entropy

MDPI

Article

Short Term Electrical Load Forecasting Using Mutual Information Based Feature Selection with Generalized Minimum-Redundancy and Maximum-Relevance Criteria

Nantian Huang *, Zhiqiang Hu, Guowei Cai and Dongfeng Yang

School of Electrical Engineering, Northeast Dianli University, Jilin 132012, China; huzhiqiang19910714@163.com (Z.H.); caigw@mail.nedu.edu.cn (G.C.); ydfnedu@126.com (D.Y.)
* Correspondence: huangnantian@nedu.edu.cn; Tel.: +86-432-6480-6691

Academic Editor: Raúl Alcaraz Martínez
Received: 21 July 2016; Accepted: 5 September 2016; Published: 8 September 2016

Abstract: A feature selection method based on the generalized minimum redundancy and maximum relevance (G-mRMR) is proposed to improve the accuracy of short-term load forecasting (STLF). First, mutual information is calculated to analyze the relations between the original features and the load sequence, as well as the redundancy among the original features. Second, a weighting factor selected by statistical experiments is used to balance the relevance and redundancy of features when using the G-mRMR. Third, each feature is ranked in a descending order according to its relevance and redundancy as computed by G-mRMR. A sequential forward selection method is utilized for choosing the optimal subset. Finally, a STLF predictor is constructed based on random forest with the obtained optimal subset. The effectiveness and improvement of the proposed method was tested with actual load data.

Keywords: short term load forecasting; generalized minimum redundancy and maximum relevance; random forest; sequential forward selection

1. Introduction

A short-term load forecasting (STLF) predicts future electric loads with a particular prediction limit from one hour extending up to several days. The primary target of smart grids, such as reducing the difference between peak and valley electric loads, large-scale renewable energy absorption, demand side response, and optimal economic operation of the power grid, needs accurate STLF results [1]. In addition, with the development of competitive electricity markets, an accurate STLF is an important basis for drafting a reasonable electricity price and improving the stability of electricity market operation [2].

The existing STLF methods can be divided into traditional methods and artificial intelligence methods. In the traditional methods, such as autoregressive integrated moving average (ARIMA) [3] and regression analysis [4], Kalman filter [5] and exponential smoothing [6] are commonly used. The combination of autoregressive and moving average in ARIMA is a better time series model for STLF [7]. According to the historical time-varying load data, the ARIMA is established and applied for predicting the forthcoming electrical load. The regression analysis uses historical data to establish simple but highly efficient regression models [8]. The Kalman filter improves the accuracy of STLF by estimating each component of load which is apportioned into random and fixed components. The exponential smoothing eliminates the noise in the load time series, and the degree of future load influenced by recent load data can be reflected by adjusting the weight of both data, which is helpful for improving the accuracy of STLF [9]. Overall, the traditional STLF methods can analyze

the linear relationships between input and output, but not the nonlinear relationships [10]. If the load presents large fluctuations caused by environmental factors, the traditional methods may provide inaccurate forecasts.

In recent years, predictors based on artificial intelligence algorithms were widely used in the STLF of power systems [10–17]. Such processes like fuzzy logic [14], expert systems [16,17], artificial neural networks (ANNs) [18,19], and support vector machines (SVMs) [20,21] are currently used in STLF. Fuzzy logic methods divide the input and the output into different kinds of membership functions, and then the relationship between input and output is established by a set of fuzzy rules for fuzzy systems for STLF [22]. However, the fuzzy systems with single if-then rules lack self-learning and adaptive ability to be able to learn the input information effectively. An ANN acquires the complicated non-linear relationship between input and output variables by learning the training samples. However, there is no scientific way of acquiring the optimal network architecture when establishing an ANN model. In addition, it also encounters the problems of falling into local optima and over-fitting [15,23]. SVMs overcome the deficiencies of ANNs by dealing with quadratic programming problems in acquiring the global optimal solution. As compared to an ANN, the SVM has many advantages. However, the SVM parameters, such as the type and variance of the kernel function, and penalty factor, are selected empirically. To achieve the optimal parameters, a SVM combined with genetic and particle swarm optimization algorithm is utilized [24,25]. The random forest (RF) is a combination of classification and regression trees (CARTs) and a bagging learning method. Randomly, by sampling from the training samples and selecting features for splitting node, the RF provides the ability to resist noise and is free from over-fitting problems [26]. Furthermore, in actual practice, there are only two parameters (the tree number and the number of the features for node splitting) that need to be set when RF is applied for STLF [15], making RF highly suitable for STLF.

Considering the effect of various factors, artificial intelligence methods analyze the complicated nonlinear relationships between power load and related factors to achieve higher precision of prediction. However, the features that the predictor employs will influence the accuracy and efficiency of STLF. Therefore, a feature selection schedule should be generated for choosing the optimal feature subset for a predictor. The common features, including historical load, time, and meteorology, are used for STLF modeling [11,27,28]. Historical load can reflect the variation of load accurately, which contains plenty of information. The features of time, such as hour point, day of week, and on/off work day, can also indirectly show the load pattern. In addition, a short-term power load is mainly affected by the changing weather conditions which have a strong correlation with load demand. The accurate meteorological information of the numerical weather prediction (NWP) can improve the accuracy of STLF effectively. Consequently, NWP errors will reduce the accuracy of STLF [29].

A feature selection is a process of choosing the most effective features from an original feature set. The optimal feature subset extracted from a given feature set can improve the efficiency and accuracy of predictor in STLF [30]. Nowadays, the manner of selecting the features has become a hot topic in short-term load forecasting research. Reference [31] adopted conditional mutual information for feature selection. The mutual information values between features and load was measured and subsequently ranked through their values. The first 50 features were used as a threshold parameter for filtering out the irrelevant and weakly relevant features. Reference [10] constructed an original feature set by using the phase space reconstruction theory. The correlation between features and load was analyzed, discovering the optimal feature subset. In reference [29], the mutual information was applied for extracting the effective features from the weather features, as well as, the historical load data features were also extracted for improving the accuracy of holiday load forecasting. Reference [32] used a memetic algorithm to extract a proper feature subset from an original feature set for medium-term load forecasting. Reference [33] analyzed the daily and weekly pattern by autocorrelation function, and chose 50 features as the best features for very short-term load forecasting. The mutual information based on feature selection was used in reference [23]. By calculating the mutual information values between feature vectors and target variable, we can temporarily define a lower boundary criterion

to filter the features. The optimal feature subset with best features was achieved for STLF. All of the researches [10,23,29,31–33] made important contributions to the feature selection in STLF. However, these feature selection methods were just carried out by analyzing the correlation between features and load and the redundancy among these features was not considered.

To improve the accuracy of STLF, the mutual information based on generalized minimum redundancy and maximum relevance feature selection and RF for STLF is proposed. First, an original feature set is formed by extracting historical load features and time features from the original load data. Second, G-mRMR is used for generating the candidate feature, which is ranked in a descending order. Third, the sequential forward selection (SFS) method and a decision criteria based on mean absolute percentage error (MAPE) are utilized for obtaining optimal feature subset by adding one feature at a time to the input feature set of RF. Finally, the RF-based predictor is constructed with the optimal feature subset to achieve the optimal predictor. The proposed method is validated through STLF experiments using the actual load data from a city in Northeast China. The experimental results are compared with different feature selection methods and predictors.

2. Methodology

2.1. Mutual Information-Based Generalized Minimal-Redundancy and Maximal-Relevance

The minimum-redundancy and maximum-relevance (mRMR) is the method which uses mutual information (MI) to measure the dependence between two variables. The MI-based mRMR not only considers the effective information between feature and target variable, but also acquires the repetitive information among features [34]. It has the advantage of obtaining helpful features accurately when dealing with high dimensional data.

Given two random variables X and Y, the MI between them can be estimated as:

$$I(X,Y) = \sum_{X,Y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)} \tag{1}$$

where $P(x)$ and $P(y)$ are the marginal density functions, and $P(x, y)$ is the joint probability density function.

The target of feature selection methods based on MI is finding a feature subset J with n features which reflect the largest dependency on the target variable l from a feature set F_m with m features ($n \ll m$).

The maximum-relevance criterion uses the mean value of MI between feature x_i and target l is described as follows:

$$\max D(J,l),\ D = \frac{1}{|J|}\sum_{x_i \in J} I(x_i,l) \tag{2}$$

The redundancy indicated by MI value describes the overlapping information among features, wherein a larger MI signifies more overlapping information and vice versa. In the process of feature selection, the features selected by maximum-relevance criterion can have more redundancy, and the redundant features have similar information as the prior selected feature cannot improve the accuracy of predictor. Therefore, the redundancy among features should also be evaluated in the process of feature selection.

The minimum-redundancy requires a minimum dependency among each feature:

$$\min R(J),\ D = \frac{1}{|J|^2}\sum_{x_i,x_j \in J} I(x_i,x_j) \tag{3}$$

The mRMR criterion combined with Equations (2) and (3) is computed as follows:

$$\max \psi(D,R),\ \psi = D - R \tag{4}$$

Generally, an incremental search method is used to search for the optimal features [34]. Supposing there is a feature set J_{n-1} with $n-1$ features that has been selected. The aim is to select the nth feature from the rest of set $\{F_m\text{-}J_{n-1}\}$ according to Equation (4). The incremental search method with respect to the condition is as follows:

$$mRMR: \quad \max_{x_j \in F_m\text{-}J_{n-1}} \left[I(x_j, l) - \frac{1}{|J_{n-1}|} \sum_{x_i \in J_{n-1}} I(x_j, x_i) \right] \tag{5}$$

where $|J_{n-1}|$ refers to the number of features in J_{n-1}.

Restructuring Equation (5) by using a weighting factor to balance the redundancy and relevance of feature subset develops into the generalized mRMR (G-mRMR) presented as follows [35]:

$$G\text{-}mRMR: \quad \max_{x_j \in F_m\text{-}J_{n-1}} \left[I(x_j, l) - \alpha \sum_{x_i \in J_{n-1}} I(x_j, x_i) \right] \tag{6}$$

2.2. Random Forest

The random forest (RF) is a kind of machine learning algorithm presented by Leo Breiman, who integrates classification and regression tree (CART) and bagging algorithm [26]. A RF generates many different CARTs by sampling with replacement, wherein each CART achieves one result. The final forecasting result is achieved by computing the average value of all CARTs' results.

2.2.1. CART

The CART employs binary recursive partitioning technology for solving classification and regression issues [36]. A CART, which consists of a root node, non-leaf nodes, branches, and leaf nodes, is shown in Figure 1. Each non-leaf node must be divided according to the Gini index when CART grows.

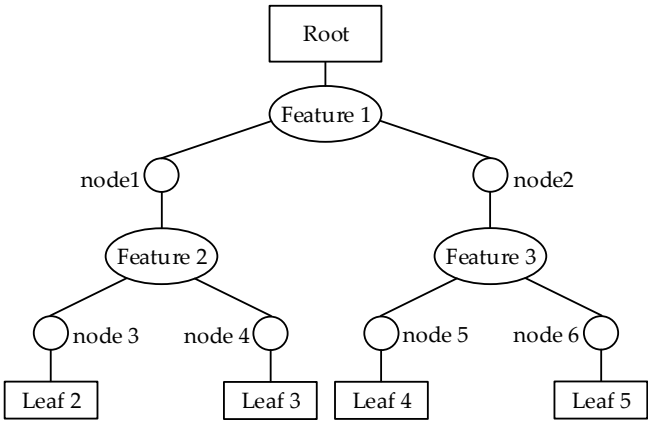

Figure 1. A simple CART.

Supposing there is a dataset D with d samples which includes C classes, the Gini index of D can be defined as:

$$G(D) = 1 - \sum_{i=1}^{C} \left(\frac{d_i}{d} \right)^2 \tag{7}$$

where d_i is the number of ith class.

Afterward, the feature f is used to divide D into D_1 and D_2 subset, wherein the Gini index after the split is:

$$G_{split}(D) = \frac{d_1}{d} G(D_1) + \frac{d_2}{d} G(D_2) \tag{8}$$

2.2.2. Bagging

The bagging is an integrated learning algorithm proposed by Leo Breiman [37]. Given dataset B with M features and learning rule H, a bootstrapping is carried out to generate training sets $\{B^1, B^2, \ldots, B^q\}$. The samples in dataset B may be appraised many times or not at all. A forecasting system consists of a group of learning rule $\{H^1, H^2, \ldots, H^q\}$ which have learned the training set is achieved. Breiman pointed out that bagging can improve the accuracy of predicting the instability of learning algorithms such as CART and ANN [37].

2.2.3. RF

The RF is a group of predictors $\{p(\boldsymbol{x}, \Theta_k), k = 1, 2, \ldots\}$, which is composed of numbers of CARTs, where $\mathbf{x}$ is the input vector and $\{\Theta_k\}$ represents the independent identically distributed random vectors. The modeling process of RF is:

(1) k training sets are sampled with replacement from the dataset B by bootstrap.
(2) Each training set grows up to a tree according to CART algorithm. Supposing dataset B has M features and *mtry* features are randomly selected from B for each non-leaf node. Afterward, the node is split by a feature selected from these *mtry* features.
(3) Each tree grows completely without pruning.
(4) The forecasting result is solved by calculating the mean value of the consequences of each tree predicted.

The flow chart of RF model is illustrated in Figure 2.

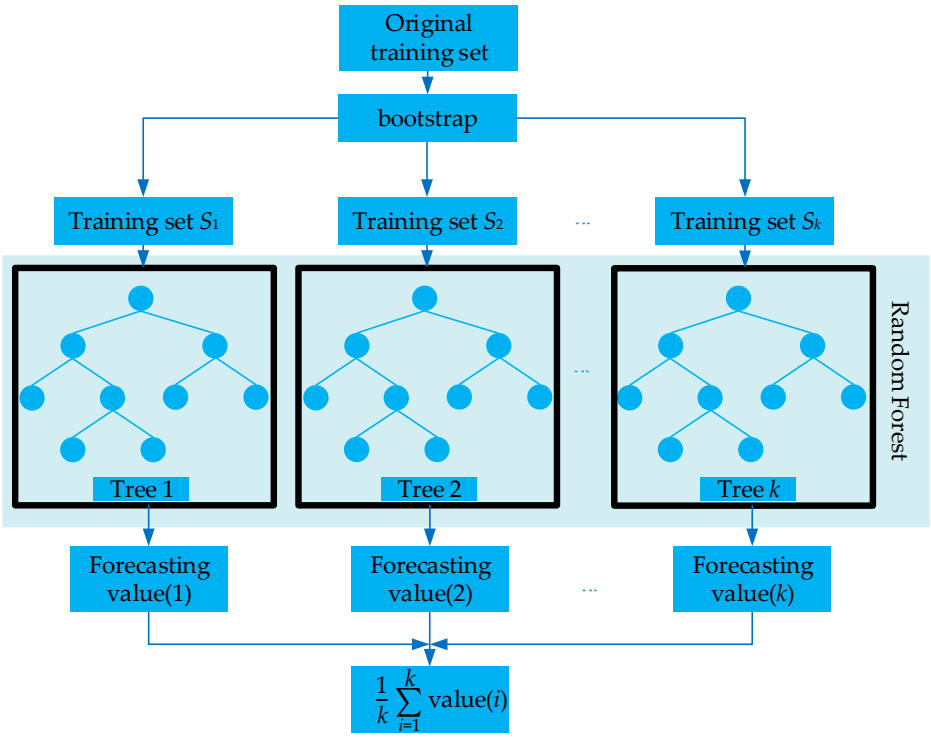

Figure 2. Random Forest modeling and predicting process.

The bagging and the random selection of feature for splitting ensure the good performance of RF, wherein:

(1) The same capacity of the training set sampled by bootstrap guarantees each sample in dataset B to be appraised equally. A situation that one sample may appear many times in the same training set and some may not causes low correlation among the trees.
(2) The manner of selecting feature for node split applies randomness, and ensures the generalized performance of RF.

The number of feature *mtry* and the number of tree of RF *nTree* should be set when applying RF. Generally, *mtry* suggested setting is either $mtry = [\log 2\,(M) + 1]$ or $mtry = \sqrt{M}$ or $mtry = M/3$. The scale of RF generally selected empirically the largest size in order to improve the diversity of trees and guarantee the performance of RF.

3. Data Analysis

The historical load data used in this paper is archived data from a city in Northeast China from 2005 to 2012. As shown in Figure 3a,b, the load demand from 2005 to 2012 increased rapidly with the increase in population and development of the local society. It is difficult to generate a highly accurate STLF in this kind of load pattern. Figure 3c shows the correlation analysis results of the historical load by autocorrelation function [38]. Evidently, the autocorrelation coefficient is reduced gradually along with the increasing of hour lag. According to Figure 3c, the load far from current has low correlation. Only the correlation of the load data from 2011 to 2012 is above the confidence interval which is positive correlation (above of the blue line). With the increasing of the load, the historic load with large lag has very low correlation with the forecasting point. Therefore, we prefer the data from 2011 to 2012 to be used for further research.

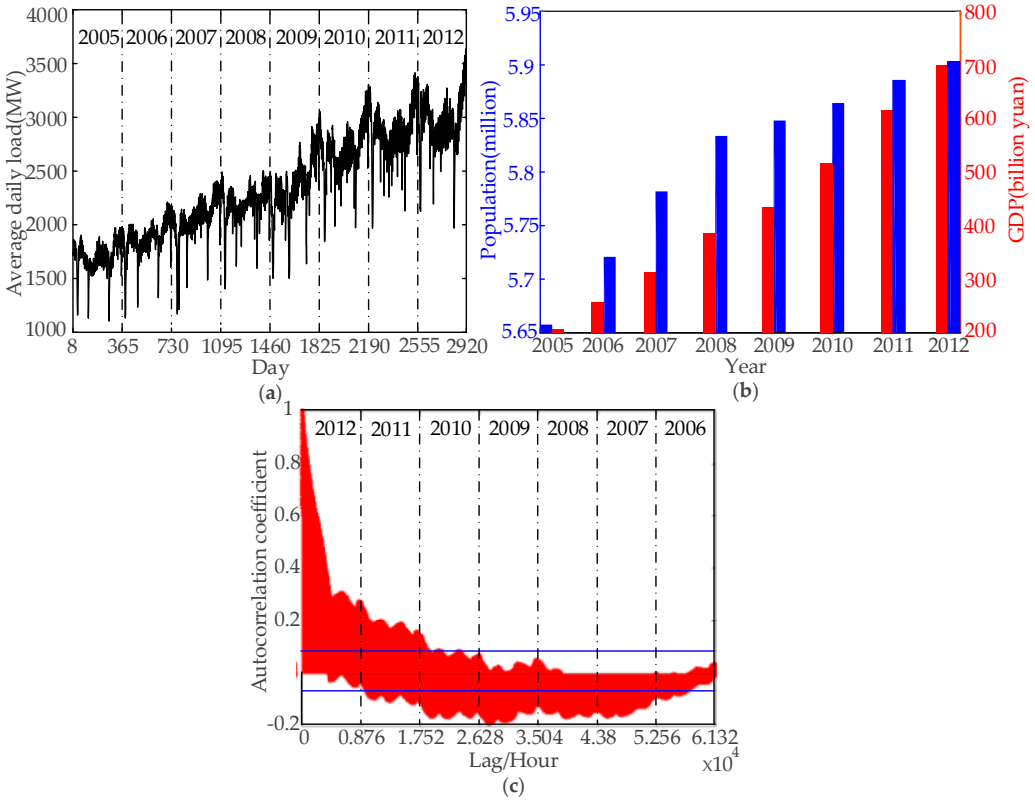

Figure 3. Yearly load curve analysis: (**a**) Average daily load from 8 January 2005 to 31 December 2012; (**b**) The population and GDP from 2005 to 2012; (**c**) Hourly load autocorrelation of historical load data.

Figure 4 shows the average daily load pattern occurring in different seasons. These loads have visibly different patterns which are caused by the varying climate.

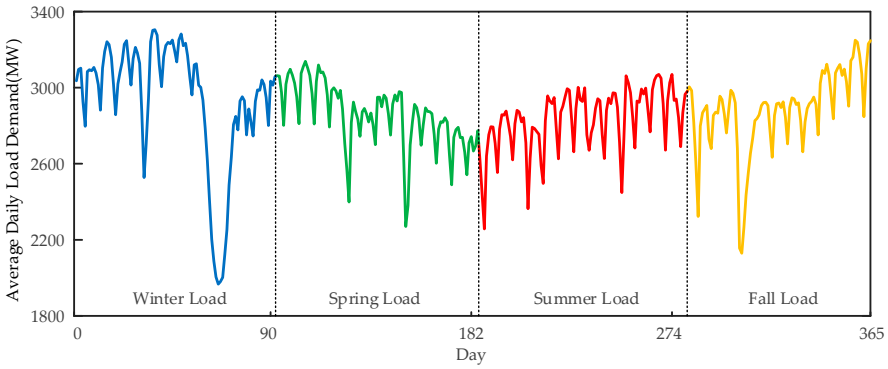

Figure 4. Four seasons average daily load profile from December 2010 to November 2011.

By observing Figure 5, it is possible to know that the load demand presents a kind of cycling mode with a period of 7 days. The load demand from Monday to Friday is similar, whereas on Saturday and Sunday they are dissimilar from each other. This pattern is due to the concurrent changing of load level with the varying electricity consumption behavior of people within a week.

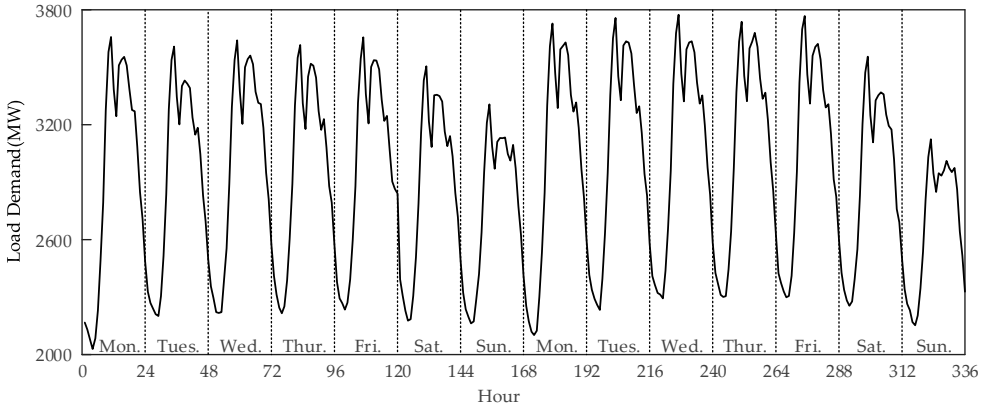

Figure 5. Load curve from 15 to 28 August 2011.

The load point predicts the highly correlated load points similar from the day before as well as relevant with previous week. As shown in Figure 6, the load points throughout the week at lag 1, lag 24, lag 48, lag 72, lag 96, lag 120, lag 144, and lag 168 have strong relevance assuming each lag is 1 h difference. Furthermore, other moment load values also have different dependence.

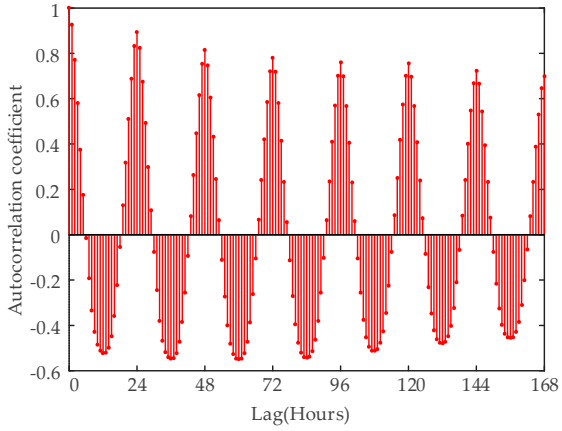

Figure 6. Autocorrelation coefficient of load with 168 lags.

The original feature set for STLF can be achieved based on the above analysis. The 168 load variables $\{L_{t\text{-}168}, L_{t\text{-}167}, \ldots, L_{t\text{-}2}, L_{t\text{-}1}\}$ are extracted as part of original feature set. When doing a day ahead load forecasting, assuming the current moment is t, the load values from the moment t-1 to t-24 are unknown. Therefore, the variables $\{L_{t\text{-}24}, L_{t\text{-}23}, \ldots, L_{t\text{-}1}\}$ are eliminated from the original feature set. In addition, the features, such as hour of day, the day is within weekday or weekend, day of week and season, are considered for constructing the original feature set.

Though meteorological factor affects the load demand, it is not considered in this paper because the error of NWP influences the accuracy of STLF [29]. If needed, the meteorological can be added into the original feature set for feature selection in the same manner. There are 168 features in the original feature set F, as shown in Table 1.

Table 1. The original feature set.

Feature Type	Original Feature
Exogenous features	$1.F_{\text{Hour}}$, $2.F_{\text{WW}}$, $3.F_{\text{DW}}$, $4.F_{\text{S}}$
Endogenous features	$5.F_{\text{L(t-25)}}$, $6.F_{\text{L(t-26)}}$, $7.F_{\text{L(t-27)}}$, $8.F_{\text{L(t-28)}}$, $\ldots$, $146.F_{\text{L(t-166)}}$, $147.F_{\text{L(t-167)}}$, $148.F_{\text{L(t-168)}}$

The meaning of features in Table 1 is:
Exogenous features:

F_{Hour} means the moment of hour, which is tagged by the numbers from 1 to 24.
F_{WW} is either weekday or weekend marked by binary numbers, wherein 0 means weekend and 1 means weekday.
F_{DW} refers to the day of week, which is labeled by the numbers from 1 to 7.
F_{S} uses the numbers from 1 to 4.

Endogenous features:

$F_{\text{L(t-25)}}$ is the load 25 h before, $F_{\text{L(t-26)}}$ means the load 26 h before, and so on.

4. The Proposed Feature Selection Method and STLF Model

A feature selection method combined with G-mRMR and RF is proposed. First, the redundancy of features in F and the relevance between features and load are measured by G-mRMR. Each feature with mRMR value is ranked in a descending order. Afterward, a SFS-based RF is used to search for the optimal feature subset. The MAPE used as a performance index in the feature subset selection process is defined as:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Z_i - \hat{Z}_i}{Z_i}\right| \times 100\% \quad (9)$$

where Z_i is the actual value of load, $\hat{Z}_i$ is the forecasting value, N is the number of sample.

4.1. G-mRMR for Feature Selection

Supposed an original feature set F_m including m features and a selected feature set J. The detail of feature selection process is enumerated below:

(1) Initialization $\varnothing \rightarrow J$.
(2) Compute the relevance between each feature and target variable l. Pick out the feature from F_m which satisfies Equation (2) and add it into J.
(3) Find the feature in the rest of $m-1$ features in F_m that satisfies Equation (4) and add it in to J.
(4) Repeat step (3) until F_m becomes $\varnothing$.
(5) Rank the features in feature set J in descending order in accordance with the measured mRMR value.

4.2. Wrapper for Feature Selection

The common wrapper is a sequential forward and backward selection, both of which do not consider the feature weighting [34,39]. Therefore, the effects of different dimensional features are must be measured, making wrapper a complex and computational feature selection method. According to the result of feature selection of G-mRMR, a wrapper for finding a feature subset can be applied in simpler manner. Considering the features selected by mRMR are ranked in a descending order, the features in the front of the ranking list contain more effective information, thus SFS is used for finding a small feature subset.

A SFS, in which features are sequentially added to an empty candidate set until the addition of another features, does not decrease the criterion. By defining an empty set S and an original feature set F_m, in the first step, the wrapper searched for the feature subset with only one feature, marked as S_1, wherein the feature x_1 selected in S_1 leads to the largest prediction error reduction. In the second step, the wrapper selects the feature x_2 from $\{F_m\text{-}S_1\}$ and combines with S_1 lead to the largest prediction error reduction. The search schedule is repeated until the prediction stops decreasing.

4.3. The Proposed STLF Model

Based on the methods in Sections 4.1 and 4.2, the method of feature selection with RF for STLF is proposed. The feature selection and short-term load forecasting process are shown in Figure 7, where p is the number of feature and α is the weighting factor from 0.1 to 0.9, with an increment of 0.1.

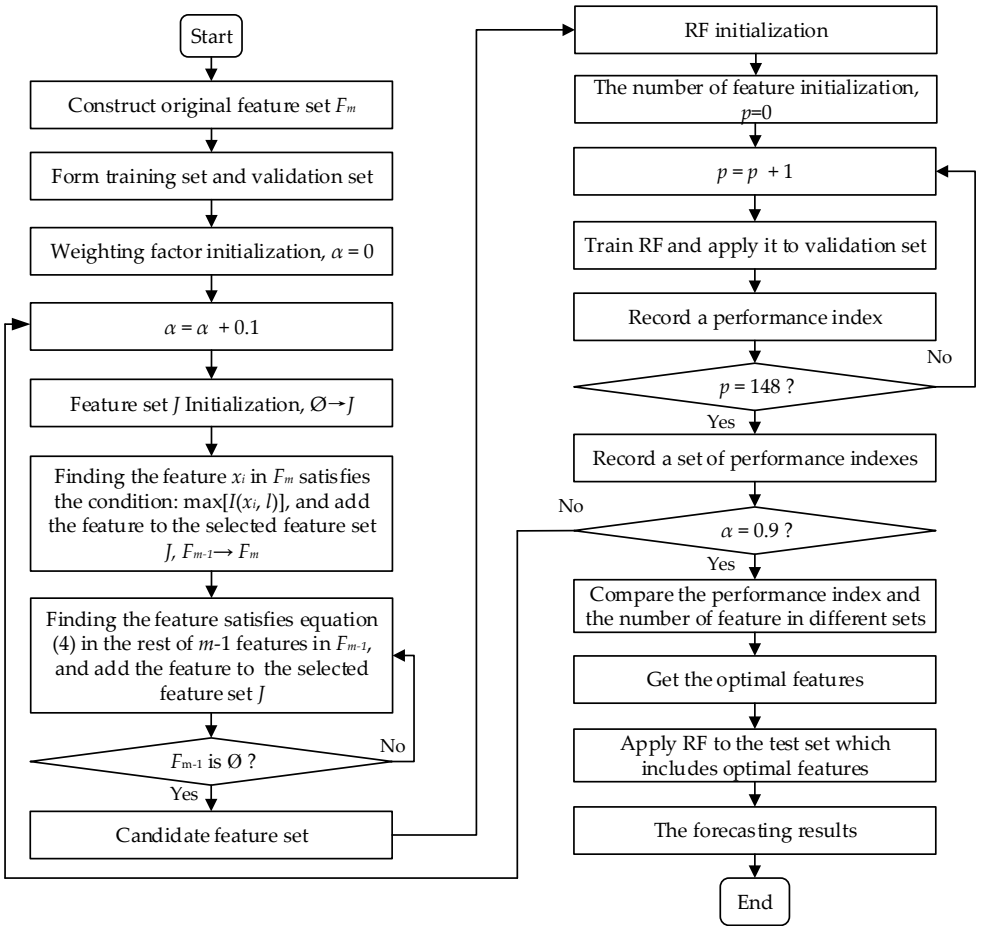

Figure 7. The feature selection process based on G-mRMR and RF for STLF.

5. Case Study and Results Analysis

The data for the experiment consists of the actual data from 2011 to 2012 from a city in Northeast China. For the purpose of feature selection and STLF, the data is divided into three parts: (1) training set (extract eight months randomly from 2011); (2) validation set (the remaining four months of 2011); and (3) test set (extract one week from each season from the data of 2012). More information about the data set is shown in Table 2.

Table 2. Detail information about the data set.

Data Set	Information	Purpose
Training Set	January, February, May, June, August, September, October, December	Train RF
Validation Set	March, April, July, November	Use for obtain the best weighting factor
Test Set	23–29 February 2012 (Winter) 13–19 May 2012 (Spring) 21–27 August 2012 (Summer) 24–30 November 2012 (Fall)	Test performance of RF

The number of variable *mtry*, which RF is not overly sensitive to, is recommended as $mtry = p/3$ [40]. The complexity of RF is affected by the number of tree. Under the premise of non-reduction of prediction accuracy, the initial number of trees *nTree* is set as 500 [15].

Let Equation (9) to be one of the criteria of RF. In addition, the root mean square error (RMSE) is also used. The RMSE is defined in the follow equation:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Z_i - \hat{Z}_i)^2} \tag{10}$$

5.1. Feature Selection Results Based on G-mRMR and RF

In this subsection, the optimal subset is achieved according to the minimum MAPE by setting different weighting factor values of G-mRMR. Figure 8 shows the MAPE curves of the results from RF predictions under different weighting factor α. As shown in Figure 8a, the MAPE is reduced and reaches a minimum value with the increase in the number of feature. Subsequently, it ceases to decrease and gradually increases, indicating that the later addition of features does not improve the performance of RF, but only brings adverse effect. As shown in Figure 8b, the error is reduced rapidly when adopting a small value of α, for instance $\alpha = 0.1$, which indicates that features have useful information for improving the performance of RF. By excessively considering the redundancy among features when using a large value of α, the selected feature subset does not provide enough relevant information for the prediction of RF-based predictor.

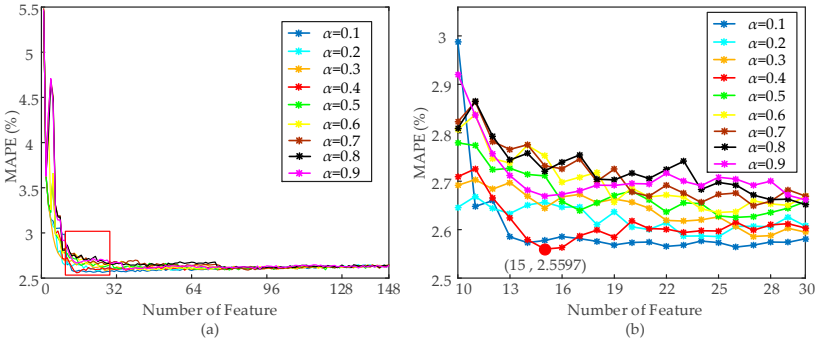

Figure 8. Prediction error curves: (**a**) Prediction error curves corresponding to different weighting factor α; (**b**) The enlarged figure of red box in (**a**).

Table 3 presents the results of feature selection. When α = 0.4, the feature subset has the least number of feature and the RF generates the minimum MAPE. The optimal feature subset is selected.

Table 3. Feature subsets selected by minimum MAPE under different weighting factors.

α	Min MAPE (%)	Number of Features	Feature Subset
0.1	2.5640	26	$F_{L(t-168)}$, $F_{L(t-25)}$, $F_{L(t-48)}$, $F_{L(t-144)}$, $F_{L(t-72)}$, F_{Hour}, $F_{L(t-47)}$, $F_{L(t-26)}$, $F_{L(t-120)}$, $F_{L(t-167)}$, F_{WW}, F_S, F_{DW}, $F_{L(t-34)}$, $F_{L(t-158)}$, $F_{L(t-103)}$, $F_{L(t-27)}$, $F_{L(t-96)}$, $F_{L(t-162)}$, $F_{L(t-132)}$, $F_{L(t-44)}$, $F_{L(t-88)}$, $F_{L(t-149)}$, $F_{L(t-153)}$, $F_{L(t-37)}$, $F_{L(t-107)}$
0.2	2.5857	25	$F_{L(t-168)}$, $F_{L(t-25)}$, $F_{L(t-48)}$, $F_{L(t-144)}$, F_{Hour}, F_S, F_{WW}, $F_{L(t-71)}$, F_{DW}, $F_{L(t-27)}$, $F_{L(t-106)}$, $F_{L(t-162)}$, $F_{L(t-38)}$, $F_{L(t-127)}$, $F_{L(t-93)}$, $F_{L(t-156)}$, $F_{L(t-88)}$, $F_{L(t-32)}$, $F_{L(t-29)}$, $F_{L(t-96)}$, $F_{L(t-44)}$, $F_{L(t-134)}$, $F_{L(t-26)}$, $F_{L(t-166)}$, $F_{L(t-59)}$
0.3	2.5858	27	$F_{L(t-168)}$, $F_{L(t-25)}$, $F_{L(t-48)}$, F_{Hour}, F_{WW}, F_S, $F_{L(t-144)}$, F_{DW}, $F_{L(t-103)}$, $F_{L(t-37)}$, $F_{L(t-162)}$, $F_{L(t-70)}$, $F_{L(t-131)}$, $F_{L(t-28)}$, $F_{L(t-88)}$, $F_{L(t-153)}$, $F_{L(t-106)}$, $F_{L(t-75)}$, $F_{L(t-159)}$, $F_{L(t-34)}$, $F_{L(t-125)}$, $F_{L(t-96)}$, $F_{L(t-43)}$, $F_{L(t-165)}$, $F_{L(t-109)}$, $F_{L(t-31)}$, $F_{L(t-26)}$
0.4	2.5597	15	$F_{L(t-168)}$, $F_{L(t-25)}$, $F_{L(t-48)}$, F_{WW}, F_S, $F_{L(t-127)}$, $F_{L(t-85)}$, $F_{L(t-139)}$, F_{DW}, $F_{L(t-34)}$, $F_{L(t-160)}$, $F_{L(t-70)}$, $F_{L(t-28)}$, $F_{L(t-120)}$, $F_{L(t-141)}$
0.5	2.5897	80	$F_{L(t-168)}$, $F_{L(t-25)}$, $F_{L(t-47)}$, F_{WW}, F_S, $F_{L(t-127)}$, $F_{L(t-86)}$, F_{DW}, $F_{L(t-139)}$, $F_{L(t-35)}$, $F_{L(t-99)}$, $F_{L(t-160)}$, $F_{L(t-69)}$, $F_{L(t-29)}$, $F_{L(t-154)}$, $F_{L(t-120)}$, $F_{L(t-41)}$, $F_{L(t-81)}$, $F_{L(t-133)}$, $F_{L(t-148)}$, $F_{L(t-166)}$, $F_{L(t-32)}$, $F_{L(t-63)}$, $F_{L(t-92)}$, $F_{L(t-26)}$, $F_{L(t-108)}$, $F_{L(t-162)}$, $F_{L(t-78)}$, ...
0.6	2.5868	46	$F_{L(t-168)}$, $F_{L(t-25)}$, F_{Hour}, $F_{L(t-47)}$, F_S, $F_{L(t-127)}$, $F_{L(t-88)}$, F_{DW}, $F_{L(t-156)}$, $F_{L(t-139)}$, $F_{L(t-76)}$, $F_{L(t-34)}$, $F_{L(t-110)}$, $F_{L(t-69)}$, $F_{L(t-149)}$, $F_{L(t-120)}$, $F_{L(t-41)}$, $F_{L(t-81)}$, $F_{L(t-27)}$, $F_{L(t-165)}$, $F_{L(t-37)}$, $F_{L(t-162)}$, $F_{L(t-98)}$, $F_{L(t-30)}$, $F_{L(t-131)}$, $F_{L(t-159)}$, $F_{L(t-104)}$, $F_{L(t-44)}$, ...
0.7	2.5891	88	$F_{L(t-168)}$, $F_{L(t-25)}$, F_{WW}, F_S, $F_{L(t-103)}$, $F_{L(t-61)}$, $F_{L(t-139)}$, F_{DW}, $F_{L(t-47)}$, $F_{L(t-160)}$, $F_{L(t-82)}$, $F_{L(t-124)}$, $F_{L(t-30)}$, $F_{L(t-93)}$, $F_{L(t-156)}$, $F_{L(t-41)}$, $F_{L(t-146)}$, $F_{L(t-33)}$, $F_{L(t-110)}$, $F_{L(t-72)}$, $F_{L(t-152)}$, $F_{L(t-164)}$, $F_{L(t-27)}$, $F_{L(t-90)}$, $F_{L(t-131)}$, $F_{L(t-39)}$, $F_{L(t-118)}$, $F_{L(t-77)}$, ...
0.8	2.6046	93	$F_{L(t-168)}$, $F_{L(t-25)}$, F_{WW}, F_S, $F_{L(t-103)}$, $F_{L(t-61)}$, $F_{L(t-139)}$, F_{DW}, $F_{L(t-47)}$, $F_{L(t-160)}$, $F_{L(t-82)}$, $F_{L(t-124)}$, $F_{L(t-30)}$, $F_{L(t-93)}$, $F_{L(t-156)}$, $F_{L(t-41)}$, $F_{L(t-146)}$, $F_{L(t-33)}$, $F_{L(t-110)}$, $F_{L(t-166)}$, $F_{L(t-75)}$, $F_{L(t-152)}$, $F_{L(t-90)}$, $F_{L(t-72)}$, $F_{L(t-44)}$, $F_{L(t-131)}$, $F_{L(t-28)}$, $F_{L(t-39)}$, ...
0.9	2.5918	35	$F_{L(t-168)}$, $F_{L(t-25)}$, F_{WW}, F_S, $F_{L(t-103)}$, $F_{L(t-67)}$, $F_{L(t-133)}$, F_{DW}, $F_{L(t-34)}$, $F_{L(t-160)}$, $F_{L(t-46)}$, $F_{L(t-148)}$, $F_{L(t-96)}$, $F_{L(t-29)}$, $F_{L(t-84)}$, $F_{L(t-140)}$, $F_{L(t-39)}$, $F_{L(t-153)}$, $F_{L(t-75)}$, $F_{L(t-114)}$, $F_{L(t-165)}$, $F_{L(t-56)}$, $F_{L(t-122)}$, $F_{L(t-62)}$, $F_{L(t-155)}$, $F_{L(t-126)}$, $F_{L(t-41)}$, $F_{L(t-119)}$, ...

The RF will do poor forecasting with less trees, while excessive trees will make it a complicated predictor. In order to obtain a reasonable number of trees of RF, an experiment is designed as follows:

(1) The training set and test set with optimal features are used for the experiment.
(2) The initial number of tree *nTree* = 1.
(3) Training RF and testing with different *nTree* value with increment of 1 until *nTree* = 500.

The experimental result is shown in Figure 9.

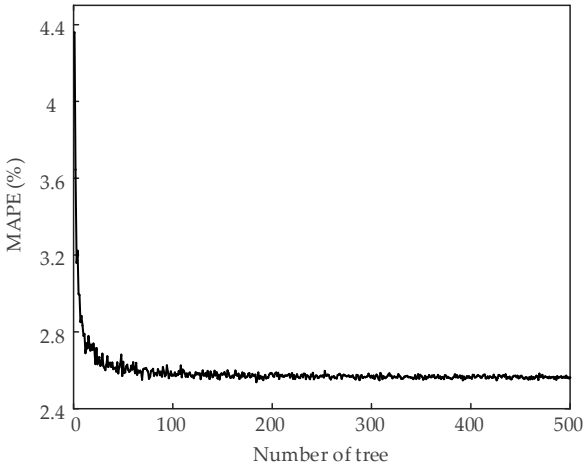

Figure 9. Correlation between tree number and prediction of RF.

The prediction error decreases with the increasing number of tree. When *nTree* > 100, the error tends to be steady. By analyzing the result, *nTree* = 184 with minimum MAPE = 2.5389% is obtained, using this number of trees as the parameter of RF in the future experiment.

5.2. Comparison Experiments for STLF

The data shown in Table 2 are used in the comparison of experiments.

5.2.1. Comparison of Different Feature Selection Methods

By using RF as the predictor, the feature selection methods such as Pearson Correlation Coefficient (PCC), MI, and SFS, are compared with the proposed method for estimating the effect of feature selection of G-mRMR. The results of these feature selection methods are presented in Figure 10.

In Figure 10, with the same predictor, the SFS provides the best performance, followed by G-mRMR (α = 0.4) and MI, and finally the PCC. The SFS, which convolves with RF, selects 22 features and achieves the minimum MAPE = 2.4925%. Considering the relevance between feature and load and the redundancy among features, G-mRMR (α = 0.4) selects 15 features with the minimum MAPE = 2.5597%. The feature subset selected by MI, which does not consider the redundancy among features, is higher than G-mRMR (α = 0.4). Only the PCC analyzes the linear relation between features and load, however the feature subset selected through this method is not as good as G-mRMR (α = 0.4).

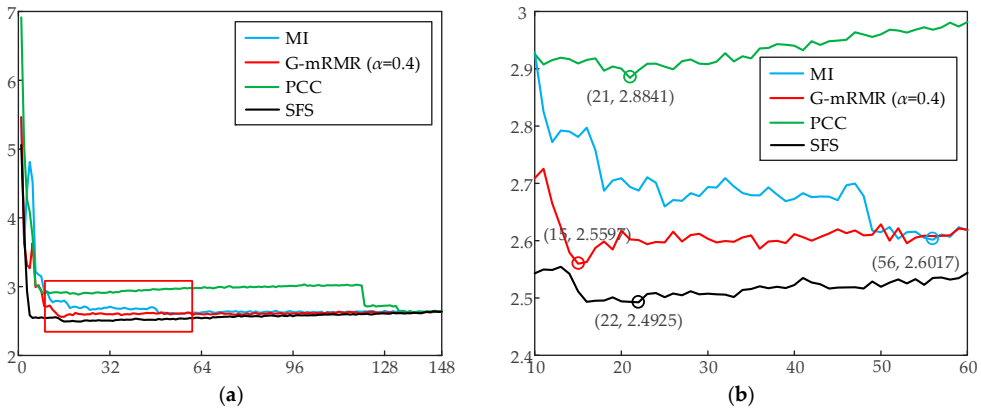

Figure 10. Prediction error curves: (**a**) Prediction error curves corresponding to different feature selection methods; (**b**) The enlarge figure of red box in (**a**).

In order to verify the validity of the feature subset applied for STLF, there were four weeks distributed among four seasons in 2012 are used to test each feature subset with RF. For comparison, the full set of features with RF is also tested. The experimental results is shown in Figure 11. By examining the results in Figure 11a–d, generalized minimum redundancy and maximum relevance-random forest (G-mRMR-RF) (α = 0.4), mutual information-random forest (MI-RF), sequential forward selection-random forest (SFS-RF), and RF (full features) can fit with true load value accurately, whereas the accuracy of pearson correlation coefficient-random forest (PCC-RF) is low. The results of fifth day prediction in Figure 11a and the seventh day in Figure 11c show G-mRMR-RF has a better fit than MI-RF, indicating the necessity of considering the redundancy among features. The results of fifth day prediction in Figure 11a show that SFS-RF has better prediction performance than G-mRMR, while the seventh day prediction results in Figure 11c indicates G-mRMR-RF predicts better.

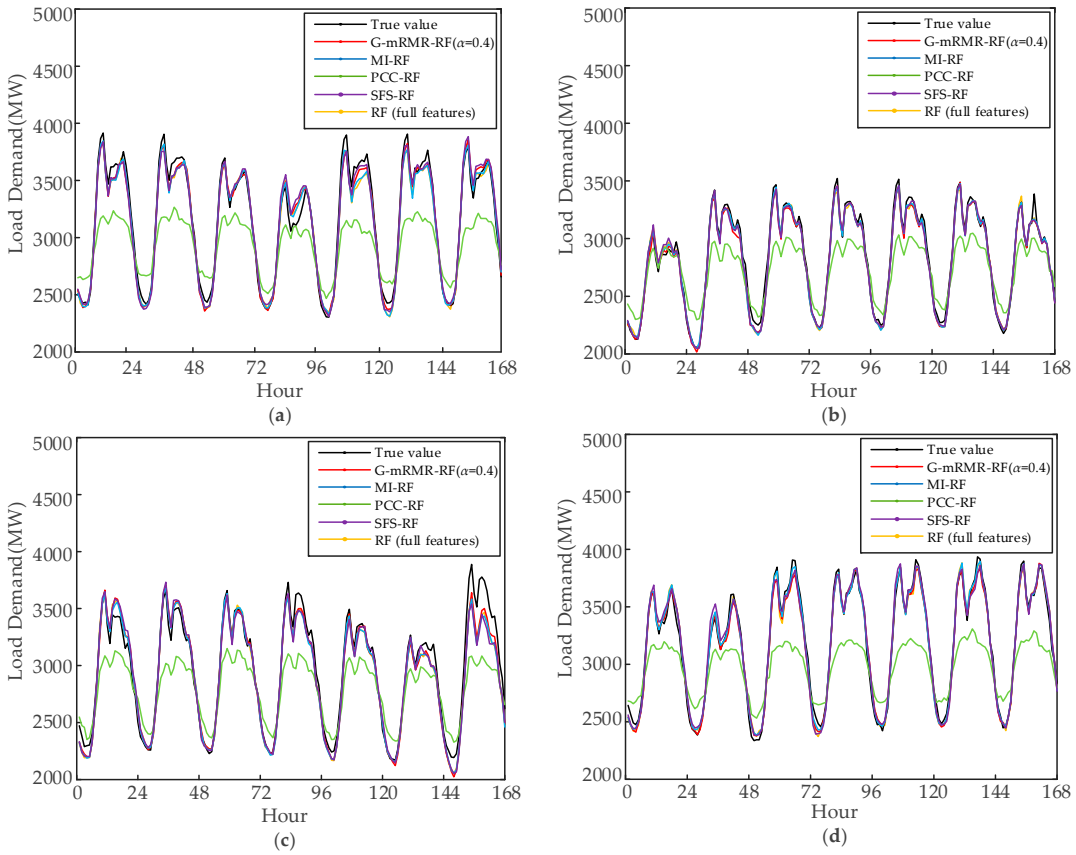

Figure 11. Load curves of forecasting results of four weeks in four seasons and the true values: (**a**) Forecasting from 23 to 29 February 2012; (**b**) Forecasting from 13 to 19 May 2012; (**c**) Forecasting from 21 to 27 August 2012; (**d**) Forecasting from 24 to 30 November 2012.

By analyzing Figures 10 and 11 and Tables 4–7 comprehensively, although SFS achieved the best forecasting results in the feature selection process, the proposed method achieved the better result in the testing schedule. When predicting the 28 days in the test set, the proposed method yields the best forecasting in 20 days and the MAPE in the remaining eight days is higher than other methods, ranging from 0.04% to 0.37%. The average MAPE and the average RMSE indicate G-mRMR-RF performs the best among the methods which demonstrates the validity and advancement of G-mRMR.

The new method also has the minimum value of the maximum error of STLF in the testing set. As shown in Table 6, the maximum MAPE and maximum RMSE of the proposed method are 6.12% and 208.00 MW. Although the maximum error of the new method is high, but compared with other

methods, the proposed method still performed better. The high prediction error can be caused by two factors. On the one hand, the load of forecasting day is much larger than the historical load data in the training set. In this paper, most features in the original feature set are extracted from the historical load data. Without the consideration of other features, the prediction results cannot advance just by improving the feature selection and forecasting method. On the other hand, with the significant economic rise of China from 2005 to 2012, the growth rate of gross domestic product of the city is more than 10%. Under this premise, the electric load of the city increases rapidly which makes STLF a challenging work.

Table 4. Comparison of prediction error (MAPE (%) and RMSE (MW)) from 23 to 29 February 2012.

Day	G-mRMR-RF (α = 0.4)		MI-RF		PCC-RF		SFS-RF		RF with Full Features	
	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
Day 1	1.93	75.24	1.79	69.42	10.28	401.01	2.07	79.58	1.91	70.74
Day 2	1.77	66.63	1.78	67.90	9.78	388.26	2.22	79.46	1.80	69.13
Day 3	1.58	53.24	1.63	51.63	7.59	285.51	1.47	49.33	1.50	50.49
Day 4	1.69	79.28	1.59	70.02	5.35	189.65	2.52	105.32	1.98	76.33
Day 5	2.26	90.72	2.66	104.16	11.14	440.91	2.04	83.68	2.91	113.32
Day 6	1.58	57.73	2.37	83.87	9.78	396.44	1.61	57.41	2.54	87.59
Day 7	1.28	51.92	0.97	36.35	9.26	362.46	1.87	73.03	1.29	44.60
Average	1.72	67.82	1.82	69.05	9.02	352.03	1.97	75.40	1.99	73.17

Table 5. Comparison of prediction error (MAPE (%) and RMSE (MW)) from 13 to 19 May 2012.

Day	G-mRMR-RF (α = 0.4)		MI-RF		PCC-RF		SFS-RF		RF with Full Features	
	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
Day 1	1.20	42.36	1.22	39.15	3.76	110.39	1.43	50.90	1.57	47.92
Day 2	1.64	60.32	1.33	50.26	8.98	273.78	1.28	46.10	1.37	53.34
Day 3	2.04	66.88	2.04	67.09	6.56	246.64	2.03	69.43	2.00	66.78
Day 4	0.94	34.38	0.96	34.48	7.04	263.29	0.89	34.98	1.11	41.54
Day 5	1.55	53.26	1.40	46.62	7.17	261.54	1.40	50.04	1.50	52.38
Day 6	1.28	41.45	1.34	44.68	6.66	237.55	1.28	40.22	1.45	40.03
Day 7	0.84	26.82	0.99	36.97	5.51	178.83	0.92	50.61	1.01	49.05
Average	1.35	46.49	1.33	48.03	6.53	224.57	1.32	48.90	1.40	50.15

Table 6. Comparison of prediction error (MAPE (%) and RMSE (MW)) from 21 to 27 August 2012.

Day	G-mRMR-RF (α = 0.4)		MI-RF		PCC-RF		SFS-RF		RF with Full Features	
	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
Day 1	2.88	92.71	2.61	83.05	6.69	258.31	3.32	104.41	2.89	90.68
Day 2	1.48	55.30	1.55	56.81	8.22	319.74	1.77	62.53	1.59	57.62
Day 3	0.91	31.93	0.82	29.02	7.04	263.33	1.00	38.68	1.07	36.28
Day 4	1.88	76.95	2.27	90.86	8.97	344.82	1.99	84.88	2.17	87.44
Day 5	1.77	54.77	1.87	56.56	6.42	227.25	2.16	70.95	1.91	58.15
Day 6	2.08	73.60	1.78	71.44	5.91	181.33	1.71	65.13	1.86	74.78
Day 7	6.12	208.00	6.77	237.00	11.26	458.19	6.98	247.66	6.57	227.17
Average	2.45	72.83	2.52	89.25	7.79	293.28	2.70	96.32	2.58	90.30

Table 7. Comparison of prediction error (MAPE (%) and RMSE (MW)) from 24 to 30 November 2012.

Day	G-mRMR-RF (α = 0.4)		MI-RF		PCC-RF		SFS-RF		RF with Full Features	
	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
Day 1	1.61	58.60	1.64	58.26	6.80	263.40	1.96	68.90	1.78	63.62
Day 2	1.05	48.24	1.12	43.26	6.97	242.74	2.11	78.43	1.09	40.30
Day 3	1.98	74.62	2.06	74.67	10.78	427.39	1.98	73.90	2.12	76.64
Day 4	1.47	57.14	1.33	48.09	9.21	387.30	1.80	67.13	1.50	57.63
Day 5	1.12	42.84	0.90	33.83	10.29	413.17	1.15	46.87	1.26	45.01
Day 6	1.31	53.79	1.33	52.33	9.03	389.52	1.32	54.74	1.23	47.40
Day 7	1.10	42.86	1.22	45.31	9.53	387.69	1.06	44.21	1.18	43.42
Average	1.38	54.01	1.37	50.82	8.93	358.74	1.63	62.02	1.45	53.43

5.2.2. Comparison of Different Intelligent Methods

For comparing the influence of different predictors to STLF, support vector regression (SVR) and back propagation neural network (BPNN) are examined with G-mRMR for feature selecting in this subsection. The parameters of SVR are set as follows: the penalty factor is C = 100, the insensitive loss function is ε = 0.1, and the kernel width is δ^2 = 2 [41].

The parameters of BPNN are set as follows: the number of neurons in hidden layer is $N_{neu} = 2p+1$ [42], and the iteration is T = 2000 [43].

Data consist of training set, validation set, and test set are similar with Section 4.2. The SVR and BPNN are used to generate the optimal feature subsets.

Table 8 presents feature subsets that different intelligent STLF methods had selected. With different predictors, the weighting factors are diverse, thus features are varying. Although the final number of feature selected by SVR and BPNN are less than RF, the RF-based predictor has higher precision of prediction which is the main target of STLF.

Table 8. The optimal subset selected by using different intelligent STLF methods.

Predictor	Min MAPE (%)	Number of Features	Feature Subset
G-mRMR-RF (α = 0.4)	2.5389%	15	$F_{L(t-168)}$, $F_{L(t-25)}$, $F_{L(t-48)}$, F_{WW}, F_S, $F_{L(t-127)}$, $F_{L(t-85)}$, $F_{L(t-139)}$, F_{DW}, $F_{L(t-34)}$, $F_{L(t-160)}$, $F_{L(t-70)}$, $F_{L(t-28)}$, $F_{L(t-120)}$, $F_{L(t-141)}$
G-mRMR-SVR (α = 0.3)	3.3293%	5	$F_{L(t-168)}$, $F_{L(t-25)}$, $F_{L(t-48)}$, F_{Hour}, F_{WW}
G-mRMR-BPNN (α = 0.1)	2.7186%	11	$F_{L(t-168)}$, $F_{L(t-25)}$, $F_{L(t-48)}$, $F_{L(t-144)}$, $F_{L(t-72)}$, F_{Hour}, $F_{L(t-47)}$, $F_{L(t-26)}$, $F_{L(t-120)}$, $F_{L(t-167)}$, F_{WW}

The test sets, with four weeks being distributed over the four seasons, are used for estimating each predictor with the features chosen above. Figure 12 shows the MAPE for comparison and Table 9 gives the predictive accuracy of each model through maximum, minimum, and average MAPE. In addition, a direct comparison between G-mRMR-RF, generalized minimum redundancy and maximum relevance-back propagation neural network (G-mRMR-BPNN), and generalized minimum redundancy and maximum relevance-support vector regression (G-mRMR-SVR), in terms of MAPE, are also presented in this figure. Except for the MAPE prediction in the seventh day, as shown in Figure 12c, the accuracy of G-mRMR-RF is between 1% and 2%; one point is above 2%. In the whole experiment, only four days show that G-mRMR-RF forecasted worse than other models. Clearly, the G-mRMR-RF is the best prediction model for its low MAPE and small fluctuation of error. The G-mRMR-BPNN shows a little better performance than G-mRMR-SVR. We can observe the maximum MAPE of these four weeks of G-mRMR-RF is 2.26%, 2.04%, 6.12%, 1.98%, respectively, which is smaller than other models. Same conclusion can be drawn by analyzing the minimum and average MAPE.

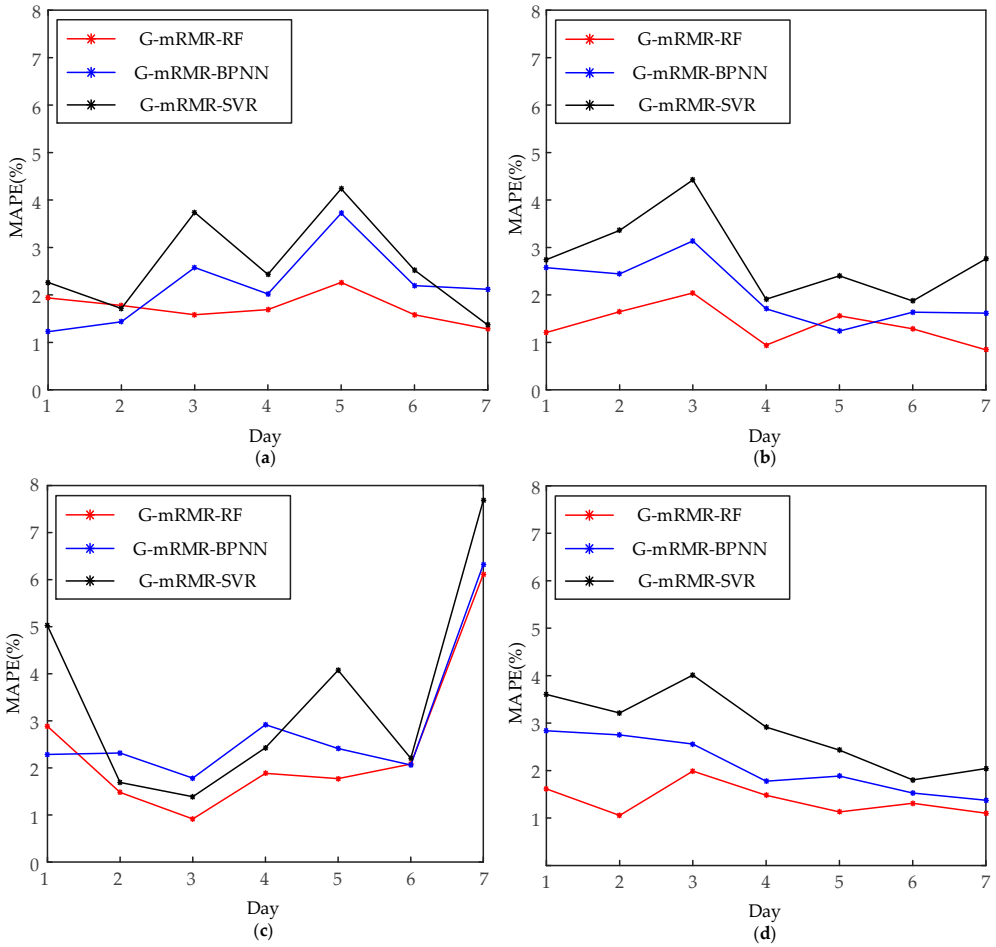

Figure 12. Forecasting error profiles of different predictors: (**a**) Forecasting from 23 to 29 February 2012; (**b**) Forecasting from 13 to 19 May 2012; (**c**) Forecasting from 21 to 27 August 2012; (**d**) Forecasting from 24 to 30 November 2012.

Table 9. Max, Min and Average daily MAPEs of test set corresponding to different predictors.

Test Set	MAPE (%)	G-mRMR-RF	G-mRMR-ANN	G-mRMR-SVR
23–29 February 2012 (Winter)	Max	2.26	3.72	4.24
	Min	1.28	1.23	1.37
	Ave	1.72	2.18	2.61
13–19 May 2012 (Spring)	Max	2.04	3.14	4.42
	Min	0.84	1.24	1.87
	Ave	1.35	2.05	2.78
21–27 August 2012 (Summer)	Max	6.12	6.32	7.68
	Min	0.91	1.78	1.69
	Ave	2.45	2.87	3.50
24–30 November 2012 (Fall)	Max	1.98	2.84	4.01
	Min	1.05	1.38	1.80
	Ave	1.38	2.10	2.86

Based on the comprehensive analysis above, as compared to BPNN and SVR, the RF combines with G-mRMR is more suitable for STLF.

6. Conclusions

For the issues regarding the selection of reasonable features for STLF, a feature selection method based on G-mRMR and RF is proposed in this paper. The experimental results show that the proposed feature selection approach can select fewer features than other feature selection methods, and the features identified by the proposed approach are useful for STLF. In addition, the experimental results show that the forecasting consequences by RF are better than other predictors.

The advantages of the proposed method are as follows:

(1) MI is adopted as the criterion to measure the relevance between features and time series of load and the dependency among features, which is the basis of quantitative analysis of feature selection by mRMR.

(2) The correlation between features and load as well as the redundancy of these features are considered. As compared to the maximum relevance method, the G-mRMR method for feature selection reduces the number of optimal feature subset and avoids the association of STLF accuracy with the redundancy of features. For the time being, the relevance and redundancy are balanced by using a variable weighting factor. The features selected by G-mRMR make the accuracy of RF more precise than mRMR.

(3) The optimal structure of RF is designed for reducing the complexity of the model and for improving the accuracy of STLF.

Acknowledgments: This work is supported by the National Nature Science Foundation of China (No. 51307020), the Science and Technology Development Project of Jilin Province (No. 20160411003XH) and the Science and Technology Foundation of Department of Education of Jilin Province (2016, No. 90).

Author Contributions: Nantian Huang put forward to the main idea and design the whole venation of this paper. Zhiqiang Hu did the experiments and prepared the manuscript. Guowei Cai guided the experiments and paper writing. Dongfeng Yang provided materials. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Moslehi, K.; Kumar, R. A reliability perspective of the smart grid. *IEEE Trans. Smart Grid* **2010**, *1*, 57–64. [CrossRef]
2. Ren, Y.; Suganthan, P.N.; Srikanth, N.; Amaratunga, G. Random vector functional link network for short-term electricity load demand forecasting. *Inf. Sci.* **2016**, 367–368, 1078–1093. [CrossRef]
3. Lee, C.-M.; Ko, C.-N. Short-term load forecasting using lifting scheme and ARIMA models. *Expert Syst. Appl.* **2011**, *38*, 5902–5911. [CrossRef]
4. Goia, A.; May, C.; Fusai, G. Functional clustering and linear regression for peak load forecasting. *Int. J. Fofrecast.* **2010**, *26*, 700–711. [CrossRef]
5. Al-Hamadi, H.M.; Soliman, S.A. Fuzzy short-term electric load forecasting using Kalman filter. *IEE Proc. Gener. Transm. Distrib.* **2006**, *153*, 217–227. [CrossRef]
6. Ramos, S.; Soares, J.; Vale, Z. Short-term load forecasting based on load profiling. In Proceedings of the 2013 IEEE Power and Energy Society General Meeting, Vancouver, BC, Canada, 21–25 July 2013; pp. 1–5.
7. Li, W.; Zhang, Z.G. Based on Time Sequence of ARIMA Model in the Application of Short-Term Electricity Load Forecasting. In Proceedings of the 2009 International Conference on Research Challenges in Computer Science, Shanghai, China, 28–29 December 2009; pp. 11–14.
8. Deshmukh, M.R.; Mahor, A. Comparisons of Short Term Load Forecasting using Artificial Neural Network and Regression Method. *Int. J. Adv. Comput. Res.* **2011**, *1*, 96–100.
9. Taylor, J.W. Short-Term Load Forecasting With Exponentially Weighted Methods. *IEEE Trans. Power Syst.* **2012**, *27*, 458–464. [CrossRef]
10. Kouhi, S.; Keynia, F.; Ravadanegh, S.N. A new short-term load forecast method based on neuro-evolutionary algorithm and chaotic feature selection. *Int. J. Electr. Power Energy Syst.* **2014**, *62*, 862–867. [CrossRef]
11. Raza, M.Q.; Khosravi, A. A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. *Renew. Sustain. Energy Rev.* **2015**, *50*, 1352–1372. [CrossRef]

12. Lin, C.T.; Chou, L.D.; Chen, Y.M.; Tseng, L.M. A hybrid economic indices based short-term load forecasting system. *Int. J. Electr. Power Energy Syst.* **2014**, *54*, 293–305. [CrossRef]
13. Yu, F.; Xu, X. A short-term load forecasting model of natural gas based on optimized genetic algorithm and improved BP neural network. *Appl. Energy* **2014**, *134*, 102–113. [CrossRef]
14. Çevik, H.H.; Çunkaş, M. Short-term load forecasting using fuzzy logic and ANFIS. *Neural Comput. Appl.* **2015**, *26*, 1355–1367. [CrossRef]
15. Lahouar, A.; Slama, J.B.H. Day-ahead load forecast using random forest and expert input selection. *Energy Convers. Manag.* **2015**, *103*, 1040–1051. [CrossRef]
16. Ho, K.L.; Hsu, Y.Y.; Chen, C.F.; Lee, T.E. Short term load forecasting of Taiwan power system using a knowledge-based expert system. *IEEE Trans. Power Syst.* **1990**, *5*, 1214–1221.
17. Srinivasan, D.; Tan, S.S.; Cheng, C.S.; Chan, E.K. Parallel neural network-fuzzy expert system strategy for short-term load forecasting: System implementation and performance evaluation. *IEEE Trans. Power Syst.* **1999**, *14*, 1100–1106. [CrossRef]
18. Quan, H.; Srinivasan, D.; Khosravi, A. Uncertainty handling using neural network-based prediction intervals for electrical load forecasting. *Energy* **2014**, *73*, 916–925. [CrossRef]
19. Hernández, L.; Baladrón, C.; Aguiar, J.M.; Carro, B.; Sánchez-Esguevillas, A.; Lloret, J. Artificial neural networks for short-term load forecasting in microgrids environment. *Energy* **2014**, *75*, 252–264. [CrossRef]
20. Ko, C.N.; Lee, C.M. Short-term load forecasting using SVR (support vector regression)-based radial basis function neural network with dual extended Kalman filter. *Energy* **2013**, *49*, 413–422. [CrossRef]
21. Che, J.X.; Wang, J.Z. Short-term load forecasting using a kernel-based support vector regression combination model. *Appl. Energy* **2014**, *132*, 602–609. [CrossRef]
22. Pandian, S.C.; Duraiswamy, K.; Rajan, C.C.A.; Kanagaraj, N. Fuzzy approach for short term load forecasting. *Electr. Power Syst. Res.* **2006**, *76*, 541–548. [CrossRef]
23. Božić, M.; Stojanović, M.; Stajić, Z.; Stajić, N. Mutual Information-Based Inputs Selection for Electric Load Time Series Forecasting. *Entropy* **2013**, *15*, 926–942. [CrossRef]
24. Ma, L.; Zhou, S.; Lin, M. Support Vector Machine Optimized with Genetic Algorithm for Short-Term Load Forecasting. In Proceedings of the 2008 International Symposium on Knowledge Acquisition and Modeling, Wuhan, China, 21–22 December 2008; pp. 654–657.
25. Gao, R.; Liu, X. Support vector machine with PSO algorithm in short-term load forecasting. In Proceedings of the 2008 Chinese Control and Decision Conference, Yantai, China, 2–4 July 2008; pp. 1140–1142.
26. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
27. Jurado, S.; Nebot, À.; Mugica, F.; Avellana, N. Hybrid methodologies for electricity load forecasting: Entropy-based feature selection with machine learning and soft computing techniques. *Energy* **2015**, *86*, 276–291. [CrossRef]
28. Wilamowski, B.M.; Cecati, C.; Kolbusz, J.; Rozycki, P. A Novel RBF Training Algorithm for short-term Electric Load Forecasting and Comparative Studies. *IEEE Trans. Ind. Electron.* **2015**, *62*, 6519–6529.
29. Wi, Y.M.; Joo, S.K.; Song, K.B. Holiday load forecasting using fuzzy polynomial regression with weather feature selection and adjustment. *IEEE Trans. Power Syst.* **2012**, *27*, 596–603. [CrossRef]
30. Viegas, J.L.; Vieira, S.M.; Melício, M.; Mendes, V.M.F.; Sousa, J.M.C. GA-ANN Short-Term Electricity Load Forecasting. In Proceedings of the 7th IFIP WG 5.5/SOCOLNET Advanced Doctoral Conference on Computing, Electrical and Industrial Systems, Costa de Caparica, Portugal, 11–13 April 2016; pp. 485–493.
31. Li, S.; Wang, P.; Goel, L. A novel wavelet-based ensemble method for short-term load forecasting with hybrid neural networks and feature selection. *IEEE Trans. Power Syst.* **2015**, 1788–1798. [CrossRef]
32. Hu, Z.; Bao, Y.; Chiong, R.; Xiong, T. Mid-term interval load forecasting using multi-output support vector regression with a memetic algorithm for feature selection. *Energy* **2015**, *84*, 419–431. [CrossRef]
33. Koprinska, I.; Rana, M.; Agelidis, V.G. Yearly and seasonal models for electricity load forecasting. In Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN), San Jose, CA, USA, 31 July–5 August 2011; pp. 1474–1481.
34. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef] [PubMed]

35. Nguyen, X.V.; Chan, J.; Romano, S.; Bailey, J. Effective global approaches for mutual information based feature selection. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24–27 August 2014; pp. 512–521.
36. Speybroeck, N. Classification and regression trees. *Int. J. Public Health* **2012**, *57*, 243–246. [CrossRef] [PubMed]
37. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
38. Sood, R.; Koprinska, I.; Agelidis, V.G. Electricity load forecasting based on autocorrelation analysis. In Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, Spain, 18–23 July 2010; pp. 1–8.
39. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [CrossRef]
40. Dudek, G. Short-Term Load Forecasting Using Random Forests. In *Intelligent Systems'2014*; Springer: Cham, Switzerland, 2015; pp. 821–828.
41. Che, J.X.; Wang, J.Z.; Tang, Y.J. Optimal training subset in a support vector regression electric load forecasting model. *Appl. Soft Comput.* **2012**, *12*, 1523–1531. [CrossRef]
42. Sheela, K.G.; Deepa, S.N. Review on Methods to Fix Number of Hidden Neurons in Neural Networks. *Math. Probl. Eng.* **2013**, *2013*, 425740. [CrossRef]
43. Rana, M.; Koprinska, I. Forecasting electricity load with advanced wavelet neural networks. *Neurocomputing* **2016**, *182*, 118–132. [CrossRef]