Distribution-driven Regression Ensemble Construction for Time Series Forecasting

Florian Wimmenuer, Evgueni Smirnov, Matúš Mihalák

Maastricht University, P.O. Box 616 6200 MD Maastricht, The Netherlands,
f.wimmenauer@student.maastrichtuniversity.nl

Abstract. This paper introduces a two-stage approach on selecting members of an ensemble generating accurate forecasts of a time series with a small amount of forecasts. In the first stage, models trained on similarly-distributed data are selected based on time series stationarity, more specially, the local stationary of sub-series. The second stage identifies the most diverse models among those selected in the first stage to compose the final ensemble. Diversity is measured either on pair-wise basis or over the complete ensemble. In both case, multiple novel diversity metrics are introduced. Additionally, the presented approach is highly modular and does not presuppose the type of comprising models of the ensemble. The experiments show that the proposed approach outperforms the base line when predicting both synthetic and real-world time series.

Keywords: Time Series Forecasting, Ensemble Forecasting, Distribution Similarity

1 Introduction

In machine learning, a forecasting or prediction task is the estimation of future events based on previous observations. Forecasting has a wide range of applications, such as weather forecasts, stock value predictions, and fine-grained forecasts of energy consumption in specific geographic areas. A special field of forecasting which has drawn significant research interest is the forecast of time series. A time series is an indexed series of data points where each index typically corresponds to a point in time. Unlike traditional data sets where the ordering of data points do not matter, a time series is ordered.

With the increasing computation power and advances in machine learning techniques, the task of time series forecasting has become more approachable. In the pursuit of higher prediction accuracy, ensemble forecasts have gained attention of both researchers and practitioners. An ensemble is a collection of models, where multiple individual predictions are combined into one single prediction with the aim of obtaining better forecasting performance than an individual model otherwise would. A key challenge in constructing ensembles is to choose a set of models that yield desirable forecasting performance when combined. An

important feature of an ensemble is the diversity among its members. The ensemble diversity can be seen as the extent of disagreement among the models. In an extreme case, if all models in the ensemble would produce identical forecasts, no performance gain could be achieved.

This paper proposes a novel framework of ensemble construction for time series forecasting from a collection of trained models. Firstly, the members of the ensemble are selected based on their relevance to the prediction task. That is, those models trained on distributions different from the encountered one are disregarded. Secondly, among the relevant models, only the most diverse ones compose the final ensemble. Last but not least, new measures of diversity for regression ensembles are proposed. It is noteworthy that this framework is highly modular and does not presuppose the type of comprising models of the ensemble.

The proposed ensemble construction approach is based on a collection of trained models, each of which concerns a part (sub-series) of a longer time series. The relative position of the sub-series in the full time series is known. When new prediction tasks are encountered, firstly, previous parts of the full time series that have the same distribution with the current sub-series are detected. Within the set of sub-series sharing the same distribution, the most diverse ones are chosen, and their corresponding trained models form the final ensemble. In this process, the selection of sub-series with distribution and the measurement of diversity are based on statistical tests rather than techniques specific to the trained models. Therefore, the proposed approach is model-independent and thus can be applied to ensemble construction of any time series without structural modifications.

The presented approach is evaluated on a synthetic time series of known underlying distribution and a real-world time series consisting of a store's sales values in five-minute buckets. For the latter, every day over the past year, models are trained based on data from the previous seven days with the goal of predicting the sales value of the next day.

This section is followed by a short survey of related work, preliminaries and introduction to the relevant concepts. In section 4, applicable methods are defined and the new measure of diversity for regression ensembles is developed. Afterwards, the experiments and results are presented in section 5. Finally, the paper closes with the conclusions in section 6.

2 Related Work

Time series forecasting can be conducted via various approaches. Statistical learning models such as the autoregressive integrated moving average (ARIMA) are widely discussed [17]. More recently, there has been a raise in machine learning approaches such as regression trees [15], regression support vector machines [16] and deep learning models [1, 24]. Notably, deep learning models became very applicable since recurrent neural networks (RNN) such as long short-term memory recurrent neural networks (LSTM) [8] were introduced. In contrast to standard feed-forward neural networks, LMTSs are capable of 'remembering' previously-encountered data and therefore learn time- or sequence-based con-

cepts. With the increasing computation power, the model architecture became deeper, achieving higher predictive power [24]. RNNs nowadays often outperforms their traditional counterparts on time series forecasting tasks [1].

A shortcoming of RNNs is the computational load and therefore time consumption. In case not all training data is available upfront and the models need to be adapted in later stages, RNNs have to be retrained on the full set of data. Retraining on the newly available data alone will lead to the loss of acquired knowledge on the older data [23]. This issue can be resolved by constructing an ensemble of models trained on different parts of the time series [23]. Ensemble learning in general is a highly active field of research, because well-chosen ensembles can increase the performance compared to single-model approaches [4, 10, 13, 21]. Error reduction via ensemble prediction can only be achieved when the ensemble is diverse, meaning that the members of the ensemble disagree with each other to a certain extent [4, 10, 13, 21]. It can be trivially seen that an ensemble of 1,000 models all producing the same forecasting result cannot lead to any performance gain over a single model. Measuring or even ensuring the diversity in an ensemble is so far mostly done for classification problems [12]. For regression problems, often specialised approaches are introduced [4, 13]. One way to increase the diversity in an ensemble is to train the models on different subsets of the data [21].

Ensemble construction for a longer time series faces additional challenges due to the possible changes in the underlying distribution of the data. In this case, a traditional approach for model selection is to compute the distances between sub-series by dynamic time warping [27] and selecting the least distant models.

3 Preliminaries

When training models with supervised machine learning techniques, one should ensure that the distribution of the data on which the models are trained has the same distribution with the data to be processed at production time [20]. A model trained on data from one distribution is not assured to make accurate predictions on data from a completely different distribution [20]. The same principle holds when constructing an ensemble from a collection of trained models. Otherwise, members of the ensemble might contribute a higher error due to incorrect assumptions on the underlying distribution. In the context of this paper, namely time series forecasting, this means that only the models trained on similarly-distributed data as the data to be forecasted are suitable to contribute to the final ensemble forecast.

Testing whether two data sets have the same underlying distribution is a well-researched topic [14]. From a time series point of view, the statistical properties of the data are closely related to stationarity. A stationary time series is defined as a time series whose statistical properties do not change over time [17]. When it comes to stationary times series, one distinguishes between first order stationary and second order stationary time series.

In a first order stationary time series, all moments of all degrees such as mean and variance do not change throughout the complete time series. This means that for a time series $T = t_1 \dots t_L$ with length L, the joint statistical distribution $X_{t_1} \dots X_{t_l}$ is the same as the joint statistical distribution $X_{t_1+\tau} \dots X_{t_l+\tau}$ for all l and τ where $l + \tau \leq L$. In contrast to this strict definition, in a second order stationary time series, the mean and the variance are constant and the auto-covariance between X_t and $X_{t+\tau}$ solely depends on the lag τ. With the first order stationarity being a mostly overly strict assumption for real-world data sets, the upcoming sections focus on the second order stationarity. A test for second order stationarity of a time series was introduced by Priestley and Subba Rao [25]. In this test, an analysis of variance is performed on the logarithm of a time-varying spectral estimate at a set of times and frequencies in the Fourier spectrum.

For a long time-series, second order stationary cannot be assumed on the complete series. A more reasonable assumption is that the long time series might contain second order stationary parts and that the complete time series is locally stationary. To test a time series for local second order stationarity and to localise these parts within the time series, a test introduced by Nason [18] can be used. This test is closely related to the work of Priesley and Subba Rao [25], but analyses the evolutionary wavelet spectrum [19] instead of the Fourier spectrum of the time series.

4 Methods

After introducing the general framework, this section details the methods used to construct an ensemble. Firstly presented are approaches for selecting models trained on similarly-distributed data. Secondly, various ensemble diversity measures are discussed. Finally, three optimisation methods for selecting an ensemble with the highest diversity are introduced.

4.1 Sub-Series with Similar Underlying Distribution

Two methods are introduced to construct an ensemble of models with distribution similarity. The former is based on the time series property of stationarity, whereas the latter uses traditional statistics techniques.

Time Series Stationarity Properties As mentioned in section 3, a stationary time series has the same underlying distribution throughout the complete series. This section focuses on testing second order stationarity. To this end, and to additionally localise the stationary parts within the time series, a test introduced by Nason [18] is used. The choice for this test was based on, besides the reported runtime of $\mathcal{O}(l \log l)$ for a time series of length l, the fact that it does not assume a normal distribution as the underlying distribution of the time series.

The test is based on evaluating the stability of the expected value $\beta_j(k/T)$ of the wavelet periodogram

$$I_\ell(k/T) = I_{j,k} = \left(\sum_{t=1}^{T} X_t \psi_{j,k-t} \right)^2 \tag{1}$$

over a finite set of wavelet scales ℓ. In general, $\ell \in \mathbb{N}$ but since T is finite, $\ell = 1, \ldots, (J = \log_2 T)$. X_t is the time series over time $t = 1, \ldots, T$, $\{\psi_{j,k}\}_{j,k}$ is a set of non-decimated discrete wavelets, $k = 1, \ldots, T$ and $j \in \mathbb{N}$ is the scale of the wavelet. The tested time series is stationary if and only if for all scales ℓ, $\beta_\ell(z)$ is a constant function for rescaled time $z = k/T \in (0, 1)$

To test if the $\beta_\ell(z)$ is a constant function, the Haar wavelet coefficients of time series are analysed. These coefficients are given as

$$v_{i,p}^{(\ell)} = \int_0^1 \beta_\ell(z) \psi_{i,p}^H(z) dz \tag{2}$$

for Haar wavelet scale $i = 1, \ldots, J$, $p = 1, \ldots, 2^i - 1$, with $\{\psi_{i,p}^H(t)\}_{i,p}$ being the Haar wavelets.

The null hypothesis H_0 is that the quantity $\beta_\ell(z)$ is a constant function of z and hence $v_{i,p}^{(\ell)} = 0$ for all ℓ, i, p, as $\int \psi(z) dz = 0$ is a defining wavelet property. The set of test statistic $\mathscr{S}$ is:

$$\mathscr{S}_{i,p}^{(\ell)} = \hat{v}_{i,p}^{(\ell)} \hat{\sigma}_{i,p}^{(\ell)-1} \tag{3}$$

As proposed by von Neuman et al. [29], an estimate for $\hat{\sigma}_{i,p}^{(l)2}$ is the variance of the estimated wavelet periodogram:

$$\mathrm{var}(\beta_\ell) = \int_{-\pi}^{\pi} f(\omega) \left| \hat{\psi}_\ell(\omega) \right|^2 d\omega \tag{4}$$

In this case the classical stationary spectrum $f(\omega)$ can be estimated by the regular periodogram in equation (1).

To evaluate the test statistic, the empirical wavelet periodogram values $I_{\ell,k} = I_\ell(k/T)$ for $k = 1, \ldots, T$ are analysed. With this, the Haar wavelet coefficents are estimated for scale ℓ by

$$\hat{v}_{i,p}^{(\ell)} = 2^{i/2} \left(\sum_{r=0}^{2^{i-1}-1} I_{\ell, 2^i p - r} - \sum_{q=2^{i-1}}^{2^i - 1} I_{\ell, 2^i p - q} \right) \tag{5}$$

Based on $\mathscr{S}$ being a set of test statistics for multiple H_0, multiple hypothesis tests must be performed. For combining these multiple hypotheses, the false discovery rate method [3] is used, because this method is less conservative than for example the Bonferroni correction [7].

The above-mentioned test on a time series not only provides knowledge about stationarity but also locates non-stationarity. For each test statistic in $\mathscr{S}$, with

ℓ, i and p, the corresponding part of the time series and thus the non-stationary parts where the related H_0's were rejected can be identified. This can, in the context of this paper, find trained models that are not suitable for contributing to the ensemble forecast. Therefore, only the parts where none of the H_0's are rejected qualify for contributing trained models to the ensemble.

Statistics Properties After discussing finding the time series parts with the same underlying distribution via local stationarity tests, the focus is shifted from time series properties to traditional statistical analyses of the corresponding training data of the models. Under the assumption that each model is trained on a sub-series of the time series, the distribution of the training data of each model is analysed.

The underlying distributions are compared between the current training data on hand and the corresponding data where existing models were trained. Two different tests are used, namely the two-sample Kolmogorov-Smirnov test (KS test) [14] and the two-sample Anderson-Darling test (AD test) [2]. The two-sample KS test tests whether two distributions differ. The null hypothesis H_0 is that this is not the case. Like in all KS tests, each instance is weighted equally. However, many distributions differ primarily in their tails, which poses a challenge to the KS test. The Anderson-Darling test is specially suited to detect differences in the tails of distributions. In this case, a two-sample AD test is required. [22] extended the standard AD test to a two sample version. Similar to the two-sample KS test, the two-sample AD test examines whether two underlying distributions differ by testing the null hypothesis H_0 that the two samples are from the same distribution.

4.2 Diversity

Once the appropriate models are selected, the diversity of the resulting ensemble needs to be evaluated. Firstly, techniques for evaluating the diversity of a pair of models are discussed, followed by diversity measures on the complete ensemble. Analogous to the methods in section 4.1, the diversity is only measured based on distributions, making the approach applicable regression tasks in general.

Pair-Wise Diversity The diversity between trained models can be measured in various ways. Let $m_u, m_v \in M$ be two models trained on the time series $Y^u = [y_1^u, y_2^u, \dots y_{n_Y}^u]$ and $Y^v = [y_1^v, y_2^v, \dots y_{n_Y}^v]$ respectively, where the training series has n_Y steps. Straightforward diversity measures include the covariance, the correlation coefficient and entropy [5]. The covariance between Y^u and Y^v is defined as

$$\mathrm{cov}(Y^u, Y^v) = \frac{1}{n^2} \sum_{i=1}^{n_Y} \sum_{j=i+1}^{n_Y} (y_i^u - y_j^u)(y_i^v - y_j^v) \tag{6}$$

A smaller covariance indicates less association between models and therefore higher diversity. As the magnitude of covariance is in general hard to interpret,

the correlation coefficient, the normalisation of the covariance, may serve as a better diversity measure than the covariance. The correlation coefficient between Y^u and Y^v is defined as

$$\operatorname{corr}(Y^u, Y^v) = \frac{n_Y \sum y_i^u y_i^v - \sum y_i^u - \sum y_i^v}{\sqrt{n_Y \sum y_i^u - (\sum y_i^u)^2}\sqrt{n_Y \sum y_i^v - (\sum y_i^v)^2}}. \tag{7}$$

Here $\sum$ abbreviates $\sum_{i=1}^{n_Y}$. Since diversity is inversely proportional to correlation, ensembles with lower correlation coefficients are preferred.

Besides the above-mentioned measurements, the entropy of a system provides an indication of diversity. The entropy between Y^u and Y^v is defined as

$$\operatorname{ent}(Y^u, Y^v) = -\frac{1}{2}\ln(\sigma_u^2\sigma_v^2(1 - \operatorname{corr}(Y^u, Y^v)^2)), \tag{8}$$

where $\operatorname{corr}(Y^u, Y^v)$ is the correlation coefficient as defined above in equation (7).

In addition to the conventional approaches, a novelty in this paper is to measure diversity via statistical tests. Recall that the aforementioned two-sample KS test and the two-sample AD test examine whether two given samples come from the same underlying distribution under the null-hypothesis H_0 that they do. This means that if the computed p-value is high or higher than a critical value, one fails to reject H_0. In this light, a lower p-value implies higher diversity.

Ensemble Diversity Based on all pair-wise diversity measures of the n_E models, measuring the complete ensemble diversity requires all training series of the models. Let all n_E models $m_j \in E$ be trained on the $n_Y \times n_E$ values

$$Y = \begin{pmatrix} y_1^{[1]} & y_1^{[2]} & \cdots & y_1^{[n_E]} \\ y_2^{[1]} & y_2^{[2]} & \cdots & y_2^{[n_E]} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n_Y}^{[1]} & y_{n_Y}^{[2]} & \cdots & y_{n_Y}^{[n_E]} \end{pmatrix} \tag{9}$$

Extending the previously introduced pair-wise diversity measure to the entire ensemble can be conveniently achieved by averaging over all pairs. For instance, the ensemble covariance can be defined based on pairwise covariance as follows:

$$\operatorname{cov}_E = \frac{1}{n_E(n_E - 1)} \sum_{i=1}^{n_E} \sum_{\substack{j=1 \\ j \neq i}}^{n_E} \operatorname{cov}(Y^{[i]}, Y^{[j]}). \tag{10}$$

Moreover, the averaging of pair-wise diversity such as equation (10) can be extended to account for different weights.

Whereas measurements like correlation coefficients and entropy can be averaged, combining p-values is less straightforward. In this case, the p-values $P = [p_{1,1}, p_{1,2}, \ldots, p_{n_E,n_E}]$ are combined using Stouffer's z-score method explained by [30]. Stouffer's z-score method was chosen over other methods such

as Fishers's method, because Stouffer's z-score method allows the extra degrees of freedom with weights $W = [w_{1,1}, w_{1,2}, \ldots, w_{n_E,n_E}]$ which specify the relative importance of pairwise p-values. The combined p-value is:

$$p_N = 1 - \Phi\left(\frac{\sum_{i=1}^{n_Y}\sum_{j=1}^{n_Y} w_{i,j}\Phi^{-1}(1-p_{i,j})}{\sqrt{\sum_{i=1}^{n_Y}\sum_{j=1}^{n_Y} w_{ij}^2}}\right) \tag{11}$$

where p_{n_Y} is the p-value of the model pair and Φ, Φ^{-1} are the standard normal cumulative distribution function and its inverse respectively. In addition to combining the p-values of the pair wise test, k-sample testes such as the k-sample AD test [28] can be performed.

A new diversity measure based on disagreement measure introduced by [12] is defined as follows: For each of the n_Y series points per model training series, the standard deviation in Y is computed. Then an $n_Y \times n_E$ error matrix ε is built from Y where $e_i^{[j]} \in [0,1], i = 1, \ldots, n_Y, j = 1, \ldots, n_E$, where $e_i^{[j]} = 0$ if $e_i^{[j]}$ falls within a margin of one standard deviation around the mean of the forecast value of all models. Otherwise $e_i^{[j]} = 1$. Based on this, the new disagreement measure of the ensemble is

$$dis = \frac{\sum_{i=1}^{n_Y}\sum_{j=1}^{n_E} e_i^{[j]}}{n_Y n_E} \tag{12}$$

4.3 Selecting Most Diverse Ensemble

The methods in section 4.1 provide possibilities to select n_M feasible models $m_k \in M$ from the pool of models. With the methods in section 4.2 the diversity of an ensemble can be measured. This section introduces methods to select an ensemble $E \subset M$ with n_E models from the feasible n_M models such that the diversity d_E is maximised.

One of the most straightforward but computational intensive possibility for selecting the best n_E models is to enumerate all $\binom{n_M}{n_E}$ unique combinations of size n_E out of the n_M feasible models and select the combination with the highest diversity d_E. For larger n_M this becomes infeasible. In these cases, an approximation of the highest diversity d_E can be found for example with simulated annealing [9].

To find the optimal subset of n_E models with the pair-wise diversity measures, a mixed integer linear programming (MILP) approach is chosen. The underlying idea is to select the models such that the minimum of the pair-wise diversities is maximised. To model this as an MILP, firstly a variable d_E representing the diversity of the ensemble is introduced. Secondly, n_E binary variables $x_1, \ldots, x_{n_E}$ representing if a model m_k is selected or not are added to the program. Finally the following optimisation problem, as similarly introduced by Kuby [11], must be solved:

$$
\begin{aligned}
\text{maximise} \quad & d_E \\
\text{subject to} \quad & \sum_{i=1}^{n_E} x_i = n_E, \\
& d_E \leq M(2 - x_i - x_j) + d_{ij},\ 1 \leq i \leq j \leq n_M \\
& \forall x, \in \{0, 1\} \\
& d_E \geq 0
\end{aligned}
$$

For the above-mentioned NP-complete MILP [6], a linear time approximation can be achieved by selecting the both most distant models and iteratively adding the most distant remaining model to the growing ensemble until n_E models are selected. Due to its greedy nature, this solution might not reach the optimum. As shown in the proof by Ravi et al. [26], for problems where the triangular inequality holds, this method is a 2-approximation of the optimal solution.

5 Experiments and Results

In order to evaluate the performance of the proposed methods, multiple experiments is conducted. Section 4.1 introduced three different approaches for model selection based on distribution similarity and 17 approaches on selecting the most diverse models, the combination of which result in 51 setups. For each experiment, all possible forecasts for the sub-sequences are performed and diversity mean (div_μ), diversity standard deviation (div_σ), mean squared forecast error (MSE), the mean MSE (MSEμ) and the MSE standard deviation (MSEσ) are computed. The experiments were conducted on two time series. The first one is a synthetic time series, generated from three different beta-distributions with known parameters. The synthetic time series is divided into 150 sub-series each coming from one of the three underlying distributions. The models to be selected are a collection of LSTMs, where each model is trained per sub-sequence. The second time series is a real-world time series representing the sales values of a store in five minutes buckets. Every day, 20 LSTMs are trained on data from the past seven days with different time lags. The data set spans over 324 continuous days. On both data sets, a fixed 10% of available models are selected into the ensemble.

In the first experiment, the three approaches for selecting sub sequences with the same distribution are compared. Table 1 presents the selection accuracy and the false positive (FPR) rate. The FPR measures the proportion of models in the ensemble that are not trained on the correct distribution and therefore can lead to higher prediction error.

	Accuracy_μ	Accuracy_σ	FPR_μ	FPR_σ
LS	0.97	0.05	0.001	0.005
KS	0.96	0.06	0.03	0.06
AD	0.95	0.07	0.02	0.03

Table 1: Performance of distribution similarity selection with local stationarity (LS), Kolmogorov-Smirnov test (KS) and Anderson-Darling test (AD)

As expected, the local-stationarity-based approach outperforms the traditional statistical measures. The higher FPR of the KS test suggests that the previously-mentioned shortcomings in detecting differences in the tails of the distributions do play a role.

For the comparison of different selection methods, three base lines are introduced. Base line A does not perform any selection and builds an ensemble of all available models. Base line B performs the distribution-based model selection but does not consider diversity. In base line C, the most diverse models are selected, but distribution similarity is not taken into account. Performance measures in terms of the mean squared error (MSE) are shown in table 2.

	Synthetic			Real-world		
	MSE_μ	MSE_σ	MSE_{median}	MSE_μ	MSE_σ	MSE_{median}
Base line A	231.63	204.84	128.85	659.10	248.40	585.64
Base line B	74.41	22.32	67.73	614.94	281.67	557.49
Base line C	203.91	122.42	190.21	630.42	264.36	562.41
Selected	69.24	20.75	69.02	603.27	246.46	556.93

Table 2: Performance in terms of MSE of the base lines and of the best selection

The performance differences among the base lines clearly suggest the importance of basing the models on similar distributions while training and forecasting. The same tendency can be observed in the real-world time series, although the difference is less salient due to more similar distributions.

Tables 3 and 4 show results obtained by the three optimisation methods for models pre-selected by diversity similarity evaluated by the local stationarity.

	Approximation				MILP			
	div_μ	div_σ	MSE_μ	MSE_σ	div_μ	div_σ	MSE_μ	MSE_σ
cov	6907.36	2191.24	79.12	22.44	6694.47	2059.24	71.01	22.02
corr	0.98	0.07	69.76	21.40	0.96	0.01	69.52	19.26
ent	7.14	0.25	71.89	20.62	7.24	0.23	68.77	20.74
p-ks_2	0.08	0.02	69.63	20.78	0.28	0.08	69.7	20.82
p-ad_2	0.13	0.20	69.24	20.75	0.18	0.03	69.87	20.10

Table 3: Performance of pair-wise diversity selections on local stationarity preselected synthetic time series

	div_μ	div_σ	MSE_μ	MSE_σ
cov	6354.32	1989.95	79.07	22.06
corr	0.95	0.07	71.61	22.76
ent	7.38	0.29	70.32	21.84
p-ks_2	0.05	0.01	70.13	20.59
p-ad_2	0.04	0.01	69.47	20.53
p-ad_k	0.02	0.01	71.28	21.01
dis	0.75	0.05	69.43	20.97

Table 4: Performance of combinatoric diversity selection on local stationarity preselected synthetic time series

For the MILP approach and the combinatorial optimisation, the covariance performs worse than the other diversity measures, whereas the correlation coefficient yields stable results for all method on both experiment data sets. The performance difference could be attributed to the correlation coefficient being a normalised measure wheareas the covariance being on a less workable scale. Throughout all tests, the two-paired AD test and the correlation coefficient yield stable results. The newly proposed diversity measure for the complete ensemble yields the best results, but due to the computation of the standard deviation for each point of the time series, it is the slowest of all tested approaches.

The differences in the optimal solutions for pair-wise diversity measure or a measure on the complete ensemble are not this prominent. In the tested cases, the pair-wise optimisation perform slightly better than the measurements on the complete ensemble. Additionally, the pair-wise evaluations are less computational intensive.

The proposed two-stage selection on diversity and distribution equality outperforms all three base lines. To evaluate if the proposed methods are performing significantly better than the base lines, two-sample t-tests are performed. At a

significance level of $\alpha = 0.02$, the proposed method achieves significantly lower MSE than base line A and C. Although the performance gain over base line B is not statistically significant, the numerical differences in the results are visible.

After evaluating the performance of the methods from the forecasting accuracy and diversity measure point of view, the performance of the approximation, the MILP method and the combinatorial approach are tested. Figures 1, 2 and 3 show a two dimensional scaling of the pair-wise distances between 17 different real-world time series measured with the described diversity metrics. A smaller random subset of the time series was chosen for the sake of clearer visualisation. It is noteworthy that the distances in the two dimensional visualisation is an approximation of the actual distances for most of the measures, as entropy is the only diversity measure where triangle inequality holds. It can be seen that both the pair-wise approaches of approximation and MILP give different selections. Visually, the MILP selection appears to be more diverse than the approximation selection. This is in line with the assumption in section 4.2 that the greedy approach might not be optimal. Even though one could have expected that the MILP approach and the combinatorial approach yield the same result, it can be seen that this is not the case. This happens because the objective is different. The MILP approximates the diversity by pair-wise measurements whereas the combinatorial approach takes the final ensemble diversity into account. Taking the entropy as measurement, MILP and combinatorial approach made the same selection.

Additionally, it can be seen that the shapes of the distance visualisations of the correlation and covariance are similar, which is as expected. Moreover, the MILP method made marginally different selections for correlation and covariance. The correlation coefficient is shown to outperform the covariance as a diversity measure for all tested cases.

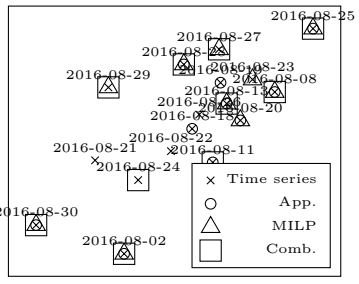

Fig. 1: Estimated two dimensional visualisation of covariance diversity selections

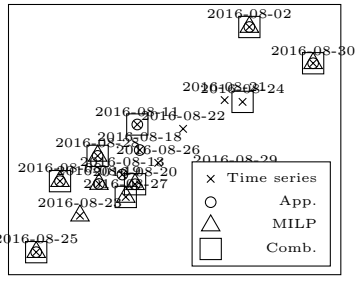

Fig. 2: Estimated two dimensional visualisation of correlation coefficient diversity selections

Finally, figure 4 shows the increase in accuracy throughout the selection and forecast process. The figure shows the range where the forecast could reside when

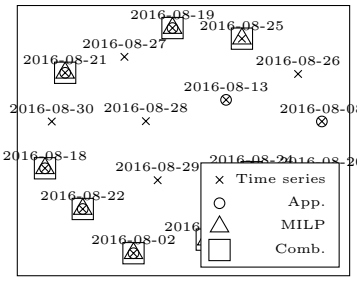

Fig. 3: Estimated two dimensional visualisation of entropy diversity selections

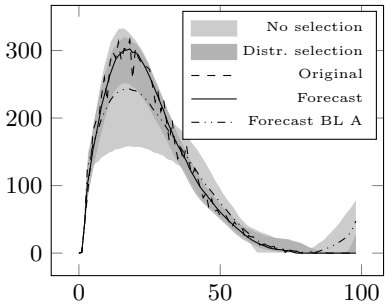

Fig. 4: Both stages of selection with final forecast of one sub-series

no selection is performed, the shrunken area which results from the diversity-based selection and the final forecast with the actual values.

6 Conclusion

This paper has shown that selecting specific models as members for an ensemble from a pool of trained models can increase the performance of the forecast. An increase in performance can be gained if models are selected based on the distributions of their training series. It is also important to select a set of diverse models for the final forecast. With the proposed methods, accurate results can be achieved by performing a smaller amount of forecasts than an ensemble of all available models would need. Testing for these distributions is challenging in the field of time series. It was shown that analysing stationary of time series yields good result, so does traditional statistical tests which consider differences in the tails of the underlying distributions.

Multiple measurements for diversity in forecasting ensembles based on the training series were introduced. Diversities in ensembles can be measured both on pair-wise basis and over the complete ensemble at once. In the tested series, the pair-wise measurement was faster in computation and marginally better in performance. The newly proposed ensemble-level measurement yielded highly promising results.

Lastly, optimising based on the size of the ensemble and the weights of its members could lead to an additional increase in performance and opens an interesting field for further research. It can be concluded that the underlying distribution of time series is an important feature in ensemble construction for time series forecasting.

Bibliography

[1] Abou-Nasr, M.: Time series forecasting with recurrent neural networks nn3 competition (2007)

[2] Anderson, T.W., Darling, D.A.: A test of goodness of fit. Journal of the American statistical association 49(268), 765–769 (1954)

[3] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological) 57(1), 289–300 (1995)

[4] Brown, G., Wyatt, J.L., Tiňo, P.: Managing diversity in regression ensembles. Journal of Machine Learning Research 6(9), 1621–1650 (2005)

[5] Dutta, H.: Measuring diversity in regression ensembles. In: IICAI. vol. 9, p. 17 (2009)

[6] Erkut, E.: The discrete p-dispersion problem. European Journal of Operational Research 46(1), 48–60 (1990)

[7] Glickman, M.E., Rao, S.R., Schultz, M.R.: False discovery rate control is a recommended alternative to bonferroni-type adjustments in health studies. Journal of clinical epidemiology 67(8), 850–857 (2014)

[8] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)

[9] Khachaturyan, A.G., Semenovskaya, S.V., Vainstein, B.: A statistical-thermodynamic approach to determination of structure amplitude phases. Sov. Phys. Crystallogr 24, 519–524 (1979)

[10] Krogh, A., Vedelsby, J.: Neural network ensembles, cross validation, and active learning. Advances in neural information processing systems 7, 231–238 (1995)

[11] Kuby, M.J.: Programming models for facility dispersion: The p-dispersion and maxisum dispersion problems. Geographical Analysis 19(4), 315–329 (1987)

[12] Kuncheva, L.I., Whitaker, C.J.: Ten measures of diversity in classifier ensembles: limits for two classifiers. In: A DERA/IEE Workshop on intelligent Sensor Processing (Ref. No. 2001/050). pp. 10–16. IET (2001)

[13] Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine learning 51(2), 181–207 (2003)

[14] Law, A.M., Kelton, W.D.: Simulation modeling and analysis, vol. 2. McGraw-Hill New York (1991)

[15] Meek, C., Chickering, D.M., Heckerman, D.: Autoregressive tree models for time-series analysis. In: Proceedings of the 2002 SIAM International Conference on Data Mining. pp. 229–244. SIAM (2002)

[16] Müller, K., Smola, A., Rätsch, G., Schölkopf, B., Kohlmorgen, J., Vapnik, V.: Using support vector machines for time series prediction. Advances in kernel methods—support vector learning pp. 243–254 (1999)

[17] Nason, G.P.: Stationary and non-stationary time series. Statistics in Volcanology. Special Publications of IAVCEI 1, 000–000 (2006)
[18] Nason, G.P.: A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75(5), 879–904 (2013)
[19] Nason, G.P., Von Sachs, R., Kroisandt, G.: Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62(2), 271–292 (2000)
[20] Ng, A.: Machine Learning Yearning - Technical Strategy for AI Engineers, In the Era of Deep Learning (2017)
[21] Oliveira, M., Torgo, L.: Ensembles for time series forecasting. In: ACML (2014)
[22] Pettitt, A.N.: A two-sample anderson–darling rank statistic. Biometrika 63(1), 161–168 (1976)
[23] Polikar, R., Upda, L., Upda, S.S., Honavar, V.: Learn++: An incremental learning algorithm for supervised neural networks. IEEE transactions on systems, man, and cybernetics, part C (applications and reviews) 31(4), 497–508 (2001)
[24] Prasad, S.C., Prasad, P.: Deep recurrent neural networks for time series prediction. CoRR abs/1407.5949 (2014)
[25] Priestley, M.B., Rao, T.S.: A test for non-stationarity of time-series. Journal of the Royal Statistical Society. Series B (Methodological) 31(1), 140–149 (1969)
[26] Ravi, S.S., Rosenkrantz, D.J., Tayi, G.K.: Heuristic and special case algorithms for dispersion problems. Operations Research 42(2), 299–310 (1994)
[27] Salvador, S., Chan, P.: Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis 11(5), 561–580 (2007)
[28] Scholz, F.W., Stephens, M.A.: K-sample anderson–darling tests. Journal of the American Statistical Association 82(399), 918–924 (1987)
[29] Von Sachs, R., MacGibbon, B.: Non-parametric curve estimation by wavelet thresholding with locally stationary errors. Scandinavian Journal of Statistics 27(3), 475–499 (2000)
[30] Whitlock, M.C.: Combining probability from independent tests: the weighted z-method is superior to fisher's approach. Journal of evolutionary biology 18(5), 1368–1373 (2005)