sustainability

MDPI

Article

GMDH-Based Semi-Supervised Feature Selection for Electricity Load Classification Forecasting

Lintao Yang [1], Honggeng Yang [1,*], Hongyan Yang [2] and Haitao Liu [2]

[1] College of Electrical Engineering and Information Technology, Sichuan University, Chengdu 610065, China; yanglt352@163.com

[2] Business School, Sichuan University, Chengdu 610065, China; hongyan_yang1994@163.com (H.Y.); haitaoliuch@gmail.com (H.L.)

* Correspondence: pqlab99@126.com or yanghonggeng@sina.cn; Tel.: +86-28-8541-7867; Fax: +86-28-8541-5628

Received: 14 December 2017; Accepted: 15 January 2018; Published: 16 January 2018

Abstract: With the development of smart power grids, communication network technology and sensor technology, there has been an exponential growth in complex electricity load data. Irregular electricity load fluctuations caused by the weather and holiday factors disrupt the daily operation of the power companies. To deal with these challenges, this paper investigates a day-ahead electricity peak load interval forecasting problem. It transforms the conventional continuous forecasting problem into a novel interval forecasting problem, and then further converts the interval forecasting problem into the classification forecasting problem. In addition, an indicator system influencing the electricity load is established from three dimensions, namely the load series, calendar data, and weather data. A semi-supervised feature selection algorithm is proposed to address an electricity load classification forecasting issue based on the group method of data handling (GMDH) technology. The proposed algorithm consists of three main stages: (1) training the basic classifier; (2) selectively marking the most suitable samples from the unclassified label data, and adding them to an initial training set; and (3) training the classification models on the final training set and classifying the test samples. An empirical analysis of electricity load dataset from four Chinese cities is conducted. Results show that the proposed model can address the electricity load classification forecasting problem more efficiently and effectively than the FW-Semi FS (forward semi-supervised feature selection) and GMDH-U (GMDH-based semi-supervised feature selection for customer classification) models.

Keywords: peak load; classification forecasting; group method of data handling; semi-supervised learning

1. Introduction

Electricity load forecasting is a major issue in the planning and operation of modern electricity networks and electricity markets [1,2]. Electricity load forecasting can be classified into long-term [3], medium-term [4], short-term [5–7] and ultra-short term [8], and the cut-off points for these four categories are three years, two weeks, and one day, respectively [9]. The short-term load forecasting (STLF), which is applied to horizons no more than one day ahead, can result in significant environmental and economic benefits for energy systems. For reliable and efficient operations, STLF is used when decision-making has significant impacts on the operations, such as scheduling generating capacity dispatches, demand side management, security assessments, and generator maintenance scheduling [5,10–16]. Unsatisfactory STLF can cause the increase in the operational cost, equipment failures, or systems blackouts, thus resulting in a waste of resources [17–19]. As the implementation of accurate and timely forecasting methods is important for environmental-friendly, economically sound operations, STLF research is essential to ensure efficient and reliable power system operations.

STLF involves the electricity load forecasting of total demand and peak demand within one day. For example, Bessec and Fouquau [19] developed an one day-ahead forecast for half-hourly electricity

loads using a combination of stationary wavelet transformations that yielded 502 daily observations for each half-hour in France. Based on the corresponding weather forecasts, Feng and Ryan [20] provided accurate day-ahead hourly load forecasting for multiple zones within a region using a temporal and weather conditional epi-splines-based load models. Tong et al. [21] developed a deep learning based model and established a support vector regression model to forecast the total day-ahead electricity load, and then refined the features by stacking the denoising auto-encoders with historical electricity load data and related temperature parameters.

In addition to total electricity load forecasting, the peak load forecasting has also been found to be related to power network dispatch centers. For instance, dispatching center operators require daily peak loads for scheduled maintenance or adequate assessments. Therefore, the forecasting of daily peak loads should be considered in the STLF. However, only a few researchers considered electricity peak load forecasting in the past. Amjady [22] presented a new time series models that could precisely forecast the daily peak loads of a power system, and obtained results from extensive tests to confirm the validity of the developed approach. In reality, because of the hysteresis in generator units, even a large number of spare generator units fail to meet immediate electricity needs when loads reach a peak and cause power restrictions. Therefore, it is essential to accurately forecast peak loads in power grids.

There is a little research focusing on electricity peak load forecasting because numerous studies only seek to predict specific electricity loads [23]. The electricity peak load interval forecasting has not been investigated so far. On the other hand, the peak load interval forecasting has greater practical value than that of specific electricity loads, since the power generation from generator sets has an interval value, which means that operators need to open spare units in advance. When the peak load lies in different intervals, the power dispatcher needs to configure the corresponding generators in advance. Therefore, this paper seeks to convert the peak power load into an interval load and then further translates it into a peak power load so as to forecast peak load classifications.

Previous research has paid close attention to the accurate forecasting of electricity loads, and multiple methods. For instance, the classical statistical methods [24] and machine learning methods [25–27] have been proposed for the electricity load forecasting. The classical statistical methods often assume that the load is a function of several explanatory variables and estimate the specified functional parameters [28,29]. One of the well-known methods is the seasonal autoregressive integrated moving average (SARIMA) proposed by Box and Jenkins [30]. To improve forecasting accuracy, there have been numerous attempts to enhance models. For example, Soares and Medeiros [31] proposed a SARIMA model for hourly electricity loads in southeast Brazil. Although SARIMA models are easy to use and are capable of forecasting accurately, they have some limitations. The machine learning methods such as artificial neural networks (ANNs) and support vector regression (SVRs) are restricted to specified functions [20]. The SVR-based electricity load forecasting methods are proposed and show good performance mainly due to the strong non-linear learning capability of SVR. The comparison between the machine learning methods (ANNs and SVRs) and the discrete-time univariate econometric models can be found in [32]. Both theoretical and empirical findings have indicated that a combination of different models could overcome the limitations of single models and improve forecasting accuracy by harnessing each mode's merits. Consequently, there have been several hybrid models developed that incorporate different energy field models for electricity load forecasting. Some researchers proposed the hybrid method which consists of a neural network and the evolutionary algorithms [33]. For instance, Mori and Takahashi [34] proposed a hybrid intelligent method for probabilistic STLF, and Xiong et al. [35] converted hourly load series into a 24 monthly interval time series and proposed a hybrid approach for forecasting the electricity demand intervals. Fan et al. [2] proposed a SVR model combining the auto regression with the differential empirical mode decomposition method for a kind of electricity load forecasting. Although the above methods may be used to resolve the continuous the electricity load forecasting problem, they are not suitable to tackle the electricity peak loads classification forecasting issue. Hence, it is necessary for academics to propose novel methods to solve the classification forecasting problem of electricity peak loads.

The group method of data handling, which is a family of inductive algorithms for the computer-based mathematical modeling of multi-parametric datasets, has been found to be an effective tool for solving the classification problem in machine learning field. It can also be used for short-term load forecasting [36,37] and traffic flow prediction [38]. The GMDH-type neural network that is the combination of GMDH and neural networks, can improve forecasting accuracy [39], and solve the classification problems efficiently [40]. Nevertheless, to the best of our knowledge, there is no research that has utilized GMDH and neural network for electricity peak load classification forecasting.

This paper investigates the one day-ahead electricity peak load classification forecasting problem. One major contribution is that it transforms the conventional continuous forecasting into a novel interval forecasting, and then further converts the interval forecasting into the classification forecasting. In addition, an indicator system of influencing the electricity load is established from three dimensions, namely the load series, calendar data and weather data. Another contribution is that a novel semi-supervised feature selection algorithm is proposed to address the electricity load classification forecasting problem based on the group method of data handling technology.

The rest of the paper is organized as follows. The related theory and the GMDH-based semi-supervised feature selection for an electricity load classification model are introduced in Section 2. Section 3 presents the experimental design and analyzing results in detail. Section 4 draws conclusions and provides suggestions for the future research.

2. GMDH-Based Semi-Supervised Feature Selection for an Electricity Load Classification Model

2.1. GMDH Network

The group method of data handling (GMDH) is a basic technique for self-organized learning. It enables the researchers to control the process of the complex model from the input set to the output data and to determine the model parameters [41–43].

The GMDH network establishes a relationship between input and output, which is referred to as the Volterra function series or the Kolmogorov–Gabor polynomial function:

$$y = a_0 + \sum_{i=1}^{m} a_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} a_{ijk} x_i x_j x_k + \ldots \quad (1)$$

Suppose that the linear function is set. All items are then taken as the $m+1$ initial input variables. The specific modeling process is as follows. From the transfer function, a new neuron is obtained to construct the first layer (see Figure 1). The specific expression is as follows:

$$y_1^k = a_1^k + a_2^k v_i + a_3^k v_j . i, j = 1, 2, \ldots, m_0, j \neq i, k = 1, 2, \ldots, t_1 \quad (2)$$

First, the parameters are calculated by the using least squares estimation and the external criterion value of every intermediate candidate model according to the model selection set. In general, the accuracy of the intermediate candidate model increases when the external criterion value decreases. When the confidence level is selected, the external criteria values are measured using the threshold value measurement. Finally, every two models are paired, which then becomes the input for the second layer:

$$y_2^k = a_1^k + a_2^k v_i + a_3^k v_j . i, j = 1, 2, \ldots, F_1, j \neq i, k = 1, 2, \ldots, t_2 \quad (3)$$

Similarly, the intermediate candidate model $t_2 = C_{F_1}^2$ is obtained in the second layer. Repeating the above steps, the model continues working until an optimal complexity model is determined. Therefore, the termination principle obeys the optimal complexity principle [44]. To identify the initial model contained in the optimal complexity model y^*, the GMDH network structure can be examined from the last layer to the initial input layer. As shown in Figure 1, v_1, v_2, v_3, v_4, v_5 are chosen as the initial input model. Then, each variable is paired with another in a group to compete with each other.

Nonetheless, y_1, y_2, y_3, y_4 are preserved by the algorithm. Note that v_1, v_3, v_4, v_5 remain in the model to participate in the subsequent competition, however, v_2 is eliminated. In other words, x_2, x_3, x_4 are selected and x_1 is deleted.

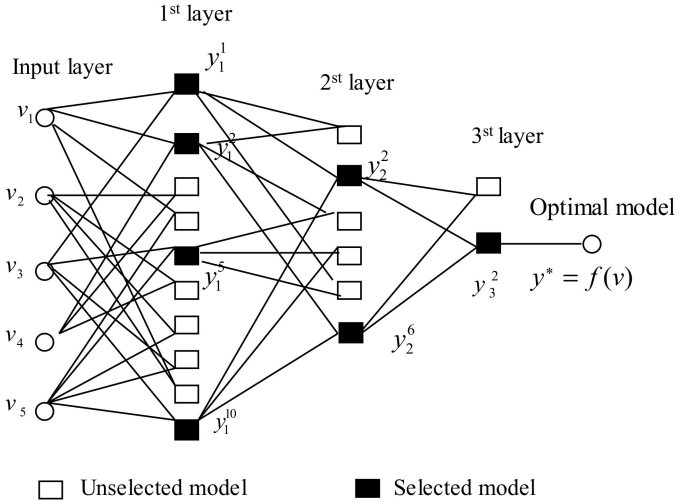

Figure 1. An illustration of modeling process for the GMDH.

2.2. Basic Modeling Idea

This paper proposes a GMDH-based semi-supervised feature selection (SSFS-GMDH) model to deal with the electricity load classification forecasting problem. In this model, the labeled and unlabeled samples are used for the feature selection. Suppose L is the original labeled training set for the electricity peak load forecasting problem, T is the labeled testing set, and U is the number of an unlabeled dataset. L is firstly divided into a simulated training set L_{train} and a simulated validation set L_{verify}. The flowchart of the proposed method is shown in Figure 2. The proposed model contains three major stages. (1) The classified dataset L is used to train N basic classification models. (2) Label the labeled samples in the dataset U by using basic classification models. A certain proportion of marked samples U_α are chosen from U^l and the samples merged in L. These two stages are repeated until the proportion of selected samples exceeds θ. (3) Train the basic classification model until the final training set L and the feature set F_s are obtained.

During the modeling process, the building of the external criteria is also crucial. A detailed description of the SSFS-GMDH model and external criteria are shown in Sections 2.2 and 2.3. The interpretation for the symbols can be found in Table 1.

Table 1. Interpretation for the symbols.

Symbols	Interpretation
L	original labeled training set
T	labeled testing set
U	unlabeled dataset
U_α	chosen unlabeled sample
U^l	marked unlabeled sample
L_{train}	training set
L_{verify}	validation set
N	the number of basic classification models
F_s	feature set
K	the number of neighboring samples
p, θ	the proportion of samples chosen to be added into training set
δ	the confidence level

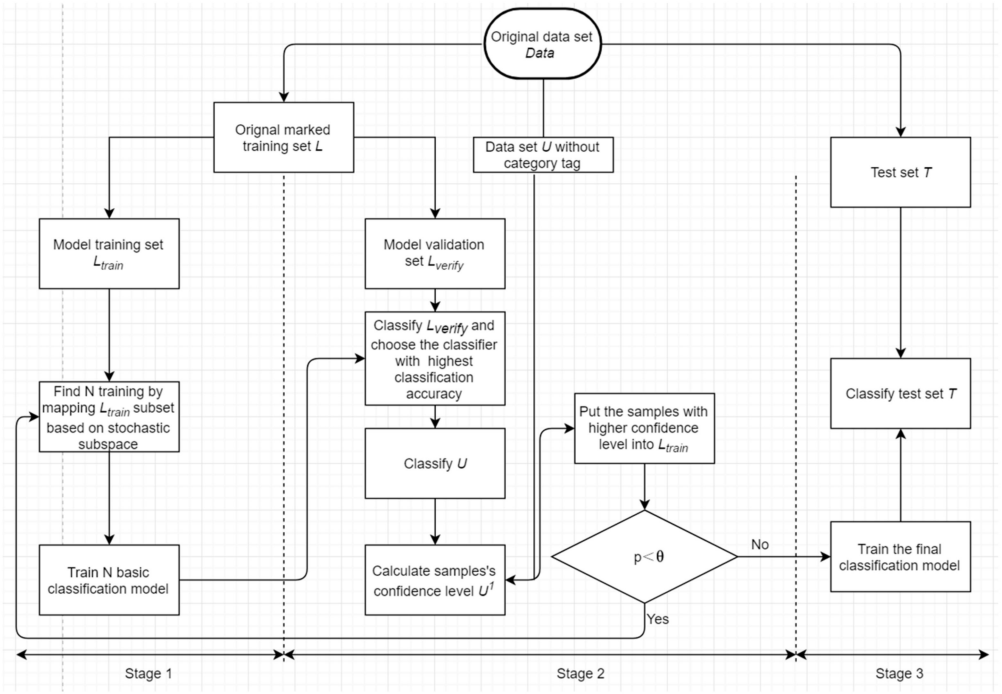

Figure 2. Flowchart of the SSFS-GMDH model.

2.3. Detailed Modeling Steps

The basic flowchart of the SSFS-GMDH model is illustrated in Figure 2, and the detailed modeling steps are as follows:

Input: L, U, T, K, θ, p.

Output: Classification results from the final training of the test set.

Step 1: Divide the original dataset into training set L with a category label and dataset U without a category label, and test dataset T with a category label. Further divide L into the simulated training set L_{train} and the simulated validation set .

Step 2: Find N training by mapping the L_{train}subsets based on the stochastic subspace, and then train the N basic classification models.

Step 3: Use the training classification model to classify L_{verify}, and then choose the classifier with highest classification accuracy.

Step 4: Use the selected classification model to mark the catalog tag on the unclassified dataset U, and find sample U^l with a catalog tag.

Step 5: Calculate and sort the confidence level of each sample; δ is defined as the confidence level of each marked sample U_i^l in set U^l, the calculation formula is:

$$\delta = \frac{k}{K} \tag{4}$$

where K is the number of neighboring samples chosen from the initial labeled training set L. k reflects the number of neighboring samples that have the same class labels as samples among K neighbors. In this paper, the Euclidean distance is used to calculate the distance between samples. It is obvious that the higher the value of $\delta \in [0, 1]$ is the higher the confidence level will be. Then, sort the marked samples based on the confidence level of each sample.

Step 6: Choose a certain proportion of the marked samples with a higher confidence level from U_i^l and put them into L_{train}.

Step 7: Repeat Steps 2 to 6. The iteration stops when the proportion p of the sample added to L_{train} in U exceeds θ.

Step 8: Train the final classification model, select the final character subset F_s and classify the samples in the testing set T.

2.4. Establishing the GMDH External Criteria

There are two fundamental types of GMDH (group method of data handling) external criteria: the accuracy criteria and the compatibility criteria. Accuracy criteria focus on the random errors in different established model parts, and are also referred to as fitting precision, while compatibility criteria highlight the consistency of the models built for the same system in datasets from different samples [45]. Ivakhnenko et al. [40] established regularization criteria and a theoretical basis for symmetric regularization criteria, and proved that the regularization criteria and symmetric regularization criteria could be used as the external criteria in GMDH theory. Because of the different application scopes, different external criteria for different GMDH have significant impacts on the model classification performance [46]. Details of the 13 types of external criteria are as follows:

SSFS-GMDH1: Symmetric mean square error.

$$d(W) = \Delta(A) + \Delta(B) \tag{5}$$

where $\triangle(A) = \sqrt{(\sum_{t\in A}(y_t - y_t^m(B))^2)/N_A}$, $\triangle(B) = \sqrt{(\sum_{t\in B}(y_t - y_t^m(A)^2)/N_B}$.

SSFS-GMDH2: Symmetric regularization criteria.

$$d^2(W) = \Delta^2(A) + \Delta^2(B) \tag{6}$$

where $\Delta^2 A = \sum_{t\in w}(y_t - y_t^m(A))^2$, $\Delta^2 B = \sum_{t\in w}(y_t - y_t^m(A))^2$.

SSFS-GMDH3: Average regularization criteria.

$$d^2(W) = \Delta^2(W) = \left(\sum_{t\in w}(y_t - y_t^m(W))^2\right)/N_w \tag{7}$$

SSFS-GMDH4: Symmetric stability criteria.

$$d^2(W) = \Delta^2(A) + \Delta^2(B) \tag{8}$$

where $\Delta^2 A = \sum_{t\in w}(y_t - y_t^m(B))^2$, $\Delta^2 B = \sum_{t\in w}(y_t - y_t^m(A))^2$.

SSFS-GMDH5: Forecasting criteria.

$$i^2(W) = i^2(A) + i^2(B) \tag{9}$$

where $i^2(A) = \sum_{t\in C}(y_t - y_t^m(A))^2$, $i^2(B) = \sum_{t\in C}(y_t - y_t^m(B))^2$.

SSFS-GMDH6: Symmetric minimum deviation criteria.

$$\eta_{bs}^2(W) = \|y_t^m(A) - y_t^m(B)\|_{t\in W}^2 \tag{10}$$

SSFS-GMDH7: Symmetric absolute interference criteria.

$$v^2(W) = v^2(A) + v^2(B) \tag{11}$$

where $v^2(A) = \sum_{t\in A}(y_t^m(A) - y_t^m(W))^2$, $v^2(B) = \sum_{t\in B}(y_t^m(B) - y_t^m(W))^2$.

SSFS-GMDH8: Combination criteria minimum deviation criteria + symmetric regularization criteria.

$$\eta_{bs}^2(A) + \eta_{bs}^2(B) + d^2(W) \tag{12}$$

where $\eta_{bs}^2(A) = \|y_t^m(A) - y_t^m(B)\|_{t\in A}^2$, $\eta_{bs}^2(B) = \|y_t^m(A) - y_t^m(B)\|_{t\in B}^2$, $d^2(W) = \Delta^2(A) + \Delta^2(B)$.

SSFS-GMDH9: Combination criteria (symmetric minimum deviation criteria + average regularization criteria).

$$\eta_{bs}^2(W) + d^2(W) \tag{13}$$

SSFS-GMDH10: Combination criteria (symmetric minimum deviation criteria + minimum square error criteria).

SSFS-GMDH11: Asymmetric regularization criteria training model on A and calculating the external criteria on B.

SSFS-GMDH12: Asymmetric stability criteria training on A and calculating the external criteria on W.

SSFS-GMDH13: Asymmetric minimum error criteria.

$$\eta_{bs}^2(A) = \|y_t^m(A) - y_t^m(B)\|_{t\in A}^2 \tag{14}$$

3. Data Description

The electricity load series were provided by the Electric Power Company in the Sichuan Province, China and the sample spanned from January 2013 to June 2017, yielding 1270 daily data. Four representative cities, namely Mianyang, Nanchong, Yibin and Panzhihua, were selected from this province. The indicator system, consisting of the weather variables, calendar variables, and load series, is used to forecast the day-ahead electricity load. There are 18 related variables—one calendar variable, six weather variables, and eleven kinds of load series (Table 2).

- *Calendar variables*: There is one calendar variable that varies across weekdays, weekends, and holidays. Calendar variables are crucial, as electricity loads show daily and weekly periodic variations [47] as well as weekday, weekend, and holiday variations [48].
- *Weather variables*: There are six weather variables: the maximum temperature, minimum temperature, maximum temperature variable rate, minimum temperature variable rate, wind speed, and weather type. As the electricity load is susceptible to changes in weather variables, it is necessary to understand electricity load volatility under various weather conditions within different timescales [49]. Weather variables have been seen as the main parameters controlling energy demand [50,51].
- *Load series:* There are eleven kinds of load series, namely the peak load, off-peak load, daily consumption, cumulative consumption, off-peak consumption, load rate, actual peak load, previous day's electricity consumption, daily consumption in the same period of the previous week, daily consumption in the same period in the previous month, and daily consumption in the same period of the previous year.
- Y: y is defined as the peak load and n as the number of categories, such that $Y \in [1, n] \wedge Y \in Z$. The specification for n is as follows:

$$n = \left\lfloor \frac{y_{\max} - y_{\min}}{S} \right\rfloor \tag{15}$$

where $y_{\max}$ and $y_{\min}$ denote the maximal peak load and the minimal peak load, respectively, and S is defined as the step length that is set based on the power of generators.

Table 2. Evaluation indicator system.

Categories of Indicators	Sub-Indicators
Weather variables	W1: Maximum temperature (°C) W2: Minimum temperature (°C) W3: Maximum temperature variable rate W4: Minimum temperature variable rate W5: Wind speed (m/s) W6: Weather type
Calendar variables	C1: Calendar, such as holidays, weekdays and weekends.
Load series	L1: Peak load L2: Off-peak load L3: Daily consumption L4: Cumulative consumption L5: Off-peak consumption L6: Load rate L7: Actual peak load L8: Previous day's electricity consumption L9: Daily consumption in the same period of the previous week L10: Daily consumption in the same period in the previous month L11: Daily consumption in the same period of the previous year

4. Empirical Analysis

To analyze the performance of the proposed SSFS-GMDH model and different criteria, the practical datasets from Mianyang, Nanchong, Yibin and Panzhihua in China are empirically analyzed. The SSFS-GMDH model is compared with the FW-SemiFS (Forward semi-supervised feature selection) [52] and GMDH-U (GMDH-based semi-supervised feature selection for customer classification) [46] models.

4.1. Experimental Setting

Table 3 shows the parameters used in the experiment. Each particular dataset is divided into three subsets: 30% of samples in the dataset is used as a training set L with a class label, and the another 30% of samples in the dataset is used as a dataset U without a class label, and the remaining 40% of samples as a testing set T with a class label. The range of K and θ is $K \in [2, 15]$ and $\theta \in [0.1, 1]$, respectively. In the SSFS-GMDH model, L is utilized to mark the labels of the samples in U. As this procedure has a significant impact on the performance of the SSFS-GMDH model, it is crucial to choose a proper basic classification model. Therefor three basic and effective classification models include the Support Vector Machines [7], Bayesian Networks [53], and Decision Trees [54,55] are employed in this paper. Each experiment is conducted 30 times via MATLAB2016b.

Table 3. Parameters setting.

Symbols	Parameters Setting
L	30%
T	40%
U	30%
N	3
K	$K \in [2, 15]$
p, θ	$\theta \in [0.1, 1]$

4.2. Model Evaluation Criteria

The most common evaluation criterion for evaluating classification forecasting models is the accuracy on the testing set. Since this is the appropriate way to evaluate the performance of models

dealing with unbalanced class distributions, the ROC (Receiver Operating Characteristic) curve can be used to evaluate the model' classification performance. However, since it is inconvenient to directly compare each model's ROC curve, the AUC (the area under the ROC curve) is usually taken as the model evaluation criterion. The ROC curve and AUC value are both capable of evenly handling the minorities and the majorities. Nevertheless, the AUC value can better weigh the minority recognition rate against the majority recognition rate, and the larger the AUC value, the better the model performance [56].

The classification evaluation matrix is then introduced. As shown in Table 4, *TP* denotes the number of correctly predicted positive classes, *FN* denotes the number of wrongly predicted negative classes, *FP* denotes the number of wrongly predicted positive classes, and FP denotes the number of correctly predicted negative classes. To deal with the dichotomy, the ROC curve is a true positive rate–false positive rate figure, in which the horizontal axis of the figure shows the fake positive rate ($=FP/(FP + TN) \times 100\%$) and the vertical axis shows the true positive rate ($=TP/(TP + FN) \times 100\%$).

Table 4. Classification evaluation matrix.

	Class Predicted to Be Positive	Class Predicted to Be Negative
Positive Class	*TP*	*FN*
Negative Class	*FP*	*TN*

4.3. Analysis of the Impacts of the GMDH External Criteria on Classification Performance of the SSFS-GMDH Model

This paper constructs 13 external criteria and then conducts tests on four datasets to determine the best external criteria by exploring the relationships between the external criteria classification performances and the model. Figure 3 shows the impacts of GMDH external criteria on classification performance of the SSFS-GMDH model in the four datasets, namely the Mianyang, Nanchong, Yibin, and Panzhihua datasets.

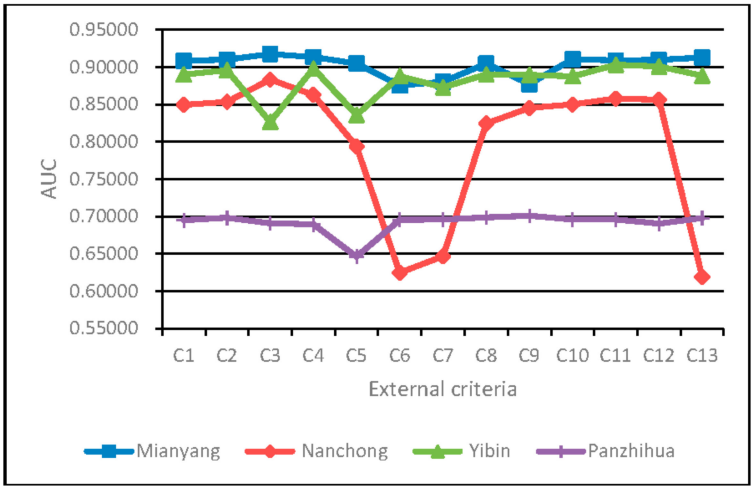

Figure 3. Impacts of GMDH external criteria on classification performance of the SSFS-GMDH model in the four datasets.

As shown in Figure 3, the SSFS-GMDH3 model on Mianyang dataset has the highest classification accuracy with a MAUC (mean AUC) value of 0.91748, followed by the SSFS-GMDH4 model with a MAUC value of 0.91339, and the SSFS-GMDH13 model with a MAUC value of 0.91265. The SSFS-GMDH6 model has the lowest classification accuracy, with a MAUC value of 0.87516. The SSFS-GMDH13 and SSFS-GMDH4 models belong to the accuracy criteria model, whereas the SSFS-GMDH13 and SSFS-GMDH6 models are the compatibility criteria model.

The SSFS-GMDH3 model on the Nanchong dataset has the highest classification accuracy with a MAUC value of 0.88320, followed by the SSFS-GMDH4 model with a MAUC value of 0.86295, and the SSFS-GMDH11 model with a MAUC value of 0.85785. The SSFS-GMDH13 model has the lowest classification accuracy with a MAUC value of 0.61927. The SSFS-GMDH13, SSFS-GMDH4, and SSFS-GMDH11 models are the accuracy criteria models, while the SSFS-GMDH13 model is the compatibility criteria model.

On the Yibin dataset, the SSFS-GMDH9 model has the highest classification accuracy with a MAUC value of 0.90309, followed by the SSFS-GMDH12 model with a MAUC value of 0.90071, and the SSFS-GMDH4 model with a MAUC value of 0.89814. The SSFS-GMDH3 model has the lowest classification accuracy, with a MAUC value of 0.82647. The SSFS-GMDH11, SSFS-GMDH12 and SSFS-GMDH4 models are the accuracy criteria model.

On the Panzhihua dataset, the SSFS-GMDH9 model has the highest classification accuracy with a MAUC value of 0.70092, followed by the SSFS-GMDH8 model with a MAUC value of 0.69865, and the SSFS-GMDH2 model with a MAUC value of 0.64622. The SSFS-GMDH5 model has the lowest classification accuracy with a MAUC value of 0.64622. The SSFS-GMDH9 and SSFS-GMDH8 models are the compatibility criteria model, and the SSFS-GMDH2 and SSFS-GMDH5 models are the accuracy criteria model.

To further examine the impacts of external criteria on classification performance of the SSFS-GMDH model, an analysis of the variance is conducted, and the results are shown in Table 5.

The SSFS-GMDH3 and SSFS-GMDH4 models perform better compared to other models on the Mianyang dataset. In particular, the MAUC value for the SSFS-GMDH3 model is the largest on the Mianyang dataset. The *p*-value for the SSFS-GMH3 and SSFS-GMDH4 significance tests is 0.073, which exceeds the significance level of 0.05. The *p*-values for the SSFS-GMDH3 and the significance tests for the other eleven external criteria are less than 0.05, and are therefore statistically different. Similarly, the SSFS-GMDH3 model is also superior to other models on the Nanchong dataset.

The SSSFS-GMDH11 model performs better compared to other models on the Panzhihua dataset, because it has the largest MAUC value. In addition, the *p*-values for the SSFS-GMDH11, SSFS-GMDH2, SSFS-GMDH4, and SSFS-GMDH12 significance testing are respectively 0.118, 0.297, and 0.616, exceeding the significance level of 0.05. On the Yibin dataset, the SSFS-GMDH9, SSFS-GMDH1, SSFS-GMDH2, SSFS-GMDH6, SSFS-GMDH7, SSFS-GMDH8, SSFS-GMDH10, SSFS-GMDH11 and SSFS-GMDH13 models have the same performance according to the statistical testing. Therefore, the SSSFS-GMDH11 model has the better robust capability on the Panzhihua and Yibin dataset.

Overall, the above analysis indicates that the performance of the SSFS-GMDH3 (average regularization criteria) is superior to the other external criteria when the dataset has a large sample size. The SSFS-GMDH 11 (asymmetric regularization criteria) is superior when the sample size is small. Therefore, the SSFS-GMDH3 and SSFS-GMDH11 models have better robustness and thus are chosen to be applied into the electricity peak load classification forecasting issue.

Table 5. Analysis of variance.

Mianyang, China													
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
C1		0.477	0.000	0.024	0.117	0.000	0.000	0.104	0.000	0.300	0.777	0.582	0.053
C2			0.001	0.122	0.023	0.000	0.000	0.020	0.000	0.745	0.669	0.872	0.222
C3				0.073	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.000	0.034
C4					0.000	0.000	0.000	0.000	0.000	0.222	0.049	0.088	0.746
C5						0.000	0.000	0.953	0.000	0.009	0.065	0.034	0.000
C6							0.023	0.000	0.366	0.000	0.000	0.000	0.000
C7								0.000	0.169	0.000	0.000	0.000	0.000
C8									0.000	0.008	0.056	0.030	0.000
C9										0.000	0.000	0.000	0.000
C10											0.452	0.627	0.370
C11												0.790	0.099
C12													0.167
C13													

Table 5. *Cont.*

Nanchong, China													
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
C1		0.608	0.000	0.080	0.000	0.000	0.000	0.001	0.595	0.940	0.276	0.376	0.000
C2			0.000	0.216	0.000	0.000	0.000	0.000	0.296	0.661	0.564	0.710	0.000
C3				0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
C4					0.000	0.000	0.000	0.000	0.023	0.094	0.508	0.386	0.000
C5						0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C6							0.005	0.000	0.000	0.000	0.000	0.000	0.476
C7								0.000	0.000	0.000	0.000	0.000	0.000
C8									0.007	0.001	0.000	0.000	0.000
C9										0.544	0.105	0.157	0.000
C10											0.310	0.418	0.000
C11												0.837	0.000
C12													0.000
C13													
Panzhihua, China													
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
C1		0.235	0.000	0.087	0.000	0.636	0.000	0.995	0.904	0.599	0.006	0.025	0.745
C2			0.000	0.602	0.000	0.097	0.000	0.238	0.191	0.087	0.118	0.288	0.131
C3				0.000	0.047	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C4					0.000	0.029	0.000	0.089	0.067	0.025	0.297	0.589	0.042
C5						0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C6							0.002	0.631	0.724	0.958	0.001	0.007	0.882
C7								0.000	0.000	0.002	0.000	0.000	0.001
C8									0.899	0.595	0.006	0.025	0.740
C9										0.685	0.004	0.018	0.838
C10											0.001	0.006	0.841
C11												0.616	0.002
C12													0.010
C13													
Yibin, China													
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
C1		0.447	0.359	0.230	0.000	0.878	0.822	0.407	0.182	0.820	0.835	0.316	0.508
C2			0.093	0.050	0.000	0.543	0.592	0.946	0.567	0.593	0.580	0.078	0.921
C3				0.777	0.000	0.284	0.253	0.081	0.024	0.253	0.261	0.932	0.114
C4					0.000	0.176	0.154	0.043	0.011	0.153	0.159	0.843	0.063
C5						0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C6							0.943	0.500	0.238	0.941	0.957	0.248	0.611
C7								0.546	0.267	0.999	0.986	0.220	0.662
C8									0.613	0.548	0.535	0.067	0.868
C9										0.268	0.260	0.020	0.501
C10											0.985	0.219	0.664
C11												0.226	0.650
C12													0.096
C13													

4.4. Analysis of the Parameter Sensitivity

θ and K are two essential parameters in the SSFS-GMGH model proposed in this paper. The two parameters need to be determined to achieve better performance. In the following section, the impact of θ and K on model performance is analyzed.

(1) Impacts of θ on model performance

Suppose that θ = 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100%. We set randomly K = 5, and the experimental results for the SSFS-GMDH3 model in the four datasets are shown in Figure 4.

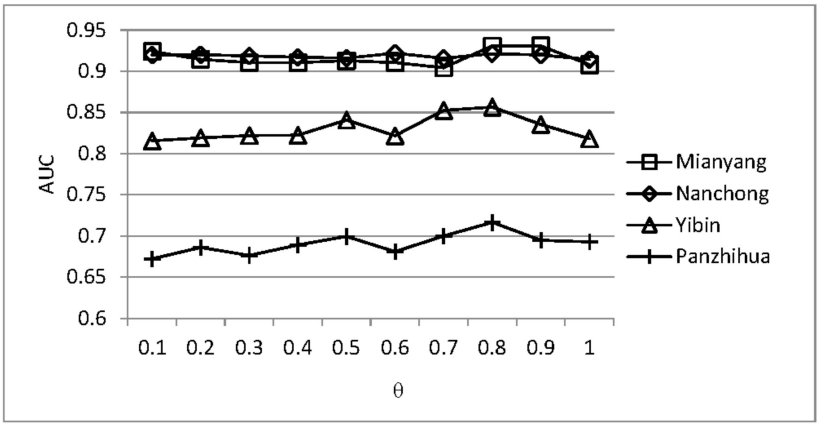

Figure 4. The performances of SSFS-GMDH3 model with different θ values.

As can be seen in Figure 4, the SSFS-GMDH3 model's performance in the four datasets gradually reaches a peak and then declines. On the Mianyang dataset, when θ reaches 0.9, the model performance is the best with a MAUC value of 0.9309. The corresponding MAUC value is 0.9308 when θ equals 0.8, so the small discrepancy can be overlooked. On the Panzhihua and Yibin datasets, when $\theta = 0.8$, both model performances are optimal. On the Nanchong dataset, parameter θ has little impact on the model performance. When θ reaches 0.6 and 0.8, the MAUC values are 0.9219 and 0.9215, respectively. Therefore, the paper suggests setting the $\theta = 0.8$.

(2) Impacts of K on model performance

The experimental results for the SSFS-GMDH3 model in the four datasets are shown in Figure 5 with $\theta = 0.8$, and $K \in [2, 15]$.

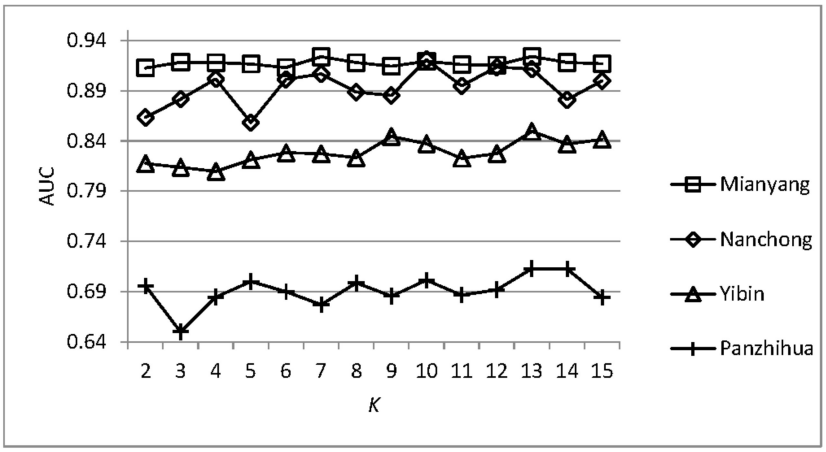

Figure 5. The performance of SSFS-GMDH3 model with different K values.

Figure 5 indicates that, with an increase in K, the MAUC value first has a fluctuating increasing tendency, after which it slowly declines. When $K = 13$, the best model performances are achieved in Mianyang, Panzhihua and Yibin datasets. When $K = 10$, the model has an optimal performance on Nanchong dataset, with an MAUC value of 0.9209. When $K = 13$, the MAUC value is 0.9114. Therefore, the SSFS-GMDH3 model has the best performance only when $\theta = 0.8$ and $K = 10$.

4.5. Comparisons with Other Models

Table 6 shows the MAUC value of the SSFS-GMDH, FW-SemiFS, and GMDH-U models on the four datasets. Symbols $\downarrow, \uparrow, \|$ indicate that the result is significantly worse, better, and similar to

that obtained by the SSFS-GMDH3 model, respectively. Symbols $\sim, +, \approx$ denote that the result is significantly worse, better, and similar to that obtained by the SSFS-GMDH11 model, respectively. On the Mianyang dataset, the MAUC values for the SSFS-GMDH3, SSFS-GMDH11, FW-SemiFS and GMDH-U are, respectively, 0.9452, 0.9381, 0.9308 and 0.9218. Therefore, the SSFS-GMDH model performs much better than the FW-SemiFS.

Table 6. Comparison between the SSFS-GMDH3, FW-SemiFS, and GMDH-U models in the four datasets.

Classification Model	Mianyang	Nanchong	Yibin	Panzhihua
SSFS-GMDH3	0.9452	0.9160	0.8640	0.7188
SSFS-GMDH11	0.9381	0.8621	0.9077	0.7491
FW-SemiFS	0.9308 $\downarrow \approx$	0.8520 $\downarrow \sim$	0.8419 $\downarrow \downarrow$	0.6188 $\downarrow \downarrow$
GMDH-U	0.9218 $\sim \downarrow$	0.8548 $\downarrow \sim$	0.8611 $\approx \downarrow$	0.6476 $\downarrow \downarrow$

On the Nanchong and Panzhihua datasets, the performance of the SSFS-GMDH model is superior to that of both the FW-SemiFS and GMDH-U models. Overall, the performance of the SSFS-GMDH model is the best compared to the FW-Semi FS and the GMDH-U models.

5. Conclusions

This paper investigates a day-ahead electricity peak load classification forecasting problem. It transforms the conventional continuous forecasting into a novel interval forecasting, and then further converts the interval forecasting into the classification forecasting. In addition, an indicator system influencing the electricity load is established from three dimensions, namely the load series, calendar data, and weather data. A novel semi-supervised feature selection algorithm based on the group method of data handling technology is proposed to address the electricity load classification forecasting problem. Furthermore, the parameters of the proposed model and the external criteria are analyzed systematically, which aims to improve the robustness of proposed model. An empirical test in real-world peak load forecasting cases shows that the proposed method has better classification forecasting performance compared to the two other state-of-the-art methods in four typical datasets, and that the peak classification forecasting problem is solved effectively. It is evident that the time interval in this paper is one day, but investigating different time intervals according to practical scheduling tasks, ranging from one hour to one week, and comparing the subtle difference are of great importance in the future. It is also urgent for researchers to develop more methods solving the short-term load classification forecasting issues.

Author Contributions: Lintao Yang proposed the problem and obtained the empirical data. Hongyan Yang and Honggeng Yang established the indicator system and wrote the initial manuscript. Haitao Liu studied and completed the proposed model. All authors read and approved the final manuscript.

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

References

1. Chen, B.-J.; Chang, M.-W. Load forecasting using support vector machines: A study on eunite competition 2001. *IEEE Trans. Power Syst.* **2004**, *19*, 1821–1830. [CrossRef]
2. Fan, G.-F.; Peng, L.-L.; Hong, W.-C.; Sun, F. Electric load forecasting by the svr model with differential empirical mode decomposition and auto regression. *Neurocomputing* **2016**, *173*, 958–970. [CrossRef]
3. Andersen, F.M.; Larsen, H.V.; Gaardestrup, R.B. Long term forecasting of hourly electricity consumption in local areas in denmark. *Appl. Energy* **2013**, *110*, 147–162. [CrossRef]
4. De Felice, M.; Alessandri, A.; Catalano, F. Seasonal climate forecasts for medium-term electricity demand forecasting. *Appl. Energy* **2015**, *137*, 435–444. [CrossRef]

5. Taylor, J.W.; McSharry, P.E. Short-term load forecasting methods: An evaluation based on european data. *IEEE Trans. Power Syst.* **2007**, *22*, 2213–2219. [CrossRef]
6. Hong, T. *Short Term Electric Load Forecasting*; North Carolina State University: Raleigh, NC, USA, 2010; pp. 3–6.
7. Liu, J.-P.; Li, C.-L. The short-term power load forecasting based on sperm whale algorithm and wavelet least square support vector machine with DWT-IR for feature selection. *Sustainability* **2017**, *9*, 1188. [CrossRef]
8. Taylor, J.W. An evaluation of methods for very short-term load forecasting using minute-by-minute british data. *Int. J. Forecast.* **2008**, *24*, 645–658. [CrossRef]
9. Hong, T.; Fan, S. Probabilistic electric load forecasting: A tutorial review. *Int. J. Forecast.* **2016**, *32*, 914–938. [CrossRef]
10. Hippert, H.S.; Pedreira, C.E.; Souza, R.C. Neural networks for short-term load forecasting: A review and evaluation. *IEEE Trans. Power Syst.* **2001**, *16*, 44–55. [CrossRef]
11. Hong, T.; Gui, M.; Baran, M.E.; Willis, H.L. Modeling and forecasting hourly electric load by multiple linear regression with interactions. In Proceedings of the Power and Energy Society General Meeting, Providence, RI, USA, 25–29 July 2010; pp. 1–8.
12. Wang, Y.; Xia, Q.; Kang, C. Secondary forecasting based on deviation analysis for short-term load forecasting. *IEEE Trans. Power Syst.* **2011**, *26*, 500–507. [CrossRef]
13. Ceperic, E.; Ceperic, V.; Baric, A. A strategy for short-term load forecasting by support vector regression machines. *IEEE Trans. Power Syst.* **2013**, *28*, 4356–4364. [CrossRef]
14. Paparoditis, E.; Sapatinas, T. Short-term load forecasting: The similar shape functional time-series predictor. *IEEE Trans. Power Syst.* **2013**, *28*, 3818–3825. [CrossRef]
15. Chitsaz, H.; Shaker, H.; Zareipour, H.; Wood, D.; Amjady, N. Short-term electricity load forecasting of buildings in microgrids. *Energy Build.* **2015**, *99*, 50–60. [CrossRef]
16. Ju, F.-Y.; Hong, W.-C. Application of seasonal svr with chaotic gravitational search algorithm in electricity forecasting. *Appl. Math. Model.* **2013**, *37*, 9643–9651. [CrossRef]
17. Desha, C.J.K.; Smith, M.; Hargroves, K.J.; Stasinopoulos, P.; Stephens, R. *Energy Transformed: Sustainable Energy Solutions for Climate Change Mitigation*; The Natural Edge Project, CSIRO, and Griffith University: Brisbane, Australia, 2007.
18. Staff, G.B. *Unlocking Energy Efficiency in the U.S. Economy*; Mckinsey & Company: Chicago, IL, USA, 2009.
19. Bessec, M.; Fouquau, J. Short-run electricity load forecasting with combinations of stationary wavelet transforms. *Eur. J. Oper. Res.* **2018**, *264*, 149–164. [CrossRef]
20. Feng, Y.; Ryan, S.M. Day-ahead hourly electricity load modeling by functional regression. *Appl. Energy* **2016**, *170*, 455–465. [CrossRef]
21. Tong, C.; Li, J.; Lang, C.; Kong, F.; Niu, J.; Rodrigues, J.J.P.C. An efficient deep model for day-ahead electricity load forecasting with stacked denoising auto-encoders. *J. Parallel Distrib. Comput.* **2017**, in press. [CrossRef]
22. Amjady, N. Short-term hourly load forecasting using time-series modeling with peak load estimation capability. *IEEE Trans. Power Syst.* **2001**, *16*, 498–505. [CrossRef]
23. Okoboi, G.; Mawejje, J. Electricity peak demand in uganda: Insights and foresight. *Energy Sustain. Soc.* **2016**, *6*, 29. [CrossRef]
24. Alani, A.Y.; Osunmakinde, I.O. Short-term multiple forecasting of electric energy loads for sustainable demand planning in smart grids for smart homes. *Sustainability* **2017**, *9*, 1972. [CrossRef]
25. Kim, Y.-J. Comparison between inverse model and chaos time series inverse model for long-term prediction. *Sustainability* **2017**, *9*, 982. [CrossRef]
26. Zhang, Z.; Song, Y.; Liu, F.; Liu, J. Daily average wind power interval forecasts based on an optimal adaptive-network-based fuzzy inference system and singular spectrum analysis. *Sustainability* **2016**, *8*, 125. [CrossRef]
27. Hu, Y.-C. Nonadditive grey prediction using functional-link net for energy demand forecasting. *Sustainability* **2017**, *9*, 1166. [CrossRef]
28. Weron, R. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *Int. J. Forecast.* **2014**, *30*, 1030–1081. [CrossRef]
29. Andrade, J.R.; Filipe, J.; Reis, M.; Bessa, R.J. Probabilistic price forecasting for day-ahead and intraday markets: Beyond the statistical model. *Sustainability* **2017**, *9*, 1990. [CrossRef]

30. Box, G.E.P.; Jenkins, G.M. *Time Series Analysis: Forecasting and Control*; John Wiley & Sons: Oakland, CA, USA, 1976; Volume 31, p. 303.
31. Soares, L.J.; Medeiros, M.C. Modeling and forecasting short-term electricity load: A comparison of methods with an application to brazilian data. *Int. J. Forecast.* **2008**, *24*, 630–644. [CrossRef]
32. Cincotti, S.; Gallo, G.; Ponta, L.; Raberto, M. Modeling and forecasting of electricity spot-prices: Computational intelligence vs. classical econometrics. *AI Commun.* **2014**, *27*, 301–314.
33. Amjady, N.; Keynia, F. Day ahead price forecasting of electricity markets by a mixed data model and hybrid forecast method. *Int. J. Electr. Power Energy Syst.* **2008**, *30*, 533–546. [CrossRef]
34. Mori, H.; Takahashi, A. Hybrid intelligent method of relevant vector machine and regression tree for probabilistic load forecasting. In Proceedings of the 2011 2nd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies (ISGT Europe), Manchester, UK, 5–7 December 2011; pp. 1–8.
35. Xiong, T.; Bao, Y.; Hu, Z. Interval forecasting of electricity demand: A novel bivariate emd-based support vector regression modeling framework. *Int. J. Electr. Power Energy Syst.* **2014**, *63*, 353–362. [CrossRef]
36. Dag, O.; Yozgatligil, C. Gmdh: An R package for short term forecasting via gmdh-type neural network algorithms. *R J.* **2016**, *8*, 379–386.
37. Chen, L.-G.; Chiang, H.-D.; Dong, N.; Liu, R.-P. Group-based chaos genetic algorithm and non-linear ensemble of neural networks for short-term load forecasting. *IET Gener. Transm. Distrib.* **2016**, *10*, 1440–1447. [CrossRef]
38. Ratrout, N. Short-term traffic flow prediction using group method data handling (gmdh)-based abductive networks. *Arab. J. Sci. Eng.* **2014**, *39*, 631–646. [CrossRef]
39. Kim, D.; Seo, S.-J.; Park, G.-T. Hybrid gmdh-type modeling for nonlinear systems: Synergism to intelligent identification. *Adv. Eng. Softw.* **2009**, *40*, 1087–1094. [CrossRef]
40. Ivakhnenko, A.G.; Ivakhnenko, G.A. The review of problems solvable by algorithms of the group method of data handling (gmdh). *Pattern Recognit. Image Anal.* **1995**, *5*, 527–535.
41. Ivakhnenko, A.G.; Ivakhnenko, G.A. Problems of further development of the group method of data handling algorithms. *Pattern Recognit. Image Anal.* **2000**, *10*, 187–194.
42. Shaghaghi, S.; Bonakdari, H.; Gholami, A.; Ebtehaj, I.; Zeinolabedini, M. Comparative analysis of gmdh neural network based on genetic algorithm and particle swarm optimization in stable channel design. *Appl. Math. Comput.* **2017**, *313*, 271–286. [CrossRef]
43. Xiao, J.; He, C.; Jiang, X.; Liu, D. A dynamic classifier ensemble selection approach for noise data. *Inf. Sci.* **2010**, *180*, 3402–3421. [CrossRef]
44. Xiao, J.; He, C.; Jiang, X. Structure identification of bayesian classifiers based on gmdh. *Knowl.-Based Syst.* **2009**, *22*, 461–470. [CrossRef]
45. McAfee, A.; Brynjolfsson, E.; Davenport, T.H. Big data: The management revolution. *Harv. Bus. Rev.* **2012**, *90*, 60–68. [PubMed]
46. Xiao, J.; Cao, H.; Jiang, X.; Gu, X.; Xie, L. Gmdh-based semi-supervised feature selection for customer classification. *Knowl.-Based Syst.* **2017**, *132*, 236–248. [CrossRef]
47. Takeda, H.; Tamura, Y.; Sato, S. Using the ensemble kalman filter for electricity load forecasting and analysis. *Energy* **2016**, *104*, 184–198. [CrossRef]
48. Bauer, M.; Scartezzini, J.L. A simplified correlation method accounting for heating and cooling loads in energy-efficient buildings. *Energy Build.* **1998**, *27*, 147–154. [CrossRef]
49. Wang, Y.; Bielicki, J.M. Acclimation and the response of hourly electricity loads to meteorological variables. *Energy* **2018**, *142*, 473–485. [CrossRef]
50. Sailor, D.J.; Muñoz, J.R. Sensitivity of electricity and natural gas consumption to climate in the U.S.A.—Methodology and results for eight states. *Energy* **1997**, *22*, 987–998. [CrossRef]
51. Valor, E.; Meneu, V.; Caselles, V. Daily air temperature and electricity load in spain. *J. Appl. Meteorol.* **2001**, *40*, 1413–1421. [CrossRef]
52. Ren, J.; Qiu, Z.; Fan, W.; Cheng, H.; Philip, S.Y. Forward semi-supervised feature selection. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining; Osaka, Japan, 20–23 May 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 970–976.

53. Lee, K.; Park, I.; Yoon, B. An approach for r&d partner selection in alliances between large companies, and small and medium enterprises (smes): Application of bayesian network and patent analysis. *Sustainability* **2016**, *8*, 117.
54. Tso, G.K.; Yau, K.K. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy* **2007**, *32*, 1761–1768. [CrossRef]
55. Huang, N.; Zhang, S.; Cai, G.; Xu, D. Power quality disturbances recognition based on a multiresolution generalized s-transform and a pso-improved decision tree. *Energies* **2015**, *8*, 549–572. [CrossRef]
56. Webb, G.I.; Ting, K.M. On the application of roc analysis to predict classification performance under varying class distributions. *Mach. Learn.* **2005**, *58*, 25–32. [CrossRef]