J.Int-2015-HJMS_Turkey.pdf

by Budi Warsito

Submission date: 18-Mar-2019 02:00PM (UTC+0700)
Submission ID: 1095202613
File name: J.Int-01-2015-HJMS_warsito.pdf (517.84K)
Word count: 5052
Character count: 23595

Hacettepe Journal of Mathematics and Statistics
Volume 44 (1) (2015), 229 – 238

Wavelet decomposition for time series: Determining input model by using mRMR criterion

Budi Warsito* , Subanar† and Abdurakhman‡

Abstract

Determining the level of decomposition and coefficients used as input in the wavelet modeling for time series has become an interesting problem in recent years. In this paper, the detail and scaling coefficients that would be candidates of input determined based on the value of Mutual Information. Coefficients generated through decomposition with Maximal Overlap Discrete Wavelet Transform (MODWT) were sorted by Minimal Redundancy Maximal Relevance (mRMR) criteria, then they were performed using an input modeling that had the largest value of Mutual Information in order to obtain the predicted value and the residual of the initial (unrestricted) model. Input was then added one based on the ranking of mRMR. If additional input no longer produced a significant decrease of the residual, then process was stopped and the optimal model was obtained. This technique proposed was applied in both generated random and financial time series data.

2000 AMS Classification: 62M10, 65T60

Keywords: time series, MODWT, Mutual Information, mRMR

Received 20/01/2014 : Accepted 22/04/2014 Doi : 10.15672/HJMS.2014117462

*Department of Statistics, Diponegoro University, Semarang, Indonesia
Email: budiwrst2@gmail.com
†Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia
Email:subanar@yahoo.com
‡Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia
Email:rachmanstat@ugm.ac.id

1. Introduction

Wavelet transform for time series analysis has been proposed in many papers in recent years. Previous researches that deserve to be references are in [6] and [10]. Several approaches have been proposed for time series prediction by wavelet transform, as in [12] that used undecimated Haar transform. The choice of Haar transform was motivated by the fact that wavelet coefficients are calculated only from data obtained previously. One of the issues raised in this modeling is the determination of lagged value as an input so that it needs a technique to obtain the optimal input. Input selection aims to select the most relevant input set for a given task. [11] proposed the input selection uses sparse modeling based on a small number of coefficients on each of the signal in autoregressive case, and it is called Multiscale Autoregressive. Wavelet transform used in the method is redundant "*â trous*" wavelet transform which is similar with Maximal Overlap Discrete Wavelet Transform (MODWT) introduced by [10], which has the advantage of being shift-invariant. In this paper, we will utilize Minimal Redundancy Maximal Relevance (mRMR) feature selection technique proposed in [8] to select the scaling and detail coefficients of wavelet decomposition MODWT up to a certain level. Selection criteria used is the Mutual Information that measures the relationship between input variables and output.

Some researches on Mutual Information have been conducted mainly deal with the feature selection as in [4], [13] and [14], while [5] used it for detection of the input time series data and [7] applied for input selection on Wavelet Neural Network Model. On wavelet modeling for time series with mRMR, the initial model is a model formed with only one input, i.e the coefficient of detail or scale generated by MODWT, which has the largest value of Mutual information criterion. Input is then added one by one based on mRMR criteria until the desired amount achieved. Restrictions on the number of coefficients based on the difference of residual are obtained from the addition of the input with the previous model. If there are no significant differences, then the addition is stopped and optimal model is obtained. This paper is organized as follows; Section 2 discusses the wavelet decomposition, especially MODWT; Section 3 discusses the Mutual Information and input selection algorithm with mRMR; and a set of experiments illustrating the method is discussed in Section 4, covers random generate and the real data in financial field.

2. Wavelet Decomposition

Wavelet is a mathematical function that contains certain properties such as oscillating around zero (such as sine and cosine functions) and is localized in time domain, meaning that when the domain value is relatively large, wavelet function will be worth zero. Wavelet is divided into two types, namely father wavelet (ϕ) and mother wavelet (ψ) which has the properties:

$$\int_{-\infty}^{\infty} \phi(x)dx = 1 \quad dan \int_{-\infty}^{\infty} \psi(x)dx = 0 \tag{2.1}$$

Father and mother wavelet will give birth wavelet family by dyadic dilation and integer translation, those are:

$$\phi_{j,k}(x) = \left(2^j\right)^{1/2} \phi\left(2^j x - k\right) \tag{2.2}$$

$$\psi_{j,k}(x) = \left(2^j\right)^{1/2} \psi\left(2^j x - k\right) \tag{2.3}$$

In this case, j is the dilation parameter and k is the translation parameter.

Base wavelet can be seen as a form of dilation and translation with j = 0 and k = 0. Dilation index j and translation index k influence the change of support and range of base wavelet. Dilation index j influences the change of support and range in reverse, i.e if the support is narrow, the range will be widened. The translation index k affects the shift in position on the horizontal axis without changing the width of the support. In this case, the support is closure of the set of points which gives the value of function domain that is not equal to zero. Suppose a mapping belongs to $f : x\epsilon\mathbb{R} \rightarrow y = f(x)\epsilon\mathbb{R}$ then $support(f) = \overline{\{x \,|\, f(x) \neq 0\}}$.

Wavelet function can build a base for $L^2\mathbb{R}$ space, or in other words every function $f\epsilon L^2\mathbb{R}$ can be expressed as a linear combination of a base built by wavelet, and can be written in the following equation.

$$f(x) = \sum_{k\in Z} c_{J,k}\phi_{J,k}(x) + \sum_{j<J}\sum_{k\in Z} d_{j,k}\psi_{j,k}(x) \tag{2.4}$$

where

$$c_{J,k} = \int f(x)\phi_{J,k}(x)dx$$

$$d_{j,k} = \int f(x)\psi_{j,k}(x)dx$$

The transformation in equation (2.4) is *Continue Wavelet Transform* (CWT) in which the wavelet coefficients are obtained through the integration process, so that the value of wavelet must be defined at each $x\epsilon\mathbb{R}$. Another form of transformation is *Discrete Wavelet Transform* (DWT) where the wavelet values are defined only at finite points. Vector containing the values of wavelet is called *wavelet filter* or *detail filter* $\{h_l : l = 0, ..., L-1\}$, where L is the length of the filter and must be an even integer. A detail filter must meet the need of the following three basic properties [10]:

(1) $\sum_{l=0}^{L-1} h_l = 0$ and $\sum_{i=0}^{L-1} g_i^2 = 1$ where L is the length of filter

(2) $\sum_{l=0}^{L-1} h_l^2 = 1$

(3) $\sum_{l=0}^{L-1} h_l h_{l+2n} = \sum_{l=-\infty}^{\infty} h_l h_{l+2n} = 0$

for all nonzero integers n. To fulfill these properties, it is required that the filter length L is even. The required second filter is the *scaling filter* $\{g_l\}$ that corresponds to $\{h_l\}$:

$$g_l \equiv (-1)^{l+1} h_{L-1-l}$$

or in the inverse relationship:

$$h_l = (-1)^l g_{L-1-l}$$

Suppose given wavelet filter $\boldsymbol{h} = (h_0, h_1, ..., h_L - 1)$ while the $\boldsymbol{f} = (f_1, f_2, ..., f_n)$ is a realization of function $\boldsymbol{f}$ on $x_1, x_2, ..., x_n$. In this case, n = 2^J for some positive integer J. DWT can be written as:

$$\mathbf{W} = \mathcal{W}\mathbf{f} \tag{2.5}$$

where $\mathbf{W}$ = result of DWT and $\mathcal{W}$ = transformation matrix with the size nxn. DWT will map the vector $\mathbf{f} = (f_1, f_2, ..., f_n)$ to the coefficient vector $\mathbf{W} = (W_1, W_2, ..., W_J)$ where $\mathbf{W}$ contains wavelet coefficients $d_{j,k}$ and scaling coefficients $c_{J,k}$, for j = 1,2, ..., J. These are an approximation of the coefficients in equation (2.4). DWT can be used to reduce or eliminate the random disturbances in a data (*de-noising* process) by the absence of wavelet coefficients which are quite small. Wavelet coefficients are great values, they

have a major contribution in reconstruction of a function, while the small coefficients contribute negligibly small (essentially zero).

Filtering by DWT as in equation (2.5) cannot be done on any sample size, which cannot be expressed in the form 2^J where J is a positive integer. As an alternative, the calculation of coefficients $d_{j,k}$ and $c_{J,k}$ can be done with *Maximal Overlap Discrete Transform* (MODWT). The advantage of MODWT is that it can eliminate data reduction by half (*down-sampling*), so that in each level there will be wavelet and scaling coefficients as much as length of the data [10]. Suppose a time series data of length N, MODWT transformation will give the column vector $w_1, w_2, ..., w_{J_0}$ and v_{J_0} each of length N.

In order to easily make relations between DWT and MODWT, it is convenient to define an MODWT wavelet filter $\{\tilde{h}_l\}$ through $\tilde{h}_l \equiv h_l/\sqrt{2}$ and scaling filter $\{\tilde{g}_l\}$ through $\tilde{g}_l \equiv h_l/\sqrt{2}$. Wavelet filter and scaling filter from MODWT must fulfill the following conditions:

$$\sum_{l=0}^{L-1} \tilde{h}_l = 0, \sum_{l=0}^{L-1} \tilde{h}_l^2 = \frac{1}{2}, \quad and \quad \sum_{l=-\infty}^{\infty} \tilde{h}_l \tilde{h}_{l+2n} = 0$$

$$\sum_{l=0}^{L-1} \tilde{g}_l = 1, \sum_{l=0}^{L-1} \tilde{g}_l^2 = \frac{1}{2}, \quad and \quad \sum_{l=-\infty}^{\infty} \tilde{g}_l \tilde{g}_{l+2n} = 0$$

In MODWT, the number of wavelet coefficients at each level is always the same, so it is more suitable for time series modeling compared with DWT. Prediction one step forward of time series data $\mathbf{X}$ is modeled linearly, based on coefficients of wavelet decomposition at previous times. Lag of coefficients that will be candidate of input to predict t are detail and scaling coefficients resulted from MODWT transformation in the form $d_{j,t-k}$ and $c_{j,t-k}$ or can be written in the following equation:

$$\hat{X}_t = \sum_{j=1}^{J} \sum_{k=1}^{A_j} (\hat{a}_{j,k} c_{j,t-k} + \hat{b}_{j,k} d_{j,t-k}) \tag{2.6}$$

the J symbol states the level of decomposition, while A_j describes the number of lag coefficients on the level of decomposition. If the number of lag coefficients at each level is the same, $A_j = A$, for each level j, then the number of variables that become candidates of input is 2AJ. Lag of coefficients which serves as inputs of the model will be determined by Minimal Redundancy Maximal Relevance criteria based on Mutual Information.

3. Maximal Relevance Minimal Redundancy

3.1. Entropy and Mutual Information. The entropy of a random variable, denoted by H(X), quantifies an uncertainty present in the distribution of X [2]. It is defined as,

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{3.1}$$

where the lowercase x denotes a possible value that the variable X can adopt from the alphabet $\mathcal{X}$. If the distribution is highly biased toward one particular event $x \in \mathcal{X}$, that is little uncertainty over the outcome, then the entropy is low. If all events are equally likely, that is maximum uncertainty over the outcome, then H(X) is maximum [2]. Following the standard rules of probability theory, entropy can be conditioned on other events. The conditional entropy of X given Y is denoted as follows.

$$H(X|Y) = -\sum_{y \in \mathcal{Y}} p(y) \sum_{x \in \mathcal{X}} p(x|y) \log p(x|y) \tag{3.2}$$

This can be thought as the amount of uncertainty remaining in X after we learn the outcome of Y. The Mutual Information (MI) between X and Y is the amount of information shared by X and Y.

$$MI(X;Y) = H(X) - H(X|Y) = \sum_{x \in X} \sum_{y \in Y} p(xy) \log \frac{p(xy)}{p(x)p(y)} \tag{3.3}$$

This is the difference of two entropies, i.e. the uncertainty before Y is known, H(X), and the uncertainty after Y is known, $H(X|Y)$. This can also be interpreted as the amount of uncertainty in X which is removed by knowing Y. Thus it follows the intuitive meaning of mutual information as the amount of information that one variable provides about another [2]. On the other words, the mutual information is the amount by which the knowledge provided by the feature vector decreases the uncertainty about the class [1]. The Mutual Information is symmetric, MI(X;Y) = MI(Y;X). If the variables are statistically independent, p(xy) = p(x)p(y), the Mutual Information will be zero.

To compute (3.1), we need an estimate of the distribution p(X). When X is discrete this can be estimated by frequency counts from data, $\hat{p}(x) = \frac{\#x}{N}$, the fraction of observations takes on value x from the total N [2]. When at least one of variables X and Y is continuous we need to incorporate data discretization as a preprocessing step. An alternative solution is to use density estimation method [9]. Given N samples of a variable X, the approximate density function $\hat{p}(x)$ has the following form:

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta\left(x - x^{(i)}, h\right) \tag{3.4}$$

where $x^{(i)}$ is the i^{th} sample, h is the window width and $\delta(z, h)$ is the Parzen Window, for example the Gaussian window:

$$\delta(z, h) = \exp\left(-\frac{z^T \sum^{-1} z}{2h^2}\right) \Big/ \left\{(2\pi)^{d/2} h^d \left|\sum\right|^{1/2}\right\} \tag{3.5}$$

where $z = x - x^{(i)}$; d is the dimension of the sample x and Σ is the covariance of z.

3.2. mRMR Selection. The feature selection's goal in terms of mutual information is to find a feature set S with k features x_i which jointly have the largest dependency to the target y called maximum dependency (Max-Dependency).

$$\max D(S, y), \quad D = MI(\{x_i, i = 1, ..., k\}; y) \tag{3.6}$$

To simplify the implementation of (3.6), [8] proposed an alternative way to select features based on maximal relevance (Max-Relevance) and minimal redundancy (Min-Redundancy) criterion. Max-Relevance is to search features satisfying (3.7), which approximates $D(S, y)$ in (3.6) with the mean value of all mutual information values between individual feature x_i and the output y:

$$\max D(S, y), \quad D = \frac{1}{|S|} \sum_{x_i \in S} MI(x_i; y) \tag{3.7}$$

According to Max-Relevance, the features selected could have rich redundancy, i.e., the dependency among these features could be large. If two features highly depend on each other, the respective class-discriminative power would not change much if one of them were removed. Therefore, minimal redundancy (Min-Redundancy) condition is added to select mutually exclusive features:

$$\min D(S), \quad R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} MI(x_i, x_j) \tag{3.8}$$

The criterion combining the eq. (3.7) and (3.8) is called "minimal-redundancy-maximal-relevance" (mRMR). The operator to combine D and R is defined as a simple form to optimize D and R simultaneously:

$$\max \Phi(D, R), \Phi = D - R \tag{3.9}$$

In order to get the near optimal features defined by $\Phi(.)$, incremental search methods can be used. If we have n candidates of input, the first input X is included into the model that has the highest $MI(X, y)$. The remaining input consists of n-1 feature. To determine the next inputs, suppose we already have S_{m-1}, the feature set with m - 1 features. The task is to select the mth feature from the set $X - S_{m-1}$. This is undertaken by selecting the feature that optimizes the following condition:

$$\max_{x_j \in X - S_{m-1}} \left[MI(x_j; y) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} MI(x_j, x_i) \right] \tag{3.10}$$

The main goal of this algorithm is to select a subset of features S from inputs X, which has minimum redundancy and has maximum relevance with the target y (output). Determination of the value of m is based on the difference between prediction accuracy of the model with m inputs and prediction accuracy of the model with m + 1 input. If the difference is smaller than the desired value, then the input is selected as m. This algorithm computes MI(x_i, x_j) and MI(x_i, y), where y is the target (output) and (x_i, x_j) are individual inputs, i.e. all of scaling and detail coefficients from MODWT until level J. Systematically, algorithm for determining the input of wavelet model uses MI as similarity measure are:

(1) Use MODWT to decompose the data up to a certain level in order to obtain detail and scaling coefficients of each level
(2) At each level of decomposition, specify the detail and scaling coefficients that would be candidates of input until a certain lag
(3) Compute the Mutual Information between candidate of input x_i and target y, MI(x_i,y)
(4) Select the initial input x_i so that MI(x_i,y) is the highest MI and $x_i \in S$ then compute MSE of the initial (unrestricted) model
(5) Sort by ascending the remaining input based on mRMR:
    $$\text{mRMR} = \left[MI(x_i, y) - \frac{1}{|S|} \sum_{x_j \in S} MI(x_i, x_j) \right]$$
(6) (a) Add selected input to S based on greatest value of mRMR then calculate the MSE
    (b) Calculate the difference in MSE from the previous model and model with the addition of input
    (c) If the difference is greater than the desired number then back to 6(a)
(7) Process of adding suspended and optimal model is obtained

The desired number in step 6(c) was chosen, the one that was small enough to the initial MSE. The addition of input was stopped when errors no longer decrease significantly. In this paper, the desired number chosen was equal to 1/100000 of MSE of the initial model.

4. Experimental Results

The using of mRMR in wavelet for time series would be applied in three types of data, they are randomly generated data from Autoregressive models, randomly generated data from GARCH model and real data in the financial fields.

4.1. Autoregressive Simulation Data. The data used is randomly generated by AR (2), AR (3) and ARMA (2,1) model of 500 respectively, the following equations are the description:

$$X_t = 1.5X_{t-1} - 0.7X_{t-2} + \varepsilon_t \tag{4.1}$$

$$X_t = 0.8X_{t-1} + 0.4X_{t-2} - 0.7X_{t-3} + \varepsilon_t \tag{4.2}$$

$$X_t = 1.5X_{t-1} - 0.7X_{t-2} + 0.5\varepsilon_{t-1} + \varepsilon_t \tag{4.3}$$

After getting the data generated from random generation, the first step taken is to decompose the data with MODWT up to 4^{th} level to obtain the detail and scaling coefficients of each level. In each level of decomposition, the lags of detail and scaling coefficients are determined as potential inputs. In this case, we choose the coefficients up to lag 16, so there will be 2x4x16 = 128 candidates of input. This value chosen on the ground can accommodate different types of past data that affects the present data. The next stage is to calculate the value of Mutual Information of each candidate and determine the highest MI used as the initial of input. From this result, modeling is executed by ordinary least squares method to obtain prediction values and the residuals.

The next stage is to sort mRMR value of each candidate without lagged value selected as the initial. One by one of the candidates is added into the model sequentially based on mRMR and then calculate the MSE. If additional input does not reduce the previous MSE by at least 1/100000 of the initial MSE, then the process of adding is suspended, and the optimal model is obtained. This stopping criteria is chosen based on the thinking that the decreasing does not affect the difference of MSE significantly. The obtained results are compared with autoregressive models. To calculate the MODWT decomposition, we use the wmtsa toolkit for Matlab while to calculate Mutual Information and mRMR we use MIToolbox package. In each model generated we repeat it five times. The calculation results are presented in table 1.

Based on table 1, it appears that for data generated from linear autoregressive models, wavelet model with MODWT decomposition combined with mRMR procedure to obtain the input always provides a more predictive results than the original models, characterized by the value of both MSE and R square. For the data generated from AR(2), there are two coefficients that are always involved as inputs in wavelet model building, i.e 1^{st} lag of 1^{st} level and 1^{st} lag of 4^{th} level from the scaling coefficients. On the data generated from the AR(3), coefficients that always come up are 1^{st} and 5^{th} lags of 1^{st} level from the scaling coefficients, as well as 1^{st} lag from 4^{th} level. While the data generated from the ARMA(2,1), the 1^{st} lag of the 1^{st}, 2^{nd}, and 4^{th} level from scaling coefficients, respectively, have always become inputs of the model. Meanwhile, for the three data types, the detail coefficients are never entered as input irrespective of levels or lags.

By considering the selected input, the resulting model yields only a few parameter from a lot of candidates. The proposed procedure has succeeded in selecting candidates which have great contribution and dismiss a lot of candidates which are not considered giving significant contribution. This gives a wavelet model for time series with a few coefficients as input and still gives good results.

Table 1. Comparison of MODWT-mRMR with autoregressive models

		autoregressive		mRMR-MODWT			
model	exp	MSE	Rsq	Input	scaling coefficients (level;lags)	MSE	Rsq
AR(2)	1	0.1670	0.6539	3	(1;1,5)(4;1)	0.1309	0.7293
	2	0.2274	0.6759	9	(1;1,6,8)(2;1,11)(3;15)(4;1,9,16)	0.0975	0.8630
	3	0.2046	0.6364	3	(1;1,5)(4;1)	0.1681	0.7019
	4	0.1786	0.6174	3	(1;1,6)(4;1)	0.1604	0.6571
	5	0.1855	0.6439	6	(1;1,12,16)(2;1,6)(4;1)	0.1400	0.7334
AR(3)	1	0.1022	0.6105	4	(1;1,5)(3;3)(4;1)	0.0866	0.6705
	2	0.1017	0.7081	6	(1;1,5)(2;5)(3;6)(4;1,2)	0.0824	0.7649
	3	0.1033	0.6540	10	(1;1,5,7)(2;1,5)(3;4)(4;1,2,7,12)	0.0710	0.7655
	4	0.1118	0.7030	4	(1;1,4,5)(4;1)	0.0852	0.7740
	5	0.1124	0.6402	8	(1;1,5,6)(2;10)(3;3)(4;1,2,15)	0.0667	0.7887
ARMA(2,1)	1	0.0349	0.6918	4	(1;1,5)(3;3)(4;1)	0.0306	0.7311
	2	0.1017	0.7081	6	(1;1,5)(2;5)(3;6)(4;1,2)	0.0824	0.7649
	3	0.0419	0.7362	6	(1;1,6)(2;1)(3;11,16)(4;1)	0.0362	0.7738
	4	0.0421	0.6866	7	(1;1)(2;1,5)(3;6,16)(4;1,5)	0.0333	0.7545
	5	0.0365	0.7424	10	(1;1)(2;1,12)(3;5,6)(4;1,6,8,11,15)	0.0278	0.8070

4.2. GARCH Simulation Data. In this section, randomly generated data with a length of 500 will be discussed, following the ARIMA (0,0,0) as a mean model and GARCH (1,1) as a variant model with the following equation:

$$y_t = 0.00001 + \varepsilon_t \qquad \sigma_t^2 = 0.00005 + 0.8\sigma_{t-1}^2 + 0.1\varepsilon_{t-1}^2 \tag{4.4}$$

Further studies will be conducted with the application of the MODWT using mRMR input selection, for which the data are generated and then carried out a comparative study of the accuracy with GARCH model. We also repeat the experiments for five times and the results obtained are as table (2).

Table 2. MODWT-mRMR Model Comparison with GARCH

	GARCH(1,1)		mRMR-MODWT				
exp	MSE ($x10^{-4}$)	Rsq	Input	scaling (level;lags)	detail	MSE ($x10^{-4}$)	Rsq
1	4.7645	0.9744	4	(1;3,4)(2;2)(3;1)	-	3.0886	0.9926
2	4.3848	0.8739	8	(1;1)(2;1,2,3)(3;1,16)(4;1,5)	-	3.7462	0.9715
3	5.0352	0.9491	4	(1;3)(2;3)(3;1)(4;13)	-	2.9922	0.9717
4	5.4082	0.9383	4	(1;1,2)(2;1)(4;16)	-	2.5345	0.9918
5	4.9960	0.9167	8	(1;1,2,7)(2;1,16)(3;2)(4;2,16)	-	2.7526	0.9847

Calculation results in table (2) indicate that the MODWT with mRMR input selection yields better predictions compared to GARCH model. It is characterized by a smaller value of MSE and R square is greater. Although the lag of selected scaling coefficients are not consistent at a certain value, but it appears that the initial lagged of each level of decomposition dominates the coefficients entrance to the model. As in the random data from AR and ARMA models, in the randomly generated data from GARCH model the coefficients entered into the model are only the scaling coefficients, none of the lag

of detail coefficients is chosen. As mentioned earlier, this procedure was successful for selecting a few coefficients included into the model.

4.3. Applications on the Financial Data. In this section we apply the method proposed in two financial time series data. The first is Quarterly Real Gross Private Domestic Investment from Bureau of Economic Analysis, U.S. Department of Commerce, January 1947 to January 2013 and the second is Monthly Price of the Indonesian Liquified Natural Gas data, from May 2003 to March 2013. The first data can be downloaded from http://research.stlouisfed.org/fred2/, while the second is http://www.indexmundi.com/commodities/. We have investigated that the first data were not stationer and after first order differencing, it would be stationer. The best linear model from the differenced data is AR(1) without constant, and by LM test we found that the residuals have an ARCH effect. The best model for the variance of residuals is GARCH(1,0) by BHHH optimization method.

In the second case, we focused on the monthly change price of the data. Investigation to the type of the data got the best linear model is ARMA(2,2) with constant, and by LM test we found that the residuals have an ARCH effect. The best model for the variance of residuals is GARCH(0,3) by BHHH optimization method. To show the efficiency of the proposed method we analyzed the both data and compared them with the appropriate models. R square value shows the influence of the price data instead of the change. The result is shown in the table (3).

Table 3. MODWT-mRMR Model of the Financial Time Series Data

Real Gross Private Domestic Investment data						
GARCH(1,0)		mRMR-MODWT				
MSE	Rsq	Input	scaling (level;lags)	detail	MSE	Rsq
1517.0429	0.9986	12	(1;1,2,8,10),(2;2,4),(3;3,4,5,7,9),(3;2)	-	1517.5143	0.9978
Indonesian Liquified Natural Gas data						
GARCH(0,3)		mRMR-MODWT				
MSE	Rsq	Input	scaling (level;lags)	detail	MSE	Rsq
40.8585	0.9995	3	(1;2)(4;1,3)	-	42.9082	0.9960

MSE value resulted from the calculation as shown in table (3) explains that in the first case, the proposed method gives result as good as the GARCH model, while in the second, the GARCH model is still superior. In both examined data, as in random generated data, only the scaling coefficients that are included into the model, but none of the detail coefficients is chosen. We can also make a conclusion that the scaling coefficients have dominant influence to the output, while the detail coefficients have almost no significant role. Overall, mRMR technique can be used to determine input of wavelet model for time series efficiently. It can be seen that the number of input selected with mRMR criteria was a few. This procedure has successfully resulted a model which was more parsimonious in the number of parameters and still gave a good description of the observed data.

5. Closing

A technique combining MODWT decomposition and mRMR criterion was proposed for constructing forecasting model for time series. In MODWT for time series we use

a linear prediction based on some coefficients of decomposition of the past values. The mRMR criteria was used as a tool to determine the input. Coefficients which have high values of mRMR were chosen as the input. By this procedure, model resulted just contained coefficients that were considered important enough to gave influence to the present value. The advantage of this technique is opening up the possibility of development by utilizing more sophisticated processing such as Neural Network that results hybrid model, which is called Wavelet Neural Network.

References

[1] R. Battiti, *Using Mutual Information for Selecting Features in Supervised Neural Net Learning*. IEEE Trans. On Neural Networks **5**, 4 (1994), 537-550

[2] G. Brown, A. Pocock, M.J. Zhao, M. Luj'an, *Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection*, Journal of Machine Learning Research **13** (2012), 27-66

[3] C. Ding, H. Peng, *Minimum Redundancy Feature Selection from Microarray Gene Expression Data*, Journal of Bioinformatics and Computational Biology **3**, 2 (2005), 185-205

[4] I. Guyon, A. Elisseeff, *An Introduction to Variable and Feature Selection*, Journal of Machine Learning Research **3** (2003), 1157-1182

[5] J. Hao, *Input Selection using Mutual Information - Applications to Time Series Prediction*, Helsinki University of Technology, MS thesis, Dep. of Computer Science and Engineering (2005)

[6] B. Kozlowski, *Time Series Denoising with Wavelet Transform*, Journal of Telecommunications and Information Technology **3** (2005), 91-95

[7] R.K. Parviz, M. Nasser, M.R.J. Motlagh, *Mutual Information Based Input Variable Selection Algorithm and Wavelet Neural Network for Time Series Prediction*, ICANN 2008, Part I, LNCS 5163 (2008), 798-807

[8] H. Peng, F. Long, C. Ding, *Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy*, IEEE Trans. on Pattern Analysis and Machine Intelligence **27**, 8 (2005), 1226-1238

[9] H. Peng, C. Ding, F. Long, *Minimum Redundancy Maximum Relevance Feature Selection*, IEEE INTELLIGENT SYSTEMS **20**, 6 (2005)

[10] D.B. Percival, A.T. Walden, *Wavelet Methods for Time Series Analysis*, Cambridge University Press, Cambridge, United Kingdom (2000)

[11] O. Renaud, J.L. Starcx, F. Murtagh, *Prediction Based on a Multiscale Decomposition*, Int. Journal of Wavelets, Multiresolution and Information Processing **1**, 2 (2003), 217-232

[12] S. Soltani, *On the Use of the Wavelet Decomposition for Time Series Prediction*, Neurocomputing **48**, (2002), 267-277

[13] G.D. Tourassi, E.D. Frederick, M.K. Markey, C.E. Floyd, Jr, *Application of the Mutual Information Criterion for Feature Selection in Computer-Aided Diagnosis*, Med. Phys **28**, December (2001), 2394-2402

[14] H.H. Yang, J. Moody, *Feature Selection Based on Joint Mutual Information*, Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis (AIDA), Computational Intelligence Methods and Applications (CIMA), International Computer Science Conventions, Rochester New York, (1999), 22-25

J.Int-2015-HJMS_Turkey.pdf

ORIGINALITY REPORT

6%	6%	0%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.jmlr.org Internet Source	6%

Exclude quotes	On	Exclude matches	< 3%
Exclude bibliography	On		