Model selection using dimensionality reduction of time series characteristics

Agus Widodo, Indra Budi
Faculty of Computer Science, University of Indonesia
Email: agus.widodo@ui.ac.id, indra@cs.ui.ac.id

Abstract
To find the best forecasting model, several methods are usually tried on training dataset and the best one is selected to forecast the testing dataset. In this paper, we propose a model which selects a forecasting method based on its previous performance on similar dataset assuming that we have historical database of predictor's performance. Thus, we need to obtain the characteristics of time series for similarity measure. Since not all of time series characteristics may necessarily be used to identify certain time series, we reduce the dimensionality of those characteristics using Principal Component Analysis. We use diverse forecasting methods ranging from simple to sophisticated ones, measure as Random Walk, ARIMA, interpolation, S-curve, and Multiple Kernel Learning. To cope with large dataset, we utilize clustering and similarity measure to select only a subset of training dataset. We use the 3003 dataset from M3 competition to construct the historical database which acts as training dataset and 111 from the M1 competition as testing dataset. Our experimental results indicate that our model selection may perform well compared to each individual predictor. The advantage of our method lies in its ability to select suitable predictors without trying all of them on current dataset.

Keywords: forecast combination, time series, dimensionality reduction, similarity measure, model selection.

1. Introduction

Methods for predicting the future values based on past and current observations have been pursued by many researchers and elaborated in many literatures in recent years. Those methods include data pre-processing, enhancing the prediction's methods, and combining those methods. Data preprocessing includes smoothing, selecting relevant features or selecting relevant training dataset.

Meanwhile, several prediction methods have been studied and used in practice. The most common ones are statistical methods based on autoregressive models of time series, as stated by (González-Romera, Jaramillo-Morán, & Carmona-Fernández, 2006) and (S. G. Makridakis, Wheelwright, & R. J. Hyndman, 1998). More advanced approaches apply nonlinear models based mainly on artificial neural networks (ANN), support vector machine (SVM), and other machine learning methods as studied by (Cao, 2003), (G. P. Zhang & Kline, 2007), (Kourentzes & Crone, 2008), (Crone & Kourentzes, 2009), and (G.-bin Huang, D. H. Wang, & Lan, 2011).

To select best model, a common approach is to train many predictors and then pick the one that guarantees the best prediction on out-of-sample (verification) data, as done by (Siwek, Osowski, & Szupiluk, 2009). Another approach is to take into account some best prediction results, and then combine them into an ensemble system to get the final forecast result as suggested by (C. Huang, Yang, & Chuang, 2008) and (Armstrong, 2001). Similarly, (Poncela, Rodríguez, Sánchez-Mangas, & Senra, 2011) combined several dimensional reduction methods for prediction and then use ordinary least squares for combination.

However, selecting the most suitable methods above for certain time series may take considerable amount of time as all methods have to be tried on all time series data. If we have a previous record of the applicability of a certain method on a certain type of time series, we can bypass the training time of current dataset. Thus, we need a way to measure the

similarity of current dataset to predict and the past dataset to infer the suitable method. Recently, (X. Wang, Smith-miles, & R. Hyndman, 2009) pioneered the model selection by matching the characteristics of time series. Those characteristics consist of trend, seasonality, serial correlation, non-linearity, skewness, kurtosis, self similarity, chaotic and periodicity.

In this paper, we extend the work of (X. Wang et al., 2009) by employing the dimensional reduction as a method of feature selection on those characteristics. We also employ decomposition and kernel learning methods as part of our predictors. Furthermore, to anticipate a long time series which may bring about large matrix training, which is usually constructed for machine learning techniques, we utilize clustering and similarity measure to select a subset of the training dataset.

Our proposed method is applied to the M1 dataset assuming that we have the forecasting records of another dataset, which in this case we use M3 Competition dataset. The accuracy of our proposed method is then compared to the prediction result of individual predictor using sMAPE (Symmetric mean absolute percentage error) as the accuracy measures.

The rest of this paper is organized as follows. In section 2 we present brief overview of forecasting method, in section 3 we present our proposed method, in section 4 we describe the experimental setup including the steps taken, dataset and tools used, in section 5 we discuss our findings, and in section 6 we draw our conclusion.

2. Overview of forecasting methods and model selection

2.1 Forecasting methods

We use diverse forecasting methods in the experiment, from simple to sophisticated ones. Simple methods rely on simple formula such as random walk, interpolation or S-curve. Thus, they are quite fast to compute. More sophisticated approach includes ARIMA and Multiple Kernel Learning (MKL).

Random walk sets the output as the last value of the time series, which means no computation involved. This method is used for benchmark. Other method shall yield better performance than the random walk to be considered useful.

Interpolation methods is used to forecast as well as to detrend the time series. For detrending purpose, we use decomposition method as described by (Weron, 2006), where the trend is calculated by polynomial interpolation on moving averaged data after subtracting it with its season, if any.

S-curve or growth curve is popular method for time series that has shape like a letter S. Mathematical function or a model is usually used for this extrapolation. This curve is usually used in modeling the growth of biological species, such as in the work of (Christodoulos, Michalakelis, & Varoutas, 2010) and (Meade & Islam, 1995).

ARIMA is a comprehensive statistical method devised by (Box & Jenkins, 1970) that consist of three parameters, namely order of autoregression (p), degree of first differencing (d) and order of moving average (q). In practice, those parameters need to be estimated by users. However, there is auto.arima function in R by (R. J. Hyndman & Athanasopoulos, 2012) that can alleviate this parameter selection.

Lastly, MKL is a classification & regression method which select suitable kernel for predictors that use kernel, such as Support Vector Regression. A described by (Rakotomamonjy & Bach, 2008), MKL can be seen as an optimization problem of learning both weight of support vector and the kernel weight the same time. Even though it is not a popular forecasting method yet, it has been successfully applied by (X.-rong Zhang, Hu, & Z.-sheng Wang, 2010) and (Yeh, C.-wei Huang, & Lee, 2011) for time series data in economic domain.

2.2 Model selection

Franses (Franses, 2008) stated that the prediction methods that need to be combined are those which contribute significantly to the increased accuracy of prediction. The selection of prediction models in the ensemble is usually

done by calculating the performance of each model toward the hold-out sample.

In addition, (Andrawis, Atiya, & El-Shishiny, 2011) used 9 best models out of 140 models to combine. The combination method used in their study is simple average. Previously, (Armstrong, 2001) stated that only five or six best models are needed to get better prediction result. Our previous study (Widodo & Budi, 2011) on the use of Neural Network for forecast combination also suggests that selecting few best models are crucial for improving the forecasting result.

To select the model, we need to identify the similarities of testing and training dataset. We use the time series characteristics devised by (X. Wang et al., 2009), in which later we reduce them using PCA.

1) Time series characteristics

The first characteristic to consider is trend and seasonality. A trend is identified as a long-term change in the mean level, upward or downward. A smooth nonparametric method can be used to estimate the trend. Meanwhile, seasonality is a repeating pattern over certain intervals of time. It can be identified by calculating autocorrelation coefficient between current and previous lag of time series. Related to seasonality, periodicity is used to examine the cyclic pattern of the time series. Similarly, serial correlation is also used to measure the relationship between a point and itself over various time intervals. The self-similarity of time series is also considered and measured using Hurst exponent.

Other characteristics include nonlinearity which measure nonlinear behavior, skewness which measures the lack of symmetry and kurtosis which captures the peak and tail of the time series. Lastly, the randomness is also used as characteristic and measured by Lyapunov Exponent.

All these characteristics are thoroughly explained by (X. Wang et al., 2009) and the related tools in R are generously provided by them.

2) Dimensional reduction

In this approach, the weight of the individual predictors is substituted with linear transformation of the data provided by a linear projection method, namely the Principal Component Analysis (PCA). To transform an N-dimensional input vector x into a K dimensional output vector z, PCA is defined as follows:

$$z = Wx \tag{1}$$

where $W=[w_1,w_2, \ldots ,w_K]^T$ is the transformation matrix formed by K eigenvectors of the covariance matrix R_x associated with K largest eigevalues. The reduced size vector z is composed of K principal components, which begins from the most important, z_1, until the least importance component, z_K.

The reconstruction of the original vector x, on the basis of principal components and the orthogonal transformation matrix W is described by the relation:

$$x = W^T z \tag{2}$$

The reconstructed vector is deprived of the least important information associated with the reduced eigenvalues of the covariance matrix R_x. The cut information usually corresponds to the noise of the measurements.

3. The proposed method

Our proposed method selects the most appropriate forecasting model based on the similarity of time series characteristics of the testing dataset and the previous dataset. Thus, we assume that there is a database of time series records along with its best predictor. Instead of training all models on the current dataset and pick the best performer, we look on that existing database.

This database can be constructed when we have a set of time series to predict and its corresponding answer such as provided by the past time series competitions. Hence, it can be continually updated and stored in a file, which can be retrieved later.

To look on the most appropriate predictors, we use a classification approach in which the time series characteristics act as the features and the corresponding best methods acts as the classification label. In this way, we do not need to have strict if-then rule which may not be able to find the most similar dataset. KNN (K-Nearest Neighbor), which classifies new cases

based on a similarity measure, such as distance functions, to the closest training samples, is utilized as classifier.

Considering that not all time series characteristics are useful as features for the classification, we opt to use PCA to reduce its dimension. Thus the smaller number of principal component (PC), which acts as features, the better as long as the prediction accuracy does not deteriorate.

4. Experimental Setup

4.1 Steps

The steps to conduct this experiment, as shown in Figure 1, are: (1) read the time series along with their forecast horizon, (2) compute the features using time series characteristics as suggested by (X. Wang et al., 2009), (3) select the most relevant features by using PCA, (4) select the most suitable predictor using classification methods, such as KNN (K-Nearest Neighbor), (5) forecast using each predictor, and (6) record and compare the performance of the prediction. Some forecast methods, such as Random Walk, ARIMA, interpolation and S-curve, can use the univariate time series directly.

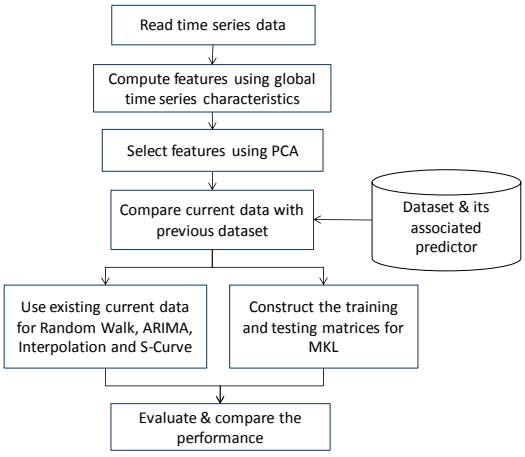

Fig. 1. Steps to evaluate the prediction result of each forecasting method.

For MKL, we need to construct matrices of input and output for testing as it needs to learn the pattern of each part of the time series. The resulting input matrix is scaled and centered so that its standard deviation is equal to one. This scale is then applied to the output vector. The result of the prediction is later converted back to the actual scale before the calculation of its accuracy.

4.2 Dataset

The datasets used in this experiment are from the M1 and M3 Competition[a]. (Theodosiou, 2011) indicates that this series have become a standard test dataset. For testing we use 111 reduced dataset from M1 Competition which consists of 20 annual, 23 quarterly and 68 monthly series. The numbers of future points to forecast are 6, 8, and 18 for yearly, quarterly, and monthly category. The detail of this category can be seen in Table 1.

Table 1. Category of 111 datasets of M1 competition.

Types	Yearly	Quarterly	Monthly	Total
Micro	6	5	22	33
Industry	4	2	21	27
Macro	7	11	17	35
Demographic	3	5	8	16
Total	20	23	68	111

In addition, the M3-Competition dataset is used to construct the look up table consisting of the features of each time series and its predictor. Thus, it acts as a training dataset. This dataset consists of 3003 time series of various type, namely, micro, industry, macro, finance, demographic, and other, as shown in Table 2. The minimum length of observation is 14 points for yearly series, 16 for quarterly, 48 for monthly, and 60 for other series. The numbers of future points to forecast are 6, 8, 18, and 8 for yearly, quarterly, monthly and other category, respectively.

Table 2. Category of M3 datasets, as compiled by (S. Makridakis, 2000).

Types	Yearly	Quarterly	Monthly	Other	Total
Micro	146	204	474	4	828
Industry	102	83	334		519
Macro	83	336	312		731
Finance	58	76	145	29	308
Demographic	245	57	111		413
Other	11		52	141	204

[a] www.forecasters.org/data/m3comp/m3comp.htm

Total	645	756	1428	174	3003

4.3 Matrix construction

For machine learning approach, a matrix needs to be constructed to learn the pattern of the time series. The number of samples that will be used as training and testing is influenced by the length of the time series. Let's suppose that we have 18 values to predict. The vector y_{test} consists of 18 values, and the matrix x_{test} consists of m×18 series, where m is the sliding window. The value of m is determined while constructing the training dataset, namely the x_{train} and y_{train}, whose matrix's size are m×n and n. The shorter the value of m the larger the dataset (which is n) that can be constructed, and vice versa. The example of x_{train} as a sliding window is shown in Figure 2.

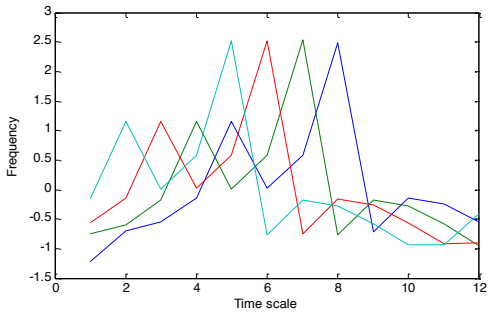

Fig. 2. Example of sliding window of training dataset.

The length of sliding window (m values) is refined further by autocorrelation function (ACF), to obtain the shortest m that still be able to record the time series pattern in a single cycle. This step is necessary since too many lags would increase the error in the forecast while too few could leave out relevant information.

4.4 Iterative forecast

In this experiment, the number of points to predict is more than one, which is commonly referred to as *k-step ahead forecast*, we use an iterative method in which the current predicted outcome will be added to the next testing data. For example, if the point *{x +1}* is predicted by the data *{xk, x-k +1, ... x-3, x-2, x-1, x}*, then the next point *{x +2}* is predicted by the data *{x -k +1, ... x-3, x-2, x-1, x, x+1}*. Recall, that *x+1* is the predicted outcome because the actual data is not yet known, and the data *xk* is deleted so that the number of features, or the size of matrix for testing, does not change. This iterative forecast is especially used for the machine learning approach as it learns based on the data pattern without specified formula as in the statistical methods.

4.5 Performance Evaluation

The accuracy measure used to evaluate the performance of each predictor is the Symmetric mean absolute percentage error (sMAPE), whose equation can be written as:

$$sMAPE = \frac{1}{N}\sum_{t=1}^{n}\frac{|X_t - F_t|}{(|X_t| - |F_t|)/2} \quad (3)$$

where X is the real value and F is the forecast. By using the symmetric MAPE, we avoid the problem of large errors when the actual, X, values are close to zero and the large difference between the absolute percentage errors when X is greater than F and vice versa. sMAPE was the primary criterion for the M3 competition.

5. Result and discussion

5.1 Construction of lookup table

Recall that we assume there is a lookup table consisting of the feature of the time series and its known best predictor. Since we do not have that yet, we must construct it using a sufficiently large dataset with diverse characteristics. The 3003 dataset from M3 Competition is chosen for this purpose. To calculate the characteristics of time series, we use toolbox in R from (X. Wang et al., 2009). Likewise, the toolbox automatic parameter selection of ARIMA in R is available from (R. J. Hyndman & Athanasopoulos, 2012).

The characteristic values of each time series are normalized into range 0 to 1 so that we have equivalent measurement. After running the algorithms of characteristic measures and forecasting methods on that dataset, we get 3003 records of 13 characteristics as features and the corresponding best methods as their classes. The average score of characteristics measures among methods are shown in Table 3.

The MKL method which uses SVR as its predictors is highly preferred than other methods. It is mostly chosen for data having strong seasonal, high auto correlation and non-chaotic. ARIMA is the second method that is mostly chosen. It is more suitable for longer seasonal length, lower trend and non-chaotic.

In addition, we expect that Random Walk would be the least chosen method. But, the fact that is still the third mostly chosen method indicates that about 18% (550/3003) of the dataset is difficult to forecast as Random Walk should be the last resort of predictor. It is mostly preferred by chaotic data, high trend, and strong auto correlation.

Interpolation on the raw data and the trend are not among highly chosen methods. It is good for high kurtosis on de-trended data and for linear data. Lastly, S-curve is the least chosen method which indicates that the M3 dataset on all categories rarely exhibits an S-shaped curve.

Table 3. The average score of characteristics measures among methods.

No	Methods	Count	Raw data									Detrended data			
			seas len	trend	seas	autocorr	nonlin	skew	kurtosis	Hurst	chaotic	autocorr	nonlin	skew	kurtosis
1	Rndm Walk	550	0.02	0.86	0.04	0.81	0.17	0.15	0.10	0.96	0.81	0.38	0.08	0.14	0.28
2	ARIMA	750	0.04	0.70	0.09	0.62	0.12	0.17	0.17	0.88	0.79	0.32	0.07	0.16	0.31
3	Interp	160	0.02	0.86	0.05	0.71	0.12	0.16	0.12	0.94	0.78	0.34	0.08	0.15	0.30
4	Interp trend	122	0.01	0.84	0.05	0.76	0.10	0.16	0.12	0.94	0.80	0.35	0.07	0.16	0.27
5	S-Curve	43	0.02	0.88	0.02	0.81	0.22	0.25	0.18	0.97	0.81	0.32	0.05	0.16	0.28
6	MKL	1378	0.03	0.83	0.15	0.80	0.12	0.15	0.11	0.94	0.76	0.37	0.10	0.14	0.27
	Total	3003													

5.2 Forecasting performance

The average of sMAPE values of each predictor on 111 M1 dataset in Table 4 indicates that MKL is the best predictor compared to the Random Walk, Interpolation, S-curve and ARIMA. Even ARIMA is on par with Random Walk. Thus, in term of achieving good accuracy, MKL alone is sufficient choice.

However, if we could utilize other simpler methods for some time series, we also may achieve efficiency. The model selection ('Modsel' method in Table 4) using the similarity of time series characteristics between the training (111 of M1 competition) and testing dataset (3003 of M3 competition) shows a very close accuracy result.

Furthermore, applying the dimensional reduction on the 13 characteristics of time series to measure the similarity between training and testing dataset ('Modsel+Dimred'), yields the best performance among predictors. The number of PCs used in the testing is determined by the best PCs used in the training. Figure 3 indicates that the sMAPE values on 111 M1 dataset excluding the testing part is pretty well when the number PCs is around 3.

In addition, to reduce the amount of records for training, which is suitable for large dataset, we try the clustering and similarity measure to select a similar subset of those records to the testing dataset. In clustering, we use K-Means to cluster the data. During the training phase we select only cluster that are similar to the dataset. Thus, if we use 2 cluster, we have reduced the dataset into half, which is quite substantial reduction. Likewise, in similarity measure, we select a certain number of records for training that are most similar to the testing dataset. Table 4 indicates that both clustering the data ('MKL+cluster' method if Table 4) and selecting similar records ('MKL+sim') yield only slight decrease in the sMAPE. Thus, it would be a reasonable option for a large dataset.

Table 4. sMAPE values among forecast methods using 111 data from M1 Competition.

No	Methods	Avg	Yearly				Quarterly				Monthly			
			Micro	Ind	Macro	Dem	Micro	Ind	Macro	Dem	Micro	Ind	Macro	Dem
1	Random Walk	19.3	16.1	30.8	22.7	6.7	22.6	11.0	11.5	38.7	23.0	16.9	14.0	23.9
2	ARIMA	19.8	23.6	38.6	27.2	12.0	21.0	10.1	12.1	31.7	17.0	17.7	16.0	30.7
3	Interp	31.9	28.2	32.4	13.0	7.2	38.8	16.7	12.0	50.6	39.3	30.0	31.6	60.6
4	Interp trend	35.2	27.7	37.0	14.5	7.1	67.7	17.2	16.5	63.2	41.7	29.2	32.4	65.1
5	S-Curve	101.7	101.3	101.7	117.4	122.7	85.7	86.8	91.0	93.7	92.0	108.7	105.3	114.2
6	MKL	17.2	17.7	29.0	11.5	5.0	27.0	9.4	10.0	45.4	17.5	15.3	12.3	23.8
7	MKL+cluster	20.5	21.6	29.9	14.0	7.9	24.8	10.1	12.3	54.7	21.3	17.6	16.5	29.3
8	MKL+sim	18.0	18.0	29.1	17.0	6.2	25.4	8.2	10.5	41.8	18.0	17.5	13.5	22.1
9	Modsel	17.7	16.9	29.2	13.4	6.9	20.7	9.5	8.7	48.1	16.2	19.1	16.2	18.1
10	Modsel+Dimred	16.1	22.9	31.6	22.0	9.9	21.8	10.0	9.2	24.7	14.3	15.8	9.5	16.5

Lastly, we observe the forecasting performance on each type of dataset. MKL is good for yearly macroeconomic and demographic data, while ARIMA is better for microeconomic data (yearly, quarterly as well as monthly). Simpler method such as interpolation can be used for yearly and quarterly macroeconomic data. Overall, the yearly demographic dataset is the easiest to predict while the quarterly one is the hardest. Looking at the sMAPE of the model selection method ('Modsel+Dimred'), it successfully select the best predictor for microeconomic (yearly, quarterly and monthly) and monthly macroeconomic, but it is not very good on yearly macroeconomic data.

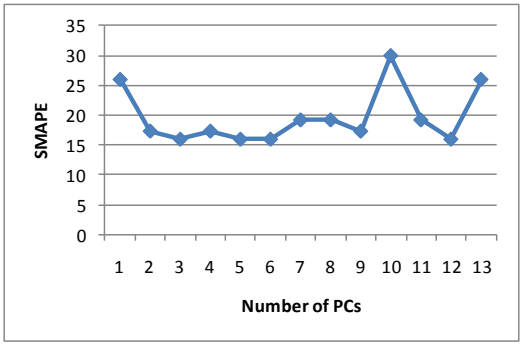

Fig. 3. Optimal number of PCs.

6. Conclusion

Our experimental result indicates that our proposed model selection by reducing the time series features and applying diverse methods, especially the Multiple Kernel Learning, may improve the forecasting accuracy. Our proposed model selection is also beneficial as we do not need to train the current dataset to get the best forecast method, but to use look up table that can be constructed beforehand using previous dataset. In addition, our subset reduction for training is also feasible for large dataset as its performance is very close to the all training data. However, we realize that even though the M1 and M3 time series competition dataset are quite large and diverse in their length, but larger and longer dataset are still needed to confirm the feasibility of our method.

In this paper, we use classification schema to select the best dataset instead of rule based used in (X. Wang et al., 2009). For future works, however, we may try other methods for this selection purpose such as fuzzy inference system. In addition, the characteristics of time series can be added such as by its polynomial degree of curve fitting or can be reduced by other feature selection methods.

References

Andrawis, R. R., Atiya, A. F., & El-Shishiny, H. (2011). Forecast combinations of computational intelligence and linear models for the NN5 time series forecasting competition. *International Journal of Forecasting*, *27*(3), 672-688. Elsevier B.V. doi:10.1016/j.ijforecast.2010.09.005

Armstrong, J. S. (2001). Combining forecasts. *Principles of forecasting: a handbook for researchers and practitioners* (pp. 417-439). Kluwer Academic Publishing.

Bach, F. R., Lanckriet, G. R. G., & Jordan, M. I. (2004). Multiple kernel learning, conic duality, and the SMO algorithm. *Twenty-first international conference on Machine learning - ICML '04*, 6. New York, New York, USA: ACM Press. doi:10.1145/1015330.1015424

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. (Holden-Day, Ed.). San Fransisco.

Cao, L. (2003). Support vector machines experts for time series forecasting. *Neurocomputing*, *51*, 321-339. doi:10.1016/S0925-2312(02)00577-5

Christodoulos, C., Michalakelis, C., & Varoutas, D. (2010). Forecasting with limited data: Combining ARIMA and diffusion models. *Technological Forecasting and Social Change*, *77*(4), 558-565. Elsevier Inc. doi:10.1016/j.techfore.2010.01.009

Crone, S. F., & Kourentzes, N. (2009). Forecasting Seasonal Time Series with Multilayer Perceptrons – an empirical Evaluation of Input Vector Specifications for deterministic Seasonality.

Franses, P. H. (2008). Model selection for forecast combination. *Erasmus*, 1-21.

González-Romera, E., Jaramillo-Morán, M. Á., & Carmona-Fernández, D. (2006). Monthly Electric Energy Demand Forecasting Based on Trend Extraction. *IEEE Transactions on Power System*, *21*(4), 1946-1953.

Huang, C., Yang, D., & Chuang, Y. (2008). Application of wrapper approach and composite classifier to the stock trend prediction. *Expert Systems with Applications*, *34*(4), 2870-2878. doi:10.1016/j.eswa.2007.05.035

Huang, G.-bin, Wang, D. H., & Lan, Y. (2011). Extreme learning machines : a survey, 107-122. doi:10.1007/s13042-011-0019-y

Hyndman, R. J., & Athanasopoulos, G. (2012). *Forecasting: Principles and practice*. Retrieved from http://otexts.com/fpp/

Kourentzes, N., & Crone, S. F. (2008). Automatic modelling of neural networks for time series prediction – in search of a uniform methodology across varying time frequencies. *Proceedings of the 2nd European Symposium on Time Series Prediction (ESTSP'08)*.

Makridakis, S. (2000). The M3-Competition : results , conclusions and implications. *International Journal of Forecasting*, *16*, 451-476.

Makridakis, S. G., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: methods and applications*. New York: John Wiley & Sons.

Meade, N., & Islam, T. (1995). Forecasting with growth curves: An empirical comparison. *International Journal of Forecasting*, *11*(2), 199-215. doi:10.1016/0169-2070(94)00556-R

Poncela, P., Rodríguez, J., Sánchez-Mangas, R., & Senra, E. (2011). Forecast combination through dimension reduction techniques. *International Journal of Forecasting*, *27*(2), 224-237. doi:10.1016/j.ijforecast.2010.01.012

Rakotomamonjy, A., & Bach, F. R. (2008). SimpleMKL. *Journal of Machine Learning Research*, 1-34.

Siwek, K., Osowski, S., & Szupiluk, R. (2009). Ensemble Neural Network Approach for Accurate Load Forecasting in a Power System. *International Journal of Applied Mathematics and Computer Science*, *19*(2), 303-315. doi:10.2478/v10006-009-0026-2

Theodosiou, M. (2011). Forecasting monthly and quarterly time series using STL decomposition. *International Journal of Forecasting*, *27*(4), 1178-1195. Elsevier B.V. doi:10.1016/j.ijforecast.2010.11.002

Varma, M., & Babu, B. R. (2009). More generality in efficient multiple kernel learning. *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, 1-8. New York, New York, USA: ACM Press. doi:10.1145/1553374.1553510

Wang, X., Smith-miles, K., & Hyndman, R. (2009). Rule induction for forecasting method selection : meta-learning the characteristics of univariate time series. *Neural Networks*, 1-34.

Weron, R. (2006). *Modeling and forecasting electricity loads and prices: a statistical approach*. West Sussex, England: John Wiley & Sons, Ltd.

Widodo, A., & Budi, I. (2011). Combination of Time Series Forecasts using Neural Network. *International Conference of Electrical Engineering and Informatics*.

Yeh, C.-yuan, Huang, C.-wei, & Lee, S.-jue. (2011). Expert Systems with Applications A multiple-kernel support vector regression approach for stock market price forecasting q. *Expert Systems With Applications*, *38*(3), 2177-2186. Elsevier Ltd. doi:10.1016/j.eswa.2010.08.004

Zhang, G. P., & Kline, D. M. (2007). Quarterly Time-Series Forecasting With Neural Networks. *IEEE Transaction on Neural Network*, *18*(6), 1800-1814.

Zhang, X.-rong, Hu, L.-ying, & Wang, Z.-sheng. (2010). Multiple Kernel Support Vector Regression for Economic Forecasting. *Science And Technology*, (70872025), 129-134.