriate models according to their human judgment. However, since large numbers of time series have to be forecasted in many practical applications, automatic model selection approaches are required. An often applied approach is calculating the forecast error of a model on a given training or validation set and choosing the model with the lowest error [23]. Another approach, which has achieved popularity in the last years, is linking the appropriateness of forecasting models to characteristic time series features. On this account, expert systems as well as data-driven approaches exist [17]. Among different other approaches, the signature work in expert systems for time series forecasting model selection is called rule-based forecasting (RBF) [5]. For this approach, 99 rules were derived from experts and used to weight four simple forecasting models. The rules based on 18 time series features. Since these features had to be manually identified by human experts, different applications were difficult to compare. Subsequently, RBF was modified to consider 28 features and to reduce the amount of needed expert judgment by automating the extraction of some features [24]. However, a number of features still had to be identified manually. In order to automate the model selection process, data-driven approaches were studied in the recent years [16], [17], [18], [19], [25], [26], [27], [28]. These methods achieve an automated selection of a forecasting model by training machine learning methods based on automatically extracted time series features and the appropriateness of different forecasting models for the regarded time series. In the machine learning community, this approach is known as 'meta-learning'. In the context of time series forecasting, the phrase meta-learning was first used in [16]. The meta-learning approaches used different configurations regarding considered features, forecasting models as well as learning methods to induce a meta-learner. While some approaches considered statistical methods, such as discriminant analysis or regression approaches to induce a meta-learner [19], [25], [27], others used decision trees and computational intelligence methods [16], [17], [18], [26], [28], such as neural networks. To characterize given time series by features, some approaches incorporated small sets of 6-13 features [16], [17], [26], while others applied larger sets of 25-38 features [18], [19], [25], [27], [28]. All existing meta-learning approaches used different feature sets consisting of simple statistical or information-theoretic as well as model-based features. However, the approach of landmarking [29] has been neglected in meta-learning for time series forecasting so far. This approach calculates the performance of simple forecasting models on the given time series and generates meta-features based on forecast errors. To analyze the impact of different features and in particular error-based features on model selection performance, this paper applies a neural network approach to meta-learning linking different feature sets to the performance of four exponential smoothing models.

III. Meta-Learning for Model Selection

A. Model Selection Problem

According to the no-free-lunch theorem [4], no universally best model exists to forecast data of a broad problem domain. Hence, suitable models have to be selected for each problem. This context was originally formulated as the algorithm selection problem (ASP) [7]. In analogy to the definitions of Rice [7] and Smith-Miles [10], we now define the model selection problem considered in this paper:

For a given time series $\mathbf{y}_{\text{Train}} = \{y_1, \ldots, y_{N_{\text{Train}}}\} \in \mathbf{Y}_{\text{Train}}$ with features $f(\mathbf{y}_{\text{Train}}) \in \mathbf{F}$, find the model selection mapping $S(f(\mathbf{y}_{\text{Train}}))$ into the forecasting model space $\mathbf{M}$, such that the selected forecasting model $m \in \mathbf{M}$ generates forecasts $m(\mathbf{y}) = \hat{\mathbf{y}}$ with minimal error (e.g. average symmetric mean absolute percentage error in (2)) on a holdout set of the time series $\mathbf{y}_{\text{Test}} = \{y_{N_{\text{Train}}+1}, \ldots, y_{N_{\text{Train}}+N_{\text{Test}}}\} \in \mathbf{Y}_{\text{Test}}$.

B. Meta-Learning for Model Selection

In order to solve the above defined model selection problem for a given time series, this paper applies a meta-learning approach. Fig. 1 illustrates the two processes of building a meta-learner based on a training set of given time series and deploying the meta-learner for selecting an appropriate forecasting model for a new time series which does not belong to the training set. Characteristic measures are computed for a given training set of time series and stored as input features (meta-features) in a knowledge base. Besides, the time series are forecasted by different forecasting models and the best model for each time series in terms of lowest forecast error is stored as output label in the knowledge base. Thus, the knowledge base comprises the features characterizing the given time series as well as the most appropriate forecasting model for each time series of the training set as label. Subsequently, the knowledge base is used to build a classifier (meta-learner) by supervised learning. In this paper, a neural network approach is applied to build the meta-learner in a multiclass classification problem with four exponential smoothing models as labels. Once this meta-learner is trained, it can be deployed to select an appropriate forecasting model for new time series. For this purpose, the same characteristic features as in the training process are also computed for the new time series. Based on these meta-features, the trained meta-learner predicts a class label representing the expected best forecasting model to select for the new time series. The concrete configuration of the meta-learning approach applied in the empirical study of this paper is specified within the experimental design in Section IV.

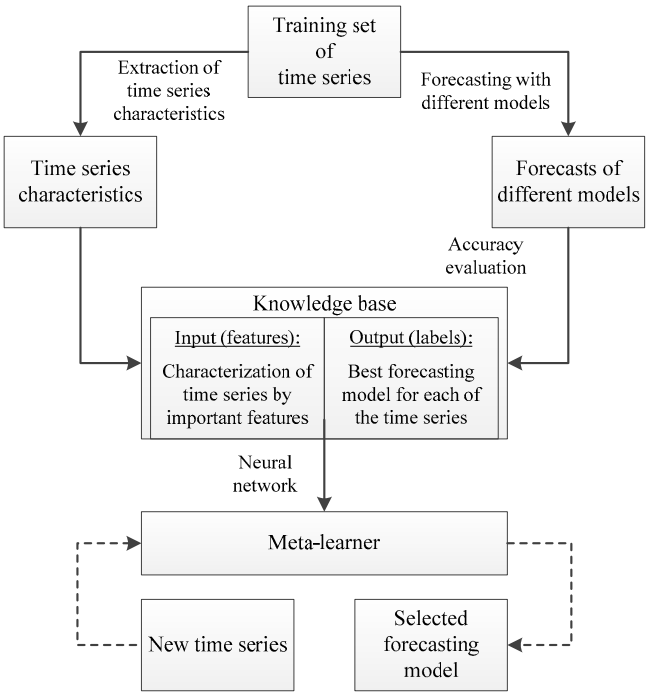

Fig. 1. Training and deployment of a meta-learner for model selection

In order to assure applicability of the meta-learning system, it must be considered if different features may be incomputable (NaN-values) for certain time series. For instance, this can be due to mathematical operations which cannot be conducted because of too short time series or for other reasons. To deal with this problem, we decided to replace each NaN-value of a feature by the mean of this feature over all considered time series if the proportion of NaN-values for this feature is at most 30%. If it is higher than 30% the feature is removed from the set and not considered for any time series in the empirical study. In addition to the modification regarding incomputable values, all features are normalized to the interval [0,1], where the smallest value of a specific feature over all considered time series is set to 0 and the highest value is set to 1. In order to study the impact of different features on the model selection performance, this paper applies seven different feature sets, which are described in the next two subsections. In this regard, only the features actually used in the empirical study are described. The features which have been removed because of a too high proportion of NaN-values are not detailed.

C. Established Meta-Feature Sets

In order characterize a given dataset by meta-features, three different approaches can be used, namely simple statistical or information-theoretic features, model-based features or landmarking [9]. Simple statistical or information-theoretic features are computed directly on the dataset. For example, this can be general information about the dataset, such as number of points, or descriptive statistics, such as mean or skewness of the data. Model-based features characterize the data indirectly by inducing a model from the data and using model characteristics as meta-features. For instance, a model-based feature can be the number of nodes of a decision tree induced from the data. The third approach of data characterization is landmarking [29], which is characterized in the next subsection. The remainder of this subsection describes simple statistical or information-theoretic features as well as model-based features, which can be regarded as established measures for meta-learning to select time series forecasting models. Three sets including different measures of these two approaches are specified in the following.

Wang et al. [17] proposed a set of 13 common metrics to quantify global characteristics of time series. These measures were also considered by Scholz-Reiter et al. [19]. Nine characteristics are measured on the raw time series. These are trend, seasonality, self-similarity, chaos, periodicity, serial correlation, nonlinearity, skewness, and kurtosis. Furthermore, four characteristics are measured on the remaining time series after detrending and deseasonalizing (dt-ds). These are serial correlation (dt-ds), nonlinearity (dt-ds), skewness (dt-ds), and kurtosis (dt-ds).

Marwan et al. [22] proposed time series complexity measures of recurrence quantification analysis (RQA), which were applied as meta-features by Scholz-Reiter et al. [19]. These measures base on the theory of phase space reconstruction and nonlinear dynamics [31], [32]. In order to reconstruct the dynamical properties of an underlying system, the method of delay coordinate embedding is applied [31]. This method maps the time series into a so-called phase space. The dynamical properties of the whole dynamical system can be reconstructed if two parameters are chosen appropriately, namely the embedding dimension m and the delay time τ [33], [34]. In general, for performing an RQA, it is recommended to compute m by the FNN-algorithm and τ as the first minimum of the average mutual information (AMI) [22], [32]. However, for specific time series, some measures of RQA may be incalculable, due to different reasons. For instance, RQA was initially introduced to characterize longer time series than the series considered in this paper. Hence, to obtain comparable characteristics, our set of features consists of twelve measures computed for $(m,\tau)=(1,1)$. These are recurrence rate (1,1), determinism (1,1), averaged diagonal length (1,1), length of the longest diagonal line (1,1), entropy of diagonal length (1,1), laminarity (1,1), trapping time (1,1), length of the longest vertical line (1,1), recurrence time of 1st type (1,1), recurrence time of 2nd type (1,1), recurrence period density entropy (1,1), and transitivity (1,1). Additionally, three characteristics are measured after calculating the embedding dimension m by the FNN-algorithm and the delay time τ as the first minimum of AMI. These are recurrence rate (FNN,AMI), length of the longest diagonal line (FNN,AMI), and length of the longest vertical line (FNN,AMI). The set of 17 considered features of RQA is completed by the computed values for m and τ.

The third feature set (Stat) consists of commonly used measures describing statistical data characteristics as well as test statistics of several statistical tests. In total, this feature set consists of the following 20 measures: length of the time series, minimum, maximum, mean, median, lower quartile, upper quartile, range, standard deviation, variance, coefficient of variation, linear coefficient trend test statistic, Mann-Kendall rank correlation coefficient, Spearman's Rho test statistic, Cox-Stuart location test statistic, Cox-Stuart dispersion test statistic,

Noether's cyclical trend test statistic, Chi-square test statistic, Kruskal-Wallis test statistic, and F-test statistic.

D. Error-Based Meta-Features (Landmarkers)

While past studies about meta-learning for selecting time series forecasting models mostly used simple statistical or information-theoretic meta-features as well as model-based meta-features [16], [17], [18], [19], [27], landmarkers have been neglected so far. Just like model-based features, landmarkers characterize the data indirectly after building a model on the data [9], [29]. In contrast to model-based features, landmarkers are not model characteristics but performance measures of the built model. In this approach, the forecasting models which can be selected as labels in the meta-learning process or simpler versions of these models are used to forecast subsets of the dataset and the forecast errors can be used as meta-features. On this account, the available forecasting models are trained on a training set of a given time series and the forecasting accuracy is measured as the average symmetric mean absolute percentage error (mean RO-sMAPE) of all forecasts until horizon h from different origins $i,\ldots,i+I$-1 on a holdout set. For this purpose, the symmetric mean absolute percentage error (sMAPE) calculated for the actuals y_t and the forecasts $\hat{y}_t$ of horizon h from origin i is defined as

$$\text{sMAPE}_i^h(\mathbf{y},\hat{\mathbf{y}}) = \frac{1}{h}\sum_{t=i+1}^{i+h}\frac{|y_t-\hat{y}_t|}{(|y_t|+|\hat{y}_t|)\cdot 0.5}. \quad (1)$$

This is a comparatively unbiased, scale independent error measure which allows comparing accuracy across multiple time series of different levels. The sMAPE is defined for the forecasts from one fixed forecasting origin i. However, forecasts calculated from a single origin can be corrupted by occurrences unique to this origin. Hence, a rolling-origin evaluation is applied [30]. For this purpose, all sMAPEs across horizons 1,…,h are calculated from multiple forecasting origins $i,\ldots,i+I$-1 and all errors are averaged to form the rolling-origin sMAPE (RO-sMAPE), which is used as performance criterion:

$$\text{RO-sMAPE}_{i,i+I-1}^h(\mathbf{y},\hat{\mathbf{y}}) = \frac{1}{I}\sum_{k=i}^{i+I-1}\text{sMAPE}_k^h(\mathbf{y},\hat{\mathbf{y}}). \quad (2)$$

In this paper, four sets of landmarkers are considered. For this purpose, the training set of time series is divided into a training set and a validation set and error measures are computed on both sets. The first set of landmarkers considers the RO-sMAPEs of the four individual forecasting models on the training set. These features are the training errors (TraE) of forecasting models 1, 2, 3, and 4 respectively. Moreover, the ranks of the models regarding the training errors are considered in the second feature set named TBR (Training error Binary Ranking). For this purpose, a binary configuration is applied. This leads to the following 16 binary meta-features: forecasting models 1, 2, 3, and 4 have lowest, second-lowest, third-lowest or highest training error respectively. Analogous to the meta-features regarding the errors on the training set, errors on the validation set are considered. Here, the validation errors of the four forecasting models are considered as the feature set ValE. Furthermore, the 16 binary meta-features relating to the ranks of the models regarding the validation errors constitute the feature set VBR (Validation error Binary Ranking). In total, 40 error-based meta-features are considered.

IV. Experimental Design

A. Experimental Setup

We conduct an empirical study to assess the efficacy of different meta-feature sets, including established meta-features and a novel type of error-based landmarking features on the model selection performance of a neural network based meta-learning approach. The empirical evaluation entails an objective experimental design and a representative benchmark dataset of the NN3 competition [3], as outlined below.

In order to study the impact of the different sets of meta-features on the model selection performance of the applied meta-learning system, we evaluate combinations of seven archetypical meta-feature sets from literature (as specified in sections III.C and III.D). These include the feature set of common metrics to quantify global characteristics of time series (Wang), features of recurrence quantification analysis (RQA), commonly used measures describing statistical data characteristics as well as test statistics of several statistical tests (Stat). Moreover, we introduce landmarkers based upon the RO-sMAPEs for training and validation errors (TraE, ValE) and respective ranks (TBR, VBR). In order to assess the accuracy derived not only from single meta-feature sets but also combinations thereof, we combine all meta-feature sets successively, creating a total of 127 meta-feature sets. These feature sets form the inputs to the meta-learner to choose an appropriate forecasting model for each time series at hand.

As base-learners, the meta-learner can select between the four class labels which represent four statistical forecasting algorithms of single (M1), seasonal (M2), seasonal-trended (M3), and trended exponential smoothing (M4) models to conduct the forecast.

As the meta-learner, we apply a single-hidden-layer Multilayer Perceptron (MLP), a feed-forward neural network using the 'nnet' function of the 'nnet' package in R [35]. This is an automatic function, which uses a quasi-Newton method (Broyden-Fletcher–Goldfarb-Shanno algorithm, BFGS), for optimization purposes. The neural network parameters were tuned within a leave-one-out cross validation of 200 iterations at a maximum with a number of units in the hidden layer between 1 and 10 and weight decay of 0, 0.01, 0.1 or 1 using the 'train' function of the 'caret' package in R [36]. The model selection problem is represented as a multiclass classification problem with the goal to predict the class with lowest RO-sMAPE on the holdout set of a given time series.

B. Empirical Dataset

We assess the model selection performance of different meta-learning approaches on the dataset of 111 empirical time series of industry sales from the NN3-competition [3]. Created as a subset of renowned M3-competition [2], it has been

established as a valid, reliable, and representative benchmark dataset for computational intelligence (CI) methods in multiple empirical studies. The dataset consists of a representative set of 50 long (L) and 50 short (S) monthly industry time series, with the shortest time series 68 months and the longest 144 months long. Furthermore, the long and short series contain an equal balance of 50 seasonal (S) and 50 non-seasonal patterns (NS), plus 11 time series exhibiting particular complexities in their time series patterns. The balanced design of the data allows an analysis of the model selection performance of different meta-learning approaches across typical industrial data conditions, testing their relative performance across long vs. short and seasonal vs. non-seasonal and complex data patterns.

For the empirical evaluation, the dataset of 111 time series was randomly but balanced divided into a training set of 78 time series (≈70%) and a test set containing the remaining 33 time series (≈30%). The training set was used to induce meta-learners. Subsequently, the model selection performance of the induced meta-learners was assessed on the test set. Furthermore, each of the 111 time series was divided into a training set (50-126 points) and a holdout set (18 points). For each time series, the training set was used to select a forecasting model and this model was applied to forecast the holdout set. No parameter estimation or model selection was conducted on the holdout data.

C. Evaluation of Accuracy

In order to evaluate accuracy, different measures were calculated. As the main accuracy measure, the mean RO-sMAPE (2) was regarded. For each holdout set of 18 points, 12-step-ahead forecasts were calculated from 6 different origins leading to 72 forecasts per time series, whose average is calculated by the mean RO-sMAPE. In addition, the classification accuracy and the average rank of the selected model were considered as further accuracy measures.

D. Benchmark Selections and Algorithms

We compare the accuracy of the proposed meta-learning approach of selecting between the four aforementioned exponential smoothing models to established approaches in forecasting model selection and simple benchmark algorithms.

For benchmarks in model selection in accordance with [23], we assess the approach of aggregate model selection (AggSel), which always applies the same statistical forecasting model on all time series. AggSel is estimated for each of the four underlying base models, resulting in four benchmarks of AggSel M1, M2, M3, and M4 for level, seasonal, trend-seasonal, and trend exponential smoothing respectively. In addition, different strategies of individual model selection are assessed, including the selection of the forecasting model with lowest error on the training set (TraSel) or the validation set (ValSel) [23], which is widely applied as best practice in industry.

In addition, knowing the final out-of-sample accuracies of each forecasting model on the holdout sets, we provide a ground truth of an optimal model selection (Opt) which provides a lower bound of the error that could have been achieved if always the best of the four exponential smoothing models had been chosen for each of the time series.

TABLE I. RO-SMAPE PERFORMANCE OF THE 10 BEST META-LEARNING METHODS AND THE BEST META-LEARNING METHOD WITHOUT USING ERROR-BASED FEATURES COMPARED TO INDIVIDUAL AND AGGREGATE MODEL SELECTION

	RO-sMAPE				
	Mean	Min	Max	Std	average rank
Opt	0.1118	0.0240	0.4158	0.0998	1.0000
nnet (TBR,ValE)	0.1208	0.0240	0.4747	0.1074	1.7273
nnet (TraE,ValE)	0.1217	0.0242	0.4747	0.1102	1.7879
nnet (ValE,VBR)	0.1222	0.0241	0.4747	0.1082	1.7879
ValSel	0.1229	0.0242	0.4747	0.1089	1.9091
nnet (VBR)	0.1238	0.0271	0.4747	0.1070	1.9394
nnet (Wang,Stat,TBR,VBR)	0.1272	0.0271	0.4857	0.1175	1.9091
nnet (ValE)	0.1285	0.0240	0.4857	0.1175	1.6970
nnet (TBR,ValE,VBR)	0.1287	0.0241	0.4747	0.1161	1.7576
nnet (RQA,TraE,ValE)	0.1294	0.0242	0.4857	0.1104	2.1515
nnet (Wang,RQA,TraE,TBR,VBR)	0.1295	0.0241	0.4747	0.1065	2.1515
nnet (Wang,TBR,VBR)	0.1300	0.0271	0.4651	0.1111	1.9394
TraSel	0.1325	0.0242	0.4857	0.1163	2.0606
nnet (Wang,RQA)	0.1374	0.0242	0.4651	0.1134	2.0909
AggSel M2	0.1409	0.0242	0.4857	0.1262	2.3030
AggSel M1	0.1445	0.0379	0.4651	0.1092	2.3939
AggSel M3	0.1546	0.0240	0.4747	0.1415	2.3939
AggSel M4	0.1633	0.0261	0.7666	0.1421	2.9091
Naive Benchmark	0.1589	0.0121	0.4763	0.1104	-
Average RO-sMAPE per feature set group	without error features	with error features	difference in sMAPE	difference in sMAPE in %	
Wang	0.1458	0.1328	-0.0130	-8.92%	
RQA	0.1422	0.1359	-0.0063	-4.43%	
Stat	0.1409	0.1360	-0.0049	-3.48%	
Wang RQA	0.1374	0.1347	-0.0027	-1.97%	
Wang Stat	0.1376	0.1344	-0.0032	-2.33%	
RQA Stat	0.1480	0.1373	-0.0107	-7.23%	
Wang RQA Stat	0.1420	0.1367	-0.0053	-3.73%	
Average over all feature sets	0.1420	0.1354	-0.0066	-5.00%	

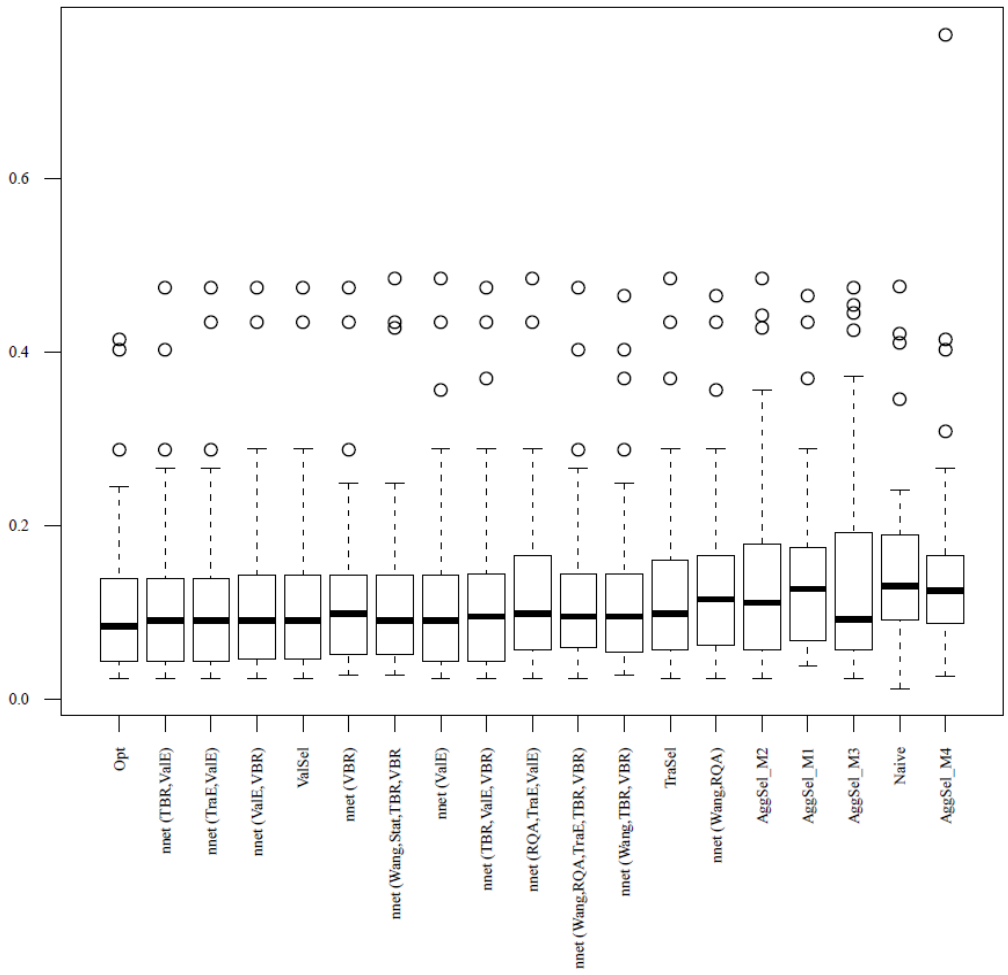

Fig. 2. Boxplots of the distributions of the RO-sMAPEs of the 10 best meta-learning methods, the best meta-learning method without using error-based features, and several benchmark methods sorted by the mean RO-sMAPEs

Furthermore, to provide an assessment of the overall accuracy of the meta-learning approach as well as conventional model selection approaches, we estimate the established benchmark of a random walk forecast (Naïve), always using the last observed actual value y_t as the horizon-forecast for the next period in time y_{t+h}.

V. Experimental Results

The results of the empirical study are illustrated within TABLES I.-II. and Fig. 2. Since it is not feasible to show the performance of all 127 meta-learners with all different combinations of feature sets, we focus on the main results. The upper part of TABLE I. shows the mean, minimum, maximum, and standard deviation of the RO-sMAPEs as well as the average ranks of the 10 best meta-learning methods, the best meta-learning method without using error features and the applied benchmark methods for model selection. It can be seen that all meta-learning methods performed better than all aggregate selection methods and the naïve random walk benchmark. The ValSel benchmark method, which selects the best model regarding the validation error, performs comparably but three meta-learning methods are closer to the optimal forecasts (Opt) which could have been reached if always the best of the four forecasting models had been selected. It can be conducted that the error features (landmarkers) have a considerable impact on model selection performance because the best meta-learning method without using error features (nnet (Wang,RQA)) performs noticeably worse. Within the range of meta-learning methods with different feature set combinations, the best method without using error features is ranked 91st of 127. The 90 better meta-learning methods all consider error features. The lower part of TABLE I. further emphasizes the positive impact of using error-based features. It shows that the average of all mean RO-sMAPEs of all combinations of feature sets with error features is always smaller than for the cases without using error features.

Fig. 2 illustrates the distributions of the RO-sMAPEs by box plots sorted by the mean RO-sMAPEs. It shall be noted that the naïve random walk, which performs as a bad benchmark on average, does have a smaller minimum value of the RO-sMAPEs due to the fact that only the four exponential smoothing models were considered for model selection. This shows that a better model selection performance (smaller value of Opt) could have been achieved if more forecasting models had been considered within the model selection process.

TABLE II. compares the RO-sMAPEs as well as the classification accuracies of meta-learning with and without using error features in detail. In this context, the classification accuracy is defined as the proportion of cases in which the meta-learner selected the true best forecasting model. For every combination of non-error feature sets (Wang, RQA, Stat), the table indicates whether an additional consideration of a single or a combination of different error feature sets (TraE, TBR, ValE, VBR) leads to an improvement in forecasting accuracy (upper part) or classification accuracy (lower part). Improvements in forecasting accuracy regarding the RO-sMAPE are indicated by a negative deviation from the approaches without considering error features since a lower error implies higher accuracy. In contrast, improvements in classification accuracy are indicated by a positive deviation from the approaches without considering error features. The cases in which considering error features lead to improved accuracy are shown in bold. The results clearly indicate that including any error feature set into the meta-learning process almost ever leads to an improvement regarding the mean RO-sMAPE as well as the classification accuracy. Moreover, the average classification accuracy values of adding TraE, TBR, ValE, VBR or the combinations respectively to the approaches without error features are always better than without using error features. For the case of forecasting accuracy regarding the RO-sMAPE, these average values lead to improvements for all combinations of error feature sets but the single TraE set. All in all, landmarkers show great potential to be considered in meta-learning for time series model selection.

VI. Conclusion and Outlook

This paper applied a meta-learning approach to select time series forecasting models. A neural network was used as meta-learner linking time series features to the forecasting accuracies of four exponential smoothing models. Within an empirical study on the univariate industry time series of the NN3-competition, the impact of different meta-feature sets on the model selection performance was evaluated. The empirical results show that the meta-learning approach achieves a better model selection performance than several benchmark methods, namely a naïve random walk, aggregate selection of one of the four exponential smoothing models as well as selecting the best model regarding training or validation errors.

TABLE II. Forecasting Accuracy in Terms of RO-sMAPE and Classification Accuracy of Meta-Learning Methods with and without Using Error Features

RO-sMAPE	no error features	TraE	TBR	ValE	VBR	TraE TBR	TraE ValE	TraE VBR	TBR ValE	TBR VBR	ValE VBR	TraE TBR ValE	TraE TBR VBR	TraE ValE VBR	TBR ValE VBR	TraE TBR ValE VBR
Wang	0.1458	0.1378	0.1369	0.1324	0.1326	0.1319	0.1356	0.1322	0.1331	0.1300	0.1322	0.1300	0.1314	0.1323	0.1318	0.1314
RQA	0.1422	0.1452	0.1353	0.1357	0.1427	0.1362	0.1294	0.1311	0.1360	0.1335	0.1311	0.1377	0.1404	0.1311	0.1345	0.1381
Stat	0.1409	0.1387	0.1359	0.1393	0.1328	0.1327	0.1460	0.1324	0.1489	0.1330	0.1328	0.1362	0.1317	0.1324	0.1329	0.1340
Wang RQA	0.1374	0.1433	0.1421	0.1341	0.1326	0.1384	0.1362	0.1322	0.1385	0.1311	0.1322	0.1349	0.1295	0.1322	0.1310	0.1316
Wang Stat	0.1376	0.1389	0.1321	0.1332	0.1328	0.1404	0.1341	0.1327	0.1467	0.1272	0.1328	0.1331	0.1352	0.1327	0.1324	0.1319
RQA Stat	0.1480	0.1483	0.1453	0.1347	0.1327	0.1325	0.1332	0.1326	0.1454	0.1361	0.1327	0.1529	0.1379	0.1322	0.1317	0.1311
Wang RQA Stat	0.1420	0.1475	0.1368	0.1433	0.1328	0.1368	0.1530	0.1327	0.1326	0.1327	0.1328	0.1426	0.1319	0.1323	0.1323	0.1312
Average	0.1420	0.1428	0.1378	0.1361	0.1341	0.1356	0.1382	0.1323	0.1402	0.1319	0.1324	0.1382	0.1340	0.1322	0.1324	0.1328
Deviation in %	-															
Wang	-	**-5.4%**	**-6.1%**	**-9.2%**	**-9.0%**	**-9.5%**	**-7.0%**	**-9.3%**	**-8.7%**	**-10.8%**	**-9.3%**	**-10.8%**	**-9.9%**	**-9.2%**	**-9.6%**	**-9.9%**
RQA	-	2.1%	**-4.8%**	**-4.5%**	0.3%	**-4.2%**	**-9.0%**	**-7.8%**	**-4.4%**	**-6.1%**	**-7.8%**	**-3.2%**	**-1.3%**	**-7.8%**	**-5.4%**	**-2.9%**
Stat	-	**-1.5%**	**-3.5%**	**-1.1%**	**-5.7%**	**-5.8%**	3.7%	**-6.0%**	5.7%	**-5.6%**	**-5.7%**	**-3.3%**	**-6.5%**	**-6.0%**	**-5.6%**	**-4.9%**
Wang RQA	-	4.3%	3.5%	**-2.4%**	**-3.5%**	0.8%	**-0.8%**	**-3.8%**	0.8%	**-4.6%**	**-3.8%**	**-1.8%**	**-5.8%**	**-3.8%**	**-4.7%**	**-4.2%**
Wang Stat	-	1.0%	**-4.1%**	**-3.2%**	**-3.5%**	2.0%	**-2.6%**	**-3.6%**	6.6%	**-7.6%**	**-3.5%**	**-3.3%**	**-1.8%**	**-3.6%**	**-3.8%**	**-4.1%**
RQA Stat	-	0.2%	**-1.8%**	**-9.0%**	**-10.3%**	**-10.5%**	**-10.0%**	**-10.4%**	**-1.8%**	**-8.0%**	**-10.3%**	3.3%	**-6.8%**	**-10.7%**	**-11.0%**	**-11.4%**
Wang RQA Stat		3.8%	**-3.7%**	0.9%	**-6.5%**	**-3.7%**	7.7%	**-6.6%**	**-6.6%**	**-6.5%**	**-6.5%**	0.4%	**-7.2%**	**-6.9%**	**-6.8%**	**-7.6%**
Average		0.6%	**-2.9%**	**-4.1%**	**-5.5%**	**-4.4%**	**-2.6%**	**-6.8%**	**-1.2%**	**-7.0%**	**-6.7%**	**-2.7%**	**-5.6%**	**-6.9%**	**-6.7%**	**-6.4%**
Classification Accuracy																
Wang	0.21	0.303	0.242	0.485	0.394	0.424	0.394	0.394	0.424	0.485	0.394	0.455	0.364	0.364	0.364	0.364
RQA	0.36	0.273	0.424	0.424	0.333	0.364	0.303	0.424	0.394	0.364	0.424	0.333	0.242	0.424	0.364	0.242
Stat	0.21	0.303	0.394	0.303	0.364	0.273	0.485	0.364	0.182	0.455	0.364	0.303	0.394	0.364	0.364	0.333
Wang RQA	0.36	0.364	0.364	0.424	0.394	0.333	0.333	0.394	0.364	0.424	0.394	0.333	0.424	0.394	0.424	0.394
Wang Stat	0.33	0.242	0.455	0.303	0.364	0.424	0.424	0.394	0.333	0.424	0.364	0.424	0.455	0.394	0.364	0.364
RQA Stat	0.33	0.333	0.212	0.455	0.364	0.303	0.485	0.394	0.212	0.424	0.364	0.121	0.364	0.394	0.394	0.515
Wang RQA Stat	0.27	0.333	0.333	0.364	0.333	0.333	0.333	0.364	0.333	0.424	0.333	0.424	0.424	0.364	0.424	0.485
Average	0.30	0.307	0.346	0.394	0.364	0.351	0.394	0.390	0.320	0.429	0.377	0.342	0.381	0.385	0.385	0.385
Deviation in %																
Wang	-	**42.9%**	**14.3%**	**128.6%**	**85.7%**	**100.0%**	**85.7%**	**85.7%**	**100.0%**	**128.6%**	**85.7%**	**114.3%**	**71.4%**	**71.4%**	**71.4%**	**71.4%**
RQA	-	-25.0%	**16.7%**	**16.7%**	-8.3%	0.0%	-16.7%	**16.7%**	**8.3%**	0.0%	**16.7%**	-8.3%	-33.3%	**16.7%**	0.0%	-33.3%
Stat	-	**42.9%**	**85.7%**	**42.9%**	**71.4%**	**28.6%**	**128.6%**	**71.4%**	-14.3%	**114.3%**	**71.4%**	**42.9%**	**85.7%**	**71.4%**	**71.4%**	**57.1%**
Wang RQA	-	0.0%	0.0%	**16.7%**	**8.3%**	-8.3%	-8.3%	**8.3%**	0.0%	**16.7%**	**8.3%**	-8.3%	**16.7%**	**8.3%**	**16.7%**	**8.3%**
Wang Stat	-	-27.3%	**36.4%**	-9.1%	**9.1%**	**27.3%**	**27.3%**	**18.2%**	0.0%	**27.3%**	**9.1%**	**27.3%**	**36.4%**	**18.2%**	**9.1%**	**9.1%**
RQA Stat	-	0.0%	-36.4%	**36.4%**	**9.1%**	-9.1%	**45.5%**	**18.2%**	-36.4%	**27.3%**	**9.1%**	-63.6%	**9.1%**	**18.2%**	**18.2%**	**54.5%**
Wang RQA Stat	-	**22.2%**	**22.2%**	**33.3%**	**22.2%**	**22.2%**	**22.2%**	**33.3%**	**22.2%**	**55.6%**	**22.2%**	**55.6%**	**55.6%**	**33.3%**	**55.6%**	**77.8%**
Average		**8.0%**	**19.8%**	**37.9%**	**28.2%**	**22.9%**	**40.6%**	**36.0%**	**11.4%**	**52.8%**	**31.8%**	**22.8%**	**34.5%**	**33.9%**	**34.6%**	**35.0%**

*improvements in RO-sMAPE and classification accuracy are in bold

Furthermore, this paper introduced error-based features (landmarkers) for meta-learning to select time series forecasting models. While landmarkers have been applied as features for meta-learning approaches to select classification algorithms, this specific type of feature was neglected so far in the time series forecasting domain. In this paper, landmarking meta-features were generated based on time series forecast errors regarding training sets and validation sets. The empirical study showed particular potential of applying landmarkers in meta-learning for selecting time series forecasting models. In addition, the results of the study suggest that feature selection as well as considering more forecasting models should potentially further improve model selection performance and hence forecasting accuracy.

REFERENCES

[1] C. Kilger, and M. Wagner, "Demand planning," in *Supply chain management and advanced planning: concepts, models, software, and case studies*, H. Stadtler, and C. Kilger, Eds. Berlin: Springer, 2008, pp. 133-160.

[2] S. Makridakis, and M. Hibon, "The M3-competition: results, conclusions and implications", *International Journal of Forecasting*, vol. 16, pp. 451-476, 2000.

[3] S. F. Crone, M. Hibon, and K. Nikolopoulos, "Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction," *International Journal of Forecasting*, vol. 27, pp. 635-660, 2011.

[4] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural computation*, vol. 8, pp. 1341-1390, 1996.

[5] F. Collopy, and J. S. Armstrong, "Rule-based forecasting: development and validation of an expert systems approach to combining time series extrapolations," *Management Science*, vol 38, pp. 1394-1414, 1992.

[6] F. Petropoulos, S. Makridakis, V. Assimakopoulos, and K. Nikolopoulos, "Horses for Courses' in demand forecasting," *European Journal of Operational Research*, vol. 237, pp. 152-163, 2014.

[7] J. R. Rice, "The algorithm selection problem," *Advances in Computers*, vol. 15, pp. 65-118, 1976.

[8] R. Vilalta, and Y. Drissi, "A perspective view and survey of meta-learning," *Artificial Intelligence Review*, vol. 18, pp. 77-95, 2002.

[9] P. Brazdil, C. Girraud-Carrier, C. Soares, and R. Vilalta, *Metalearning: applications to data mining*, Berlin: Springer, 2008.

[10] K. A. Smith-Miles, "Cross-disciplinary perspectives on meta-learning for algorithm selection," *ACM Computing Surveys*, vol. 41, pp. 6:1-6:25, 2009.

[11] C. Lemke, M. Budka, and B. Gabrys, "Metalearning: a survey of trends and technologies," *Artificial Intelligence Review*, pp. 1-14, 2013.

[12] P. Brazdil, J. Gama, and B. Henery, "Characterizing the applicability of classification algorithms using meta-level learning," in *Machine Learning: Proceedings of the European Conference on Machine Learning, ECML-94*, Catania, Italy: Springer, pp. 83-102, 1994.

[13] M. Reif, F. Shafait, M. Goldstein, T. Breuel, and A. Dengel, "Automatic classifier selection for non-experts," *Pattern Analysis and Applications*, vol. 17, pp. 83-96, 2014.

[14] J. N. van Rijn, G. Holmes, B. Pfahringer, and J. Vanschoren, "Algorithm selection on data streams," in *Discovery Science: 17th International Conference, DS 2014*, Bled, Slovenia: Springer, pp. 325-336, 2014.

[15] J. N. van Rijn, G. Holmes, B. Pfahringer, and J. Vanschoren, "Having a blast: meta-learning and heterogeneous ensembles for data streams," in Proceedings of the 15th IEEE International Conference on Data Mining, ICDM 2015, Atlantic City, New Jersey, USA: IEEE Computer Society, pp.1003-1008, 2015.

[16] R. B. C. Prudêncio, and T. B. Ludermir, "Meta-learning approaches to selecting time series models," *Neurocomputing*, vol. 61, pp. 121-137, 2004.

[17] X. Wang, K. Smith-Miles, and R. Hyndman, "Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series," *Neurocomputing*, vol. 72, pp. 2581-2594, 2009.

[18] C. Lemke, and B. Gabrys, "Meta-learning for time series forecasting and forecast combination," *Neurocomputing*, vol. 73, pp. 2006-2016, 2010.

[19] B. Scholz-Reiter, M. Kück, and D. Lappe, "Prediction of customer demands for production planning – Automated selection and configuration of suitable prediction methods," *CIRP Annals - Manufacturing Technology*, vol. 63, pp. 417-420, 2014.

[20] M. Matijaš, J. A. K. Suykens, and S. Krajcar, "Load forecasting using a multivariate meta-learning system," *Expert Systems with Applications*, vol. 40, pp. 4427-4437, 2013.

[21] A. L. D. Rossi, A. C. Ponce de Leon Ferreira de Carvalho, C. Soares, and B. F. de Souza, "MetaStream: A meta-learning based method for periodic algorithm selection in time-changing data," *Neurocomputing*, vol. 127, pp. 52-64, 2014.

[22] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, "Recurrence plots for the analysis of complex systems," *Physics Reports*, vol. 438, pp. 237-329, 2007.

[23] R. Fildes, and F. Petropoulos, "Simple versus complex selection rules for forecasting many time series," *Journal of Business Research*, vol. 68, pp. 1692-1701, 2015.

[24] M. Adya, F. Collopy, J. S. Armstrong, and M. Kennedy, "Automatic identification of time series features for rule-based forecasting," *International Journal of Forecasting*, vol. 17, pp. 143-157, 2001.

[25] C. Shah, "Model selection in univariate time series forecasting using discriminant analysis," *International Journal of Forecasting*, vol. 13, pp. 489-500, 1997.

[26] A. R. Venkatachalam, and J. E. Sohl, "An intelligent model selection and forecasting system," *Journal of Forecasting*, vol. 18, pp. 167-180, 1999.

[27] N. Meade, "Evidence for the selection of forecasting methods," *Journal of Forecasting*, vol. 19, pp. 515-535, 2000.

[28] C. Lemke, and B. Gabrys, "Meta-learning for time series forecasting in the NN GC1 competition," in *Proceedings of the 2010 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2010*, Barcelona, Spain: IEEE Computer Society, pp. 1-5, 2010.

[29] B. Pfahringer, H. Bensusan, and C. Giraud-Carrier, "Tell me who can learn you and I can tell you who you are: Landmarking various learning algorithms," in *Proceedings of the 17th international conference on machine learning, ICML 2000*, pp. 743-750, 2000.

[30] L. J. Tashman, "Out-of-sample tests of forecasting accuracy: an analysis and review," *International Journal of Forecasting*, vol. 16, pp. 437-450, 2000.

[31] F. Takens, "Detecting strange attractors in turbulence," *Dynamical Systems and Turbulence*, vol. 898, pp. 366-381, 1981.

[32] H. Kantz, and T. Schreiber, *Nonlinear Time Series Analysis*, Cambridge: Cambridge University Press, 2004.

[33] M. Kück, and B. Scholz-Reiter, "A Genetic Algorithm to Optimize Lazy Learning Parameters for the Prediction of Customer Demands," in *Proceedings of the 12th IEEE International Conference on Machine Learning and Applications, ICMLA 2013*, Miami, Florida, USA: IEEE Computer Society, pp. 160-165, 2013.

[34] M. Kück, B. Scholz-Reiter, and M. Freitag, "Robust Methods for the Prediction of Customer Demands based on Nonlinear Dynamical Systems," *Procedia CIRP*, vol. 19, pp. 93-98, 2014.

[35] B. Ripley, W. Venables, and M. B. Ripley, "Package 'nnet'", CRAN, 2015.

[36] M. Kuhn, "Caret package," *Journal of Statistical Software*, vol. 28, pp. 1-62, 2008.