Mach Learn
DOI 10.1007/s10994-014-5460-1

Using causal discovery for feature selection in multivariate numerical time series

Youqiang Sun · Jiuyong Li · Jixue Liu · Christopher Chow · Bingyu Sun · Rujing Wang

Received: 12 October 2013 / Accepted: 16 June 2014

Abstract Time series data contains temporal ordering, which makes its feature selection different from the normal feature selection. Feature selection in multivariate time series has two tasks: identifying the relevant features and finding their effective window sizes of lagged values. The methods extended from normal feature selection methods do not solve this two-dimensional feature selection problem since they do not take lagged observations of features into consideration. In this paper, we present a method using the Granger causality discovery to identify causal features with effective sliding window sizes in multivariate numerical time series. The proposed method considers the influence of lagged observations of features on the target time series. We compare our proposed feature selection method with several normal feature selection methods on multivariate time series data using three well-known modeling methods. Our method outperforms other methods for predicting future values of target time series. In a real world case study on water quality monitoring data, we show that the features selected by our method contain four out of five features used by domain experts,

Editors: Vadim Strijov, Richard Weber, Gerhard-Wilhelm Weber, and Süreyya Ozogur Akyüz.

Y. Sun (✉) · B. Sun · R. Wang
School of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui, China
e-mail: syouq@mail.ustc.edu.cn

Y. Sun · B. Sun · R. Wang
Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui, China

J. Li · J. Liu
School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, SA, Australia

C. Chow
Australian Water Quality Centre, SA Water, Adelaide, SA, Australia

C. Chow
SA Water Centre for Water Management and Reuse, University of South Australia, Adelaide, SA, Australia

Published online: 09 July 2014

and prediction performance on our features is better than that on features of domain experts using three modeling methods.

Keywords Feature selection · Multivariate time series · Causal discovery · Prediction and regression · Granger causality

1 Introduction

There has been a growing interest in time series data mining due to the explosive increase in size and volume of time series data in recent years (Ratanamahatana et al. 2010). Multivariate time series data containing a large number of features becomes more and more common in various applications, such as in biology, multimedia, medicine, finance, and so on. For example, an Electro Encephalogram (EEG) from multiple electrodes placed on the scalp are measured to examine the correlation of genetic predisposition to alcoholism. If the EEG device utilizes 64 electrodes and the sampling rate is 256 Hz (Yoon et al. 2005), the EEG data will come at a rate of 983,040 values per minute with 64 features. Multivariate time series pose many challenges in storage, query, prediction and efficiency. Feature selection removing irrelevant and/or redundant features/variables[1] provides a solution for many challenges.

Prediction is a primary task of time series data mining, which uses known historical values to forecast future values or trends. Feature selection is essential and crucial for accurate predictions (Tsai and Hsiao 2010). However, feature selection in time series data is different from feature selection in normal static data. The target value of the latter only relates to the current values of features, while the target value of the former relates to the values of features in the previous time stamps as well as in the current time stamp.

For example, given 7 feature time series X_1, X_2, ..., X_7 from time t_1, t_2, ..., t_n, and a target time series Y from time t_1, t_2, ..., t_i, our objective is to identify relevant features to predict the remaining values of Y, as shown in Fig. 1. We need to select features that are related to the target Y and the window sizes of these features indicating the effect of historical observations on the values of Y. Therefore, the feature selection in multivariate time series is a two-dimensional problem and contains two tasks: identifying the features and finding the window sizes of the features. The existing approaches of feature selection in time series either consider selecting relevant features while keeping time window sizes invariant (or the same for all features) or selecting suitable windows sizes while maintaining the same set of features (see Sect. 2 for detailed discussions).

The discovery of causal relationships is beneficial for good prediction and interpretation. In many prediction tasks, we need not only the accurate predictive results but also good interpretation of the prediction, i.e. which variables cause the change of the target. When past values of a time series have a significant influence on the future values of another time series, the relationship of the two is called "causality". For accurate and interpretable predictions, it is necessary to identify causal relationships (Shibuya et al. 2009).

In this paper, we adopt causal discovery as a means for feature selection to improve the prediction accuracy of multivariate numerical time series data. We define feature selection in time series as a two-dimensional task, and show that two-dimensional selection produces better predictions than single dimensional selection extended from normal feature selection methods. We make use of the Granger causality (Granger 1969) as a means to uncover

[1] Each variable is regarded as a feature for multivariate time series (Lal et al. 2004). Hence, the terms feature and variable, multidimensional time series and multivariate time series, are interchangeably used throughout this paper.

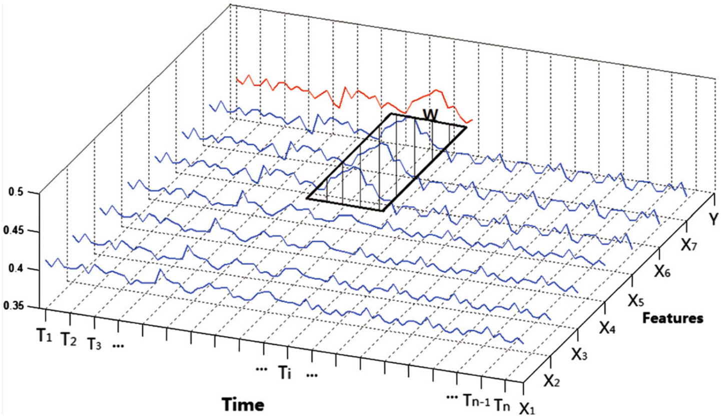

Fig. 1 Time series X in blue color contains 7 variables and Y in *red color* is the target time series. A subset of X as well as the sliding window size w are both needed to predict unknown values of Y (This is a schematic figure. In practice, w of each feature varies, and we will discuss this in Sect. 3.1.)

the causal feature subset because it considers the influence of lagged observations. Granger causality is based on the idea that a cause should be helpful in predicting the future effects, beyond that predicted solely based on their own past values. Therefore, the selected feature subset is not only useful for accurate predictions but also beneficial for interpretations. We have conducted the experiments on three data sets and a real-world case study, in comparison with existing feature selection methods and features selected by domain experts. The experimental results support that our proposed feature selection method is effective.

This work makes the following contributions.

- We make use of the Granger causality model for feature selection in multivariate time series data. In our proposed method, we not only consider causal influence in the feature selection but also consider the window of effective lagged values for each feature.
- We use a real world case study to show that the predictions based on our selected features are better than or the same as predictions based on features selected by domain experts.

The remainder of this paper is organized as follows: Sect. 2 provides background work in feature selection, causal discovery for time series, and a description of Granger causality. Section 3 describes our method using causal discovery for feature selection. Section 4 presents experimental comparisons on several data sets and a real-world case study. Section 5 concludes the paper and points out the future work.

2 Background

In this section, we provide a brief review of feature selection methods and causal discovery approaches in time series. We also discuss the main concept of Granger causality.

2.1 Brief review of feature selection and causal discovery methods in time series

The dimensionality of time series is usually very large. For example, data points of EEG in 1 h from one electrode approximate 1 million. Thus dimensionality reduction is an important processing step before a given time series data set is fed to a data mining algorithm (Chizi and Maimon 2010). Some approaches reduce the dimension by representing the high-dimensional data to a low-dimensional space, such as transformation (Ravi Kanth et al. 1998; Chan and Fu 1999) and aggregation (Keogh et al. 2001; Zhao and Zhang 2006). While feature selection approaches reduce the dimension by selecting a subset of input time series variables based on some measure of feature relevance. In this paper, we are mainly interested in feature selection and the following reviews are about feature selection.

Feature selection methods in time series can commonly fall into the filter and wrapper approaches (Kohavi and John 1997). The filter approach utilizes the data alone to decide which features should be kept. On the other hand, the wrapper approach wraps the feature search around a learning algorithm and utilizes its performance to select the features.

Filter approaches for feature selection are typified by principal component analysis (PCA) based methods (Rocchi et al. 2004; Lu et al. 2007; Zhang et al. 2009). The PCA-based feature selection normally involves two phases: (1) mapping the variables to a lower dimensional feature space by transformation, (2) selecting the features by the computation on principal components. The main problem is that the interpretation of features is difficult after the transformation. Beside that, some other filter approaches have been proposed. A feature selection based on Pearson correlation coefficients was introduced in Biesiada and Duch (2007). But this method is used only for nominal features. In Zoubek et al. (2007), the method uses data driven methods to select relevant features to perform accurate sleep/wake stages classification. In Han and Liu (2013), a filter method uses the mutual information matrix between variables as the features and ranks the variables according to the class separability. Fisher Criterion (FC) is also used in multivariate time series feature selection for classification and regression (Yoon et al. 2005). Filter approaches are flexible, and can be used for any classification and regression models. However, it needs experience to find matched feature selection method and model for the best classification/regression results.

The recursive feature elimination (RFE) is a typical wrapper method for feature selection and was introduced in Support Vector Machine (SVM) classifiers (SVMRFE) (Guyon et al. 2002). It has been used for multivariate time series (Yang et al. 2005; Yoon and Shahabi 2006; Maldonado et al. 2011). The procedure of the RFE-based method can be briefly described as follows: (1) training a classifier using SVM, (2) calculating the ranks of all features by the RFE, and (3) removing the feature with the lowest rank. This procedure is then iteratively performed until the required number of features remains. Main disadvantages of the RFE-based methods include that the number of remaining features is need to specified and the algorithm is time consuming. Beside that, some other wrapper approaches have been proposed. The method in Huang and Wu (2008) utilizes genetic algorithm to prune irrelevant and noisy features for forecasting. Neural network based feature selection methods were present in Crone and Kourentzes (2010), Wong and Versace (2012). These methods identify the feature subset according to a network's predictive performance. But modeling or training a appropriate network is challenging.

The above feature selection methods only consider selecting relevant variables with the same time window size for all variables. In Hido and Morimura (2012), the proposed method chooses the optimal time windows and time lags of each variable using least-angle regression for time series prediction. It considers the temporal effects but maintains the same set of features. Overall, the existing methods do not consider selecting relevant variables and choosing the suitable time windows for variables together. They are one-dimensional feature selection methods. Our proposed method is a two-dimensional feature selection method.

In many prediction tasks, both accuracy of predictions and the knowledge of variables causing the changes in the target variable are important. Causality (also referred to as causation) is commonly defined based on a widely accepted assumption that an effect is preceded by its cause (Haufe et al. 2010). A correlation does not imply causation. Causal discovery is a process to infer the causes of an effect from observational data. A number of approaches have been proposed for causal discovery in time series. Granger causality (Granger 1969) is a well-known method and some of its extensions (Arnold et al. 2007; Lozano et al. 2009; Chen et al. 2004; Eichler 2012) have gained successes across many domains (especially in economics) due to their simplicity, robustness and extendibility (Brovelli et al. 2004; Hiemstra and Jones 1994; Kim 2012; Qiu et al. 2012).

A causal feature selection (short for feature selection based on causal discovery) algorithm solves the feature selection problem by directly or indirectly inducing causal structure and by exploiting formal connections between causation and prediction. Some prior research has been done on the causal feature selection for normal data recently (Aliferis et al. 2010; Guyon et al. 2007; Cawley 2008), but little work has considered causal feature selection for time series.

2.2 Granger causality

The Granger causality (Granger 1969), named after Nobel Prize laureate Clive Granger has been widely used to identify the causal interactions between continuous-valued time series. Granger causality is based on statistical hypothesis test, and the notion is that a cause helps predict its effects in the future, beyond what is possible with auto-regression. More specifically, a variable x is said to Granger cause y, if the auto-regressive model for y in terms of past values of both x and y is significantly more accurate than that based just on the past values of y. For illustration, let x_t and y_t be two stationary time series sequences, x_{t-k} and y_{t-k} are the past k values of x_t and y_t respectively. Then, the Granger causality is performed by firstly conducting the following two regressions:

$$\widehat{y}_{t1} = \sum_{k=1}^{l} a_k y_{t-k} + \varepsilon_t, \tag{1}$$

$$\widehat{y}_{t2} = \sum_{k=1}^{l} a_k y_{t-k} + \sum_{k=1}^{w} b_k x_{t-k} + \eta_t, \tag{2}$$

where $\widehat{y}_{t1}$ and $\widehat{y}_{t2}$ respectively represent fitting values by the first and second regression, l and w are the maximum numbers of lagged observations of x_t and y_t respectively, $a_k, b_k \in \mathbf{R}$ are regression coefficient vectors determined by least squares, ε_t, η_t are white noises (prediction errors). Note that w can be infinity but in practice, due to the finite length of the available data, w will be assumed finite and much shorter than the length of the given time series, usually determined by a model selection method such as Akaike information criterion (AIC) (Akaike 1974) or Bayesian information criterion (BIC) (Schwarz 1978). The F test (or some other similar tests such as conditional variance) is then applied to obtain a p value to indicate whether Eq. (2) results in a better regression model than Eq. (1) with statistically significant advantage. If yes, we call x “Granger causes” y. We denote x causes y by $x \rightarrow y$ means that if and only if x “Granger causes” y but y does not “Granger cause” x.

3 Using causal discovery as a means for feature selection

In this section we first define a two-dimensional feature selection problem for multivariate time series prediction, and then propose a new feature selection method using Granger causality, and finally describe our algorithm.

3.1 Two-dimensional feature selection

As motivated in Fig. 1, we define feature selection in multivariate time series data as a two-dimensional problem, which includes the following two tasks.

1. Identify a variable subset to improve the prediction accuracy of target time series over the prediction on the whole set;
2. Find effective window sizes of these variables which have influence on target value.

Some prior research has demonstrated that causality discovery plays a crucial role in feature selection (Aliferis et al. 2010; Guyon et al. 2007; Cawley 2008). Inspired by this, we explore how to use causality discovery to achieve the tasks.

In our setup, we deal with multidimensional time series sequences $\mathbf{X} = \{X_t^i : t \in T\}_{i=1,\ldots d}$, numerical target time series sequence $Y = \{Y_t : t \in T\}$, where T is the set of all time stamps. The dimension size d of input features is assumed to be large. The main goal is stated as follows: we attempt to find a low dimensional subsequence $\mathbf{X}' = \{X_t^j : t \in T'\}_{j=1\ldots m}$ with m features where $m \ll d$ and $|T'| \ll |T|$. The prediction accuracy of Y using $\mathbf{X}'$ is not worse than using $\mathbf{X}$.

Specifically, we assume numerical target time series Y, multivariate time series $\mathbf{X} = \{X_t^i\}_{t=1,2,\ldots,n;i=1,2,\ldots,d}$, where n is the length of time series and d is the number of time series, i.e. x_1^1 means the value of variable X^1 at time point 1. $\mathbf{X}' = \{X_{t-k}^j\}_{k=0,1,\ldots,w_j;j\in[1,2,\ldots,m]}$ is the feature subset identified from $\mathbf{X}$, where w_j denotes the sliding window size of past observations of X^j and [1, 2, …, m] is the set of indexes of features picked. The result produced by our method is a two-dimensional matrix as following:

$$\begin{pmatrix} x_t^1 & x_t^2 & \cdots & x_t^j & \cdots & x_t^m \\ x_{t-1}^1 & x_{t-1}^2 & \cdots & x_{t-1}^j & \cdots & x_{t-1}^m \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{t-w_j}^1 & x_{t-w_j}^2 & \cdots & x_{t-w_j}^j & \cdots & x_{t-w_j}^m \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{t-w}^1 & x_{t-w}^2 & \cdots & x_{t-w}^j & \cdots & x_{t-w}^m \end{pmatrix} \tag{3}$$

where each column denotes one selected variable and its lagged values. Note that, we assume columns have the same maximum length w in Matrix (3). However, in practical computation, the w_j of each x_j varies. To deal with the various lengths of w_j, we set 0 to elements in the matrix when w_j is greater than its real lagged value and smaller than w.

In many real world multidimensional time series applications, we do not have the historical values of target time series except in the training data. For example, in our experimental data set, ultraviolet data is collected in real time using sensors, but associated chlorine values have to be obtained from a laboratory. In training data, both ultraviolet and chlorine data are collected, but in practice, we wish to use ultraviolet data to derive real time chlorine values. That is why our aim is to predict unknown values of target time series using known values of variables time series.

Now, for training data, we know historical values $y_{t-w}, y_{t-w+1}, \ldots, y_{t-1}$ of y_t, and we know $x_{t-w}^i, x_{t-w+1}^i, \ldots, x_{t-1}^i, x_t^i$ for all features. We will show that our prediction model using Matrix (3) is better than the model using all X_t by predicting y_t.

Firstly, for all features, we can obtain the fitting value $\widehat{y_t}_1$ of y_t using all features X^i at time t in the following equation:

$$\widehat{y_t}_1 = \sum_{i=1}^{d} a_i x_t^i + \varepsilon_t. \tag{4}$$

Alternatively, we can also predict y_t using its historical values as the following:

$$\widehat{y_t}_1 = \sum_{k=1}^{l} b_k y_{t-k} + \eta_t. \tag{5}$$

The above lists two estimations from different aspects. We may not have y_{t-k} in most cases. In our previous example, we have sensor data but we do not have chlorine data. So we will need to use Eq. (4) to estimate the level of chlorine. We assume that they give equally good estimations as our starting point.

$$\sum_{i=1}^{d} a_i x_t^i + \varepsilon_t = \sum_{k=1}^{l} b_k y_{t-k} + \eta_t. \tag{6}$$

We now use this equation to identify which features could derive a better regression. The following equation is obtained after both sides of Eq. (6) are multiplied by m:

$$m \sum_{i=1}^{d} a_i x_t^i + m\varepsilon_t = m\left(\sum_{k=1}^{l} b_k y_{t-k} + \eta_t\right). \tag{7}$$

Adding the same expression $\sum_{i\in[1,\dots,m]} \sum_{j=1}^{w_i} c_j x_{t-j}^i$ on both sides, we obtain:

$$\begin{aligned} &m \sum_{i=1}^{d} a_i x_t^i + \sum_{i\in[1,\dots,m]} \sum_{j=1}^{w_i} c_j x_{t-j}^i + m\varepsilon_t \\ &\quad = m \sum_{k=1}^{l} b_k y_{t-k} + \sum_{i\in[1,\dots,m]} \sum_{j=1}^{w_i} c_j x_{t-j}^i + m\eta_t. \end{aligned} \tag{8}$$

Because the number of elements in $\sum_{i\in[1,\dots,m]}$ is equal to m and independent of y_t, Eq. (8) is transformed to:

$$\begin{aligned} &m \sum_{i=1}^{d} a_i x_t^i + \sum_{i\in[1,\dots,m]} \sum_{j=1}^{w_i} c_j x_{t-j}^i + m\varepsilon_t \\ &\quad = \sum_{i\in[1,\dots,m]} \left(\sum_{k=1}^{l} b_k y_{t-k} + \sum_{j=1}^{w_i} c_j x_{t-j}^i + \eta_t\right). \end{aligned} \tag{9}$$

Then replacing the righthand side of Eq. (9) on the basis of Eq. (2), we obtain the following equation:

$$m \sum_{i=1}^{d} a_i x_t^i + \sum_{i\in[1,\dots,m]} \sum_{j=1}^{w_i} c_j x_{t-j}^i + m\varepsilon_t = \sum_{i\in[1,\dots,m]} \widehat{y_t}_2 = m\widehat{y_t}_2. \tag{10}$$

According to the definition of Granger causality, the prediction error of $\sum_{i\in[1,\dots,m]} \left(\sum_{k=1}^{l} b_k y_{t-k} + \sum_{j=1}^{w_i} c_j x_{t-j}^i + \eta_t\right)$ in Eq. (9) is smaller than that of $m\left(\sum_{k=1}^{l} b_k y_{t-k} + \eta_t\right)$

in Eq. (7). In other words, $\widehat{y_t}_2$ results in a better regression model than $\widehat{y_t}_1$. Note that, there are d features in the leftmost expression $m \sum_{i=1}^{d} a_i x_t^i$ of Eq. (10), but after pruning the non-causal features, the number of remaining causal features is reduced to m. As a result, the final prediction formula extracted from the leftmost side and rightmost side of Eq. (10) can be expressed as following:

$$\widehat{y_t} = \sum_{i \in [1,\ldots,m]} \left(a_i x_t^i + \frac{1}{m} \sum_{j=1}^{w_i} c_j x_{t-j}^i \right) + \varepsilon_t, \tag{11}$$

where x^i ($i \in [1,\ldots,m]$) are the causal features and x_{t-j}^i are their j-lagged values while w_i denotes the sliding window size of each x^i, m is the size of feature subset, a_i and c_j are regression coefficients and ε_t is the prediction error. Therefore, the two tasks we defined at the beginning of this section are integrated by Eq. (11).

3.2 Proposed process and algorithm

We use Granger causality to do feature selection and we make some assumption as Granger causality: (a) no confounding, the variable and the target are not driven by some hidden or common factors simultaneously; (b) the error terms of the variables and the target are normally distributed.

The reasons for using the Granger causality discovery are listed as the following. (a) Granger causality considers the influence of past lagged observations; (b) a time series sequence is normally autocorrelated with its past values which is conformable to the concept of the Granger causality; (c) the features selected by the Granger causality are useful for prediction and interpretation. In general, Granger causality meets the requirements of the two-dimensional feature selection task by uncovering subset and the window size.

The process of our method consists of four steps: (1) transformation for stationary; (2) building an auto-regression model and an augmented auto-regression model for the target; (3) testing the cause and effect relationship between a variable time series and the target time series; (4) repeat the above three steps to remaining variable time series and obtain the feature subset with their window sizes. Next, we describe the process in detail.

Step 1: Transformation for stationary. We test whether a variable time series and the target time series are covariance stationary or not. The common methods include Augmented Dickey–Fuller (ADF) test (Said and Dickey 1984) and Phillips–Perron (PP) test (Phillips and Perron 1988). If the time series is non-stationary, a first (or higher) order difference transformation is utilized to make them stationary. Error correction model (ECM) (Engle and Granger 1987) is used to avoid the spurious regression caused by the transformation.

Step 2: Building an auto-regression model and an augmented auto-regression model for the target series. An univariate auto-regression of the target series is listed as the following.

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_l y_{t-l} + \varepsilon_t.$$

An appropriate number of lags (l) to be included is determined by an information criterion, such as BIC (Schwarz 1978), in our algorithm. Next, we obtain an augmented auto-regression by adding the lagged values of variable X^i:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_l y_{t-l} + b_1 x_{t-1} + \cdots + b_w x_{t-w} + \varepsilon_t.$$

A lagged value of x is retained in the regression if it significantly improves the prediction of y_t by a t test. In the notation of the above regression, w is the lag length of x determined by BIC.

Step 3: Testing the cause and effect relationship between variable X^i and the target Y. The null hypothesis that X^i does not Granger cause Y is accepted if and only if no lagged values of X^i are retained in the augmented regression. Next, we set Y as the variable and X^i as the target to test whether Y Granger causes X^i or not. X^i is selected to add into feature subset when X^i Granger causes Y but Y does not Granger cause X^i (based on the definition of the Granger causality, the latter variables are not helpful to predict the target).

Step 4: Repeat the above three steps to all remaining variable series and obtain feature subset and their window sizes.

Algorithm 1 Feature selection using Granger causality

```
Input: variable time series set $\mathbf{X}$, target time series $Y$, significance level value $p$ and the maximum window
    size $w$;
Output: feature subset $S$;

1: $S \leftarrow \emptyset$;
2: $d \leftarrow$ the number of variables of $X$;
3: $m \leftarrow 0$;  // $m$ is the number of selected variables
4: for $i$=1 to $d$ do
5:    $[F_{X^i \rightarrow y}, w_i] \leftarrow$ Granger_cause($X^i$, $Y$, $p$, $w$);
6:    if $F_{X^i \rightarrow y}$ is 'Yes' then
7:       $[F_{y \rightarrow X^i}, w_j] \leftarrow$ Granger_cause($Y$, $X^i$, $p$, $w$);
8:       if $F_{y \rightarrow X^i}$ is 'No' then
9:          $m \leftarrow m + 1$;
10:          $S \leftarrow S + \{X^i_{w_i}\}$;
11:       end if
12:    else
13:       break;
14:    end if
15: end for
16: return $S$
```

The feature selection algorithm is described in Algorithm 1. Function Granger_cause in Lines 5 and 7 is described in Steps 2 $\sim$ 3 above. $F_{X^i \rightarrow y}$ flags whether X^i Granger causes Y and $F_{y \rightarrow X^i}$ indicates whether Y Granger causes X^i. Note that in our feature selection task, X is the variables time series set and Y is the target time series. We select the variables that cause the target. We set the maximum window size of w to reduce the computation time in regression. To speed up this algorithm, we adopt an early abandon strategy. We continue if the first flag is yes. We break to the next loop when the flag is no. So we do not need to do Granger_cause test for all variable time series.

In our algorithm, we adopt ADF (Said and Dickey 1984) to test covariance stationary of time series. If a series is covariance stationary then it has the deterministic trending mean and variance.

We adopt BIC to determine the lagged values l and w using the following formula (taking w as an example):

$$BIC(w) = n \cdot log(\sigma_\epsilon(w)) + w \cdot log(n), \tag{12}$$

where n is the number of data points in X, $\sigma_\epsilon(w)$ is the regression error variance and calculated by:

$$\sigma_\epsilon(w) = \frac{1}{n} \sum_{t=1}^{n} (y_t - \widehat{y_t(w)})^2,$$

Table 1 A description of data sets

Data sets	# Dimensions	# Instances
Parkinsons telemonitoring	26	5,875
Toms hardware	96	28,179
Communities and crime	123	1,994
Water quality monitoring	218	299,095

where $\widehat{y_t(w)}$ is the prediction value calculated by regression function with different w. We compute the BIC scores using Eq. (12), and select the w that yields the smallest BIC score as the lagged value.

The time complexity of Algorithm 1 is $O(dw)$, where d is the number of variable time series before the feature selection and the w is the maximum search length of window size of lagged observations. The significance level value p is for significance test. The size of feature subset is different for different p values.

4 Experiments and results

In this section, firstly we introduce the data sets used in our experiments. Secondly, we present the evaluation methods and some indicators. Then, we examine and discuss the performances of the proposed approach and some existing non-causal approaches. Finally, we study a real-world case.

4.1 Data sets and experimental setup

In order to evaluate the effectiveness of our method, we conducted experiments on several data sets and study a real-world case. The data sets are described in Table 1. The data sets used in our experiments vary in size and dimension. We have removed the rows with the majority of missing values. The remaining missing values are replaced by the column means. Parkinsons Telemonitoring, Toms Hardware (a data set in Buzz in social media) and Communities and Crime data sets have been downloaded from the UCI Machine Learning data repository (Bache and Lichman 2013). Water Quality Monitoring is a real-world data set. For the first three data sets, we employ the 10 fold cross validation to perform the evaluation. For the real-word data set, the split of training and test data sets will be introduced in the following.

For the real-world time series data set, an on-line Ultraviolet (UV)–Visible (Vis) spectrolyser (S::CAN^{TM}) and a chlorine residual analyser were setup at a drinking water treatment plant to collect data for investigating the relationship between chlorine residuals in water and UV–Vis absorption spectra. Absorption spectra were collected every 2 min with 1 chlorine residual reading (in mg/L). 217 absorbance readings (in m^{-1}) across the range of 200–740 nm by 2.5 nm interval, the absorbance readings against wavelengths are denoted as $A200$, $A202.5$, $A205$, …, $A737.5$, $A740$. We divided the water quality monitoring data set into 4 training sets and 4 test sets, and the way of division is shown in Table 2. In order to reduce the errors caused by time fluctuation, the training and test data sets are matched by the same quarters from 2 years.

All experiments were performed on a machine running 32-bit Windows Operating System with 2.53 GHz processor and 4 GB RAM.

Table 2 A split of water quality monitoring data

No.	# Training set	# Testing set
First quarter	3,600	3,598
Second quarter	3,720	3,719
Third quarter	3,600	3,659
Fourth quarter	3,600	3,599

4.2 Evaluation methods and indicators

To compare the quality of subsets of time series selected by our approach, by comparison approaches, and all variable time series for prediction, we employ several common predictive models including Linear regression, K-Nearest Neighbor (KNN) (Cleary and Trigg 1995), Support vector machine (SVM) (Chang and Lin 2011). These models are exemplars of practical and scalable methods and frequently used in many domains. The use of multiple models in the evaluation removes possible biases of some models with some data sets. The subsets of time series selected by different methods are not the same, so comparisons with the same parameters of a prediction model are not acceptable. For example, for the KNN method, the number of best k varies in different subsets. k was searched from [1, 20] in the experiments; for SVM, different kernel types such as linear and radial basis function kernel types with different parameters suit different subsets. The cost c, gamma g was searched from the combination of [1, 10, 20, 30] and [0, 0.05, 0.1] for the SVM in the experiments. We use the training data sets to search the best parameters, and report the prediction performances on the test data sets using these parameters.

The performances of different selection methods are assessed by the following four criteria:

1. The number of features selected;
2. The proportion of features selected relative to the original number of features;
3. Time for feature selection in seconds;
4. Performances of different predictive models;

For the fourth criterion, we measure the performance of predictive model using the following indexes (the lower the values, the better the performance):

- MAE: mean absolute error;

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \widehat{y_t}|.$$

- RMSE: root mean square error;

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \widehat{y_t})^2}.$$

- RAE: relative absolute error;

$$RAE = \frac{\sum_{t=1}^{n}|y_t - \widehat{y_t}|}{\sum_{t=1}^{n}|y_t - \overline{y}|}.$$

Table 3 The number of, the proportion of selected features and the computation time in different data sets by different methods

Date set	Method	Number	Proportion (%)	Time
Parkinsons telemonitoring	RFE-based	4	15.38	82.26
	PCA-based	5	19.23	2.98
	FC	4	15.38	3.15
	Our method	4	15.38	3.28
Toms hardware	RFE-based	22	22.92	1,028.96
	PCA-based	8	8.33	18.87
	FC	22	22.92	19.08
	Our method	22	22.92	23.88
Communities and crime	RFE-based	42	34.15	7.48
	PCA-based	13	10.57	3.29
	FC	42	34.15	3.44
	Our method	42	34.15	3.95

- RRSE: root relative square error;

$$RRSE = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \widehat{y_t})^2}{\sum_{t=1}^{n}(y_t - \overline{y})^2}}.$$

where n represents the length of time series, y_t means the actual value, $\widehat{y_t}$ is the prediction value of y_t and $\overline{y}$ is the average value of y.

4.3 Comparison with existing methods and original features

We implemented RFE-based (Yoon and Shahabi 2006), PCA-based (Lu et al. 2007), FC (Bishop 1995) and our method in Matlab. The functions of RFE and FC in a machine learning library The Spider (Weston et al. 2005) were employed in our implementation. Then we compare their performances under the criteria listed above.

We use p value = 0.05 as the significance level value in our method. RFE-based and FC methods require the numbers of features to be selected as parameters. For a fair comparison, we use the number of features of our method as the input umber of selected features for the RFE-based and FC methods. For the PCA-based method, the confidence coefficient is set as 95 %.

Firstly, the number of and the proportion of features selected and the computation time are shown in Table 3. The PCA-based method selects fewer features than our method except in data set Parkinsons Telemonitoring which has a low dimensionality. The RFE-based method takes the longest time, especially when the number of instances is large. Other three methods take similar time.

Then, we compare the prediction performances of the selected feature subsets using Linear regression, KNN and SVM methods. The average results of the three methods on the three data sets are summarized in Table 4, where entries with the lowest value in every row are highlighted. "Causality without lag" means that we make predictions using the current values of variables selected by our method, "Causality with lag" means that we make predictions using the current values and their lagged values of variables selected by our method, the

Table 4 Prediction performances using linear regression, KNN and SVM with features selected by different methods

Index	Model	RFE-based	PCA-based	FC	Causality without lag	Causality with lag	Original
MAE	Linear Regression	0.27	0.30	0.29	0.28	**0.24**	0.29
	KNN	0.27	0.29	0.28	0.27	**0.25**	0.28
	SVM	**0.21**	0.27	0.24	0.23	**0.21**	0.24
RMSE	Linear regression	0.38	0.43	0.41	0.39	**0.33**	0.41
	KNN	0.40	0.41	0.41	0.40	**0.37**	0.41
	SVM	0.32	0.38	0.36	0.34	**0.31**	0.37
RAE	Linear regression (%)	50.97	56.59	54.31	52.22	**44.87**	54.61
	KNN (%)	51.35	53.40	52.97	50.90	**47.91**	52.51
	SVM (%)	36.72	44.62	41.20	39.39	**35.64**	40.67
RRS	Linear regression (%)	54.63	61.70	59.31	56.53	**48.35**	59.01
	KNN (%)	58.20	59.28	59.23	57.20	**52.99**	58.95
	SVM (%)	41.67	49.11	45.78	43.58	**39.29**	47.18

"Original" means that we make predictions using the current values of all variables. Our method achieves lower prediction errors than other feature selection methods and the method without a feature selection. Figure 2 visualize the results in Table 4.

Through the experiments and analyses, we have the following observations and conclusions.

1. Our proposed method consistently achieves the best performance among all methods compared, especially in the linear regression and SVM. We note that using features selected by our method but without the lagged values does not improve the prediction performance. This shows the validity of our proposed two-dimensional feature selection problem identifying relevant variables and effective lagged time windows of variables.
2. The SVM is the best predictive method among the three methods, both before and after feature selection. Radial based function is the best-performing kernel when we using all features, however, linear kernel function performs best after feature selection by our method.

4.4 A real-world case study

In this section, we apply our method to a real-world case Water Quality Monitoring. Since the aim of our work is to find a set of possible causal time series of the target time series, one way to evaluate the validity of the selected variables is to compare them with domain expert's choice and compare the prediction performance of our method with that of expert's choice.

217 UV absorbance wavelengths are collected by different scan sensors in this data set. According to the empirical knowledge of domain experts, four transformation of wavelengths: $\frac{A255}{A202.5}$, $\frac{dA290}{d\lambda}$, $\frac{d^2A310}{d\lambda^2}$ and $\ln A350$ have been used to predict the chlorine residual (Byrne et al. 2011). In our experiment, they were labeled as four feature subsets Num 1–Num 4. We combined all wavelengths used in the above features: A202.5, A255, A290, A310 and A350, as a feature subset, labeled as Num 5.

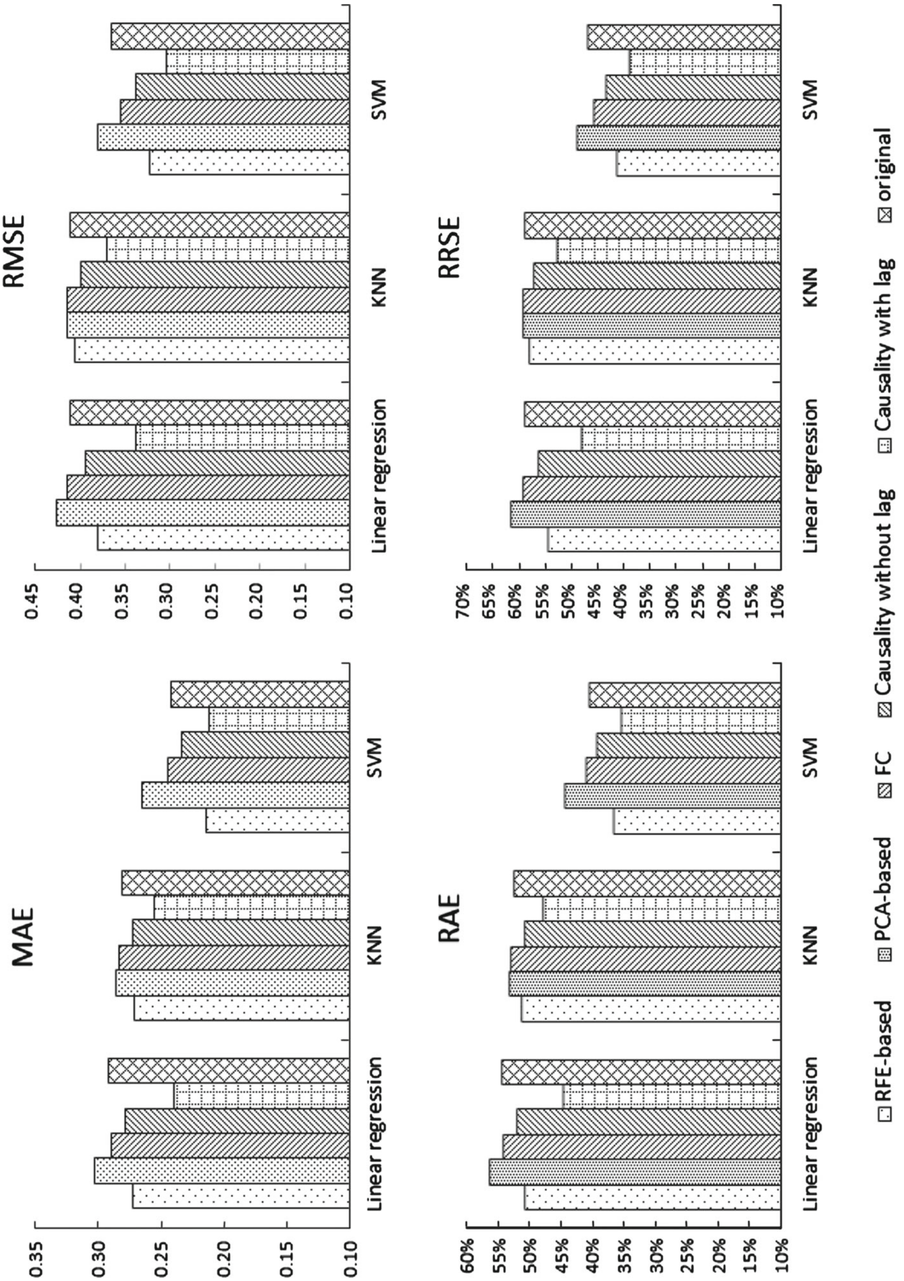

Fig. 2 Prediction performance comparisons of the features selected by different methods and original features using MAE, RMSE, RAE and RRSE

Table 5 The number of, the proportion of selected features and the computation time in the real-world data set by different methods

Method	Number	Proportion (%)	Time
RFE-based	47	21.66	42.02
PCA-based	17	7.83	7.14
FC	47	21.66	7.28
Our method	47	21.66	7.46

The number of and the proportion of features selected by existing methods and our method, and their computation time are shown in Table 5. The results are consistent with observations in the previous subsection.

The prediction performances of feature subsets using Linear regression, KNN and SVM are shown in Table 6 and visualized in Fig. 3, where entries with the lowest error in every row are highlighted. The features are selected by existing methods, our method, without selection and experts, respectively. Our method obtains the lowest prediction errors on the most cases.

The feature subset selected by our method with p value = 0.05 is $\{A202.5 - A257.5\} \cup \{A300 - A357.5\}$, which contains 4 out of 5 wavelengths used by domain experts. The features selected by our method are mainly in the low frequency wavelength end, and this is also consistent with expert's conclusion (Byrne et al. 2011). Our method shows beneficial for finding the potentially causal features for domain experts assuming that they have no knowledge about the data yet, because the range of search is significantly narrowed down.

In conclusion of the experiments, our method achieves the goals of feature selection in multivariate time series: (a) improving the model prediction accuracy; (b) reducing the cost of obtaining and storing time series for predictions; (c) narrowing the candidates for causal-effect relationships between the variables and the target.

5 Conclusions and future work

In this paper, we have presented a feature selection method for multivariate numerical time series using the Granger causality discovery. We define the feature selection in multivariate time series as a two-dimensional problem containing two tasks: selecting features and determining the window sizes of effective lagged values of features. Our proposed method solves the two-dimensional feature selection problem. The features selected by our method also give potential causal relationships between the variables and the target time series. Experimental results on three time series data sets and a real-world case study show that the prediction performance of three modeling methods on features selected by the proposed method is better than that on features selected by other three existing feature selection methods. In the real world case study, the features selected by our method contain four out of five features used by domain experts and the prediction performance on our selected features is better than that on the feature sets used by domain experts.

Feature section based on the non-linear extension of the Granger causality and causal discovery based feature selection in multivariate nominal time series will be two directions of our future work.

Table 6 Prediction performances using linear regression, KNN and SVM with features selected by different methods

Index	Model	RFE-based	PCA-based	FC	Causality without lag	Causality with lag	Original	Num 1	Num 2	Num 3	Num 4	Num 5
MAE	Linear regression	0.28	0.48	0.26	0.26	**0.15**	0.38	0.23	0.27	0.26	0.22	0.25
	KNN	0.22	0.61	0.21	0.34	0.19	0.41	0.56	0.22	0.47	0.21	**0.17**
	SVM	0.28	0.35	0.33	0.37	**0.18**	0.35	0.24	0.22	0.28	0.30	0.20
RMSE	Linear regression	0.31	0.50	0.31	0.30	**0.21**	0.46	0.31	0.30	0.30	0.28	0.28
	KNN	0.30	0.66	0.29	0.40	**0.22**	0.46	0.60	0.26	0.52	0.28	0.25
	SVM	0.35	0.65	0.36	0.52	0.26	0.60	0.32	0.26	0.35	0.37	**0.23**
RAE	Linear regression (%)	69.14	120.52	63.82	65.87	**39.03**	96.18	57.24	68.08	64.98	54.47	62.97
	KNN (%)	56.07	153.06	53.02	86.19	48.30	101.68	139.75	56.17	118.20	52.51	**43.52**
	SVM (%)	71.41	87.67	81.78	91.63	**44.94**	86.59	60.23	55.36	69.79	76.08	49.14
RRSE	Linear regression (%)	69.31	109.83	67.54	66.08	**46.96**	102.44	67.62	67.43	66.90	62.00	62.68
	KNN (%)	66.60	145.24	64.57	89.39	**49.16**	101.04	132.19	56.50	115.16	60.99	54.25
	SVM (%)	77.32	144.53	79.47	114.25	56.74	132.20	71.65	57.76	76.92	81.07	**50.81**

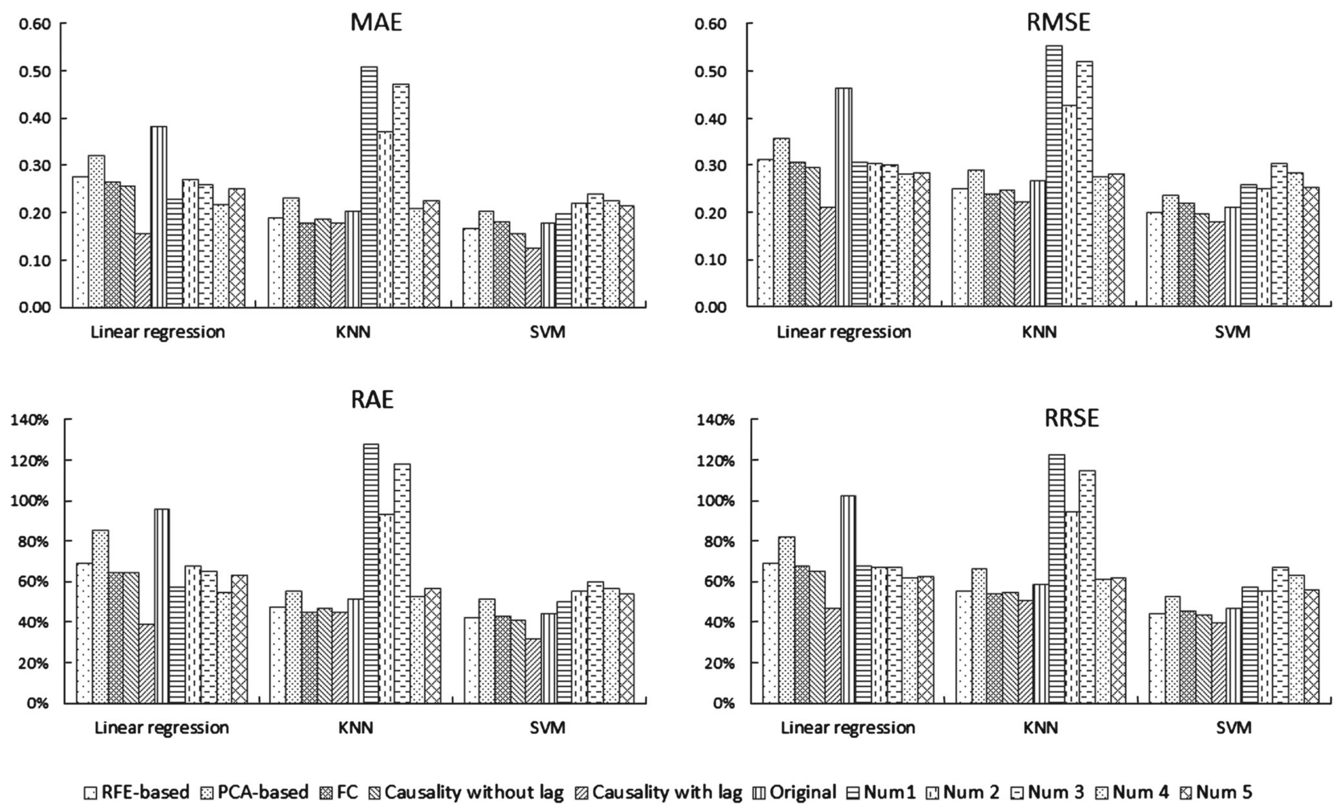

Fig. 3 Prediction performance comparisons of features selected by existing methods, our method, without selection and experts using MAE, RMSE, RAE and RRSE

Acknowledgments The authors would like to thank the SA Water Corporation and the SA Water Centre for Water Management and Reuse for supporting the work. The work has also been partially supported by Australian Research Council Grant DP140103617 and the National Natural Science Foundation of China (No: 31171456).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., & Koutsoukos, X. D. (2010). Local causal and markov blanket induction for causal discovery and feature selection for classification part I: Algorithms and empirical evaluation. *The Journal of Machine Learning Research*, *11*, 171–234.

Arnold, A., Liu, Y., & Abe, N. (2007). Temporal causal modeling with graphical Granger methods. In: *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, ACM* (pp. 66–75).

Bache, K., & Lichman, M. (2013). *UCI machine learning repository*. http://archive.ics.uci.edu/ml.

Biesiada, J., & Duch, W. (2007). Feature selection for high-dimensional dataa pearson redundancy based filter. In: *Computer recognition systems 2* (pp. 242–249). Springer.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Brovelli, A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., & Bressler, S. L. (2004). Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by Granger causality. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(26), 9849–9854.

Byrne, A. J., Chow, C., Trolio, R., Lethorn, A., Lucas, J., & Korshin, G. V. (2011). Development and validation of online surrogate parameters for water quality monitoring at a conventional water treatment plant using a UV absorbance spectrolyser. *The 7th IEEE international conference on intelligent sensors* (pp. 200–204). IEEE: Sensor Networks and Information Processing.

Cawley, G. C. (2008). Causal and non-causal feature selection for ridge regression. *Journal of Machine Learning Research-Proceedings Track*, *3*, 107–128.

Chan, K. P., & Fu, A. C. (1999). Efficient time series matching by wavelets. In *Proceedings of the 15th international conference on data engineering, IEEE* (pp. 126–133).

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 27.

Chen, Y., Rangarajan, G., Feng, J., & Ding, M. (2004). Analyzing multiple nonlinear time series with extended Granger causality. *Physics Letters A*, *324*(1), 26–35.

Chizi, B., & Maimon, O. (2010). Dimension reduction and feature selection. In: *Data mining and knowledge discovery handbook* (pp. 83–100). Springer.

Cleary, J. G., & Trigg, L. E. (1995). K*: An instance-based learner using an entropic distance measure. In: *Proceedings of the international conference on machine learning* (pp. 108–114).

Crone, S. F., & Kourentzes, N. (2010). Feature selection for time series prediction: A combined filter and wrapper approach for neural networks. *Neurocomputing*, *73*(10), 1923–1936.

Eichler, M. (2012). Graphical modelling of multivariate time series. *Probability Theory and Related Fields*, *153*(1–2), 233–268.

Engle, R.F., & Granger, C.W. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica: Journal of the Econometric Society*, 251–276.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424–438.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1–3), 389–422.

Guyon, I., Elisseeff, A., & Aliferis, C. (2007). *Computational methods of feature selection, chapter causal feature selection*. London: Chapman and Hall/CRC.

Han, M., & Liu, X. (2013). Feature selection techniques with class separability for multivariate time series. *Neurocomputing*, *110*, 29–34.

Haufe, S., Nolte, G., Mueller, K.R., & Krämer, N. (2010). Sparse causal discovery in multivariate time series. In *NIPS causality: Objectives and assessment* (pp. 97–106).

Hido, S., & Morimura, T. (2012). Temporal feature selection for time-series prediction. In *21st International conference on pattern recognition (ICPR), IEEE* (pp. 3557–3560).

Hiemstra, C., & Jones, J. D. (1994). Testing for linear and nonlinear Granger causality in the stock price–volume relation. *The Journal of Finance*, *49*(5), 1639–1664.

Huang, S. C., & Wu, T. K. (2008). Integrating ga-based time-scale feature extractions with svms for stock index forecasting. *Expert Systems with Applications*, *35*(4), 2080–2088.

Keogh, E., Chakrabarti, K., Pazzani, M., & Mehrotra, S. (2001). Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and information Systems*, *3*(3), 263–286.

Kim, M. (2012). Time-series dimensionality reduction via Granger causality. *IEEE Signal Processing Letters*, *19*(10), 611–614.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, *97*(1), 273–324.

Lal, T. N., Schroder, M., Hinterberger, T., Weston, J., Bogdan, M., Birbaumer, N., et al. (2004). Support vector channel selection in BCI. *IEEE Transactions on Biomedical Engineering*, *51*(6), 1003–1010.

Lozano, A. C., Abe, N., Liu, Y., & Rosset, S. (2009) Grouped graphical Granger modeling methods for temporal causal modeling. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, ACM* (pp. 577–586).

Lu, Y., Cohen, I., Zhou, X. S., & Tian, Q. (2007). Feature selection using principal feature analysis. In *Proceedings of the 15th international conference on multimedia, ACM* (pp. 301–304).

Maldonado, S., Weber, R., & Basak, J. (2011). Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences*, *181*(1), 115–128.

Phillips, P. C., & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, *75*(2), 335–346.

Qiu, H., Liu, Y., Subrahmanya, N. A., & Li, W. (2012). Granger causality for time-series anomaly detection. In *Proceedings of the 12th IEEE international conference on data mining, IEEE* (pp. 1074–1079).

Ratanamahatana, C. A., Lin, J., Gunopulos, D., Keogh, E., Vlachos, M., & Das, G. (2010). Mining time series data. In *Data mining and knowledge discovery handbook* (pp. 1049–1077). Springer.

Ravi Kanth, K., Agrawal, D., & Singh, A. (1998). Dimensionality reduction for similarity searching in dynamic databases. *ACM SIGMOD Record, ACM*, *27*, 166–176.

Rocchi, L., Chiari, L., & Cappello, A. (2004). Feature selection of stabilometric parameters based on principal component analysis. *Medical and Biological Engineering and Computing*, *42*(1), 71–79.

Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, *71*(3), 599–607.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Shibuya, T., Harada, T., & Kuniyoshi, Y. (2009). Causality quantification and its applications: structuring and modeling of multivariate time series. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, ACM* (pp. 787–796).

Tsai, C. F., & Hsiao, Y. C. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems*, *50*(1), 258–269.

Weston, J., Elisseeff, A., BakIr, G., & Sinz, F. (2005). *SPIDER: object-orientated machine learning library*. http://www.kyb.tuebingen.mpg.de/bs/people/spider.

Wong, C., & Versace, M. (2012). Cartmap: A neural network method for automated feature selection in financial time series forecasting. *Neural Computing and Applications*, *21*(5), 969–977.

Yang, K., Yoon, H., & Shahabi, C. (2005). A supervised feature subset selection technique for multivariate time series. In *Proceedings of the workshop on feature selection for data mining: Interfacing machine learning with statistics* (pp. 92–101).

Yoon, H., & Shahabi, C. (2006). Feature subset selection on multivariate time series with extremely large spatial features. In *Workshops of the 12th IEEE international conference on data mining, IEEE* (pp. 337–342).

Yoon, H., Yang, K., & Shahabi, C. (2005). Feature subset selection and feature ranking for multivariate time series. *IEEE Transactions on Knowledge and Data Engineering*, *17*(9), 1186–1198.

Zhang, M. L., Peña, J. M., & Robles, V. (2009). Feature selection for multi-label naive bayes classification. *Information Sciences*, *179*(19), 3218–3229.

Zhao, Y., & Zhang, S. (2006). Generalized dimension-reduction framework for recent-biased time series analysis. *IEEE Transactions on Knowledge and Data Engineering*, *18*(2), 231–244.

Zoubek, L., Charbonnier, S., Lesecq, S., Buguet, A., & Chapotot, F. (2007). Feature selection for sleep/wake stages classification using data driven methods. *Biomedical Signal Processing and Control*, *2*(3), 171–179.