entropy

MDPI

Article

Prediction Model of the Power System Frequency Using a Cross-Entropy Ensemble Algorithm

Yi Tang, Han Cui and Qi Wang *

Department of Electrical Engineering, Southeast University, Nanjing 210096, China; tangyi@seu.edu.cn (Y.T.); cuihan@seu.edu.cn (H.C.)

* Correspondence: wangqi@seu.edu.cn; Tel.: +86-186-5291-2305

Received: 20 September 2017; Accepted: 17 October 2017; Published: 19 October 2017

Abstract: Frequency prediction after a disturbance has received increasing research attention given its substantial value in providing a decision-making foundation in power system emergency control. With the advancing development of machine learning, analysis power systems with machine-learning methods has become completely different from traditional approaches. In this paper, an ensemble algorithm using cross-entropy as a combination strategy is presented to address the trade-off between prediction accuracy and calculation speed. The prediction difficulty caused by inadequate numbers of severe disturbance samples is also overcome by the ensemble model. In the proposed ensemble algorithm, base learners are selected following the principle of diversity, which guarantees the ensemble algorithm's accuracy. Cross-entropy is applied to evaluate the fitting performance of the base learners and to set the weight coefficient in the ensemble algorithm. Subsequently, an online prediction model based on the algorithm is established that integrates training, prediction and updating. In the Western System Coordinating Council 9-bus (WSCC 9) system and the Institute of Electrical and Electronics Engineers 39-bus (IEEE 39) system, the algorithm is shown to significantly improve the prediction accuracy in both sample-rich and sample-poor situations, verifying the effectiveness and superiority of the proposed ensemble algorithm.

Keywords: frequency prediction; machine learning; cross-entropy; ensemble algorithm

1. Introduction

With the construction of the alternating current (AC)/direct current (DC) hybrid power grid, grid operating characteristics have undergone fundamental changes. AC/DC transmission line failure may cause transmission power to fluctuate greatly, so that the problem of frequency stability is exacerbated [1]. One measure to reduce the risk of a frequency problem occurring is to perform online frequency prediction to determine whether the system is operating under an urgent condition subject to potential power disturbance [2]. Typically, the dynamic process of power system frequency involves multiple time scales ranging from milliseconds to hours. In this paper, frequency response in the emergent state, mainly in seconds, is focused on without considering automatic generation control (AGC).

The factors that affect the dynamic characteristics of frequency after a disturbance primarily include fault type, fault location, the operation state of the power grid, generator and load parameters and network topology. Currently, the analysis of the dynamic frequency response is performed using time-domain (T-D) simulation, equivalence models and machine learning.

The T-D method depicts the entire power system with high-order nonlinear equations to accurately obtain the frequency dynamic change process of each node of the grid. However, the calculation speed is slow. Thus, the method is unsuited for rapid prediction of the frequency after a disturbance.

The equivalence model method, which is represented by average system frequency (ASF) and system frequency response (SFR), is a mainstream method of online application with limited precision due to excessive simplification [3]. Based on physical models, both methods involve a trade-off between calculation speed and precision.

Different from these two methods, machine learning, as a model-free method, is devoted to analyzing the numerical relevance among operation state variables and research targets using data science and computer science [4]. Machine learning has solved many problems in various fields, such as animal behavior detection [5,6], emotion recognition [7] and face detection [8]. With the rapid development of smart grids, power system data are exhibiting explosive growth [9], enabling the application of machine learning to power systems in today's context of "big data". A smart grid provides detailed information that can optimize uncertain parameters and improve the stability of the microgrid [10]. As an emerging method in power system stability analysis, machine learning has been successfully applied to many problems, such as power angle stability [11,12], voltage stability [13,14] and frequency stability [15,16]. When high-quality and large-quantity samples are available, machine learning can achieve a satisfactory balance between accuracy and speed [17].

Research on power grid frequency prediction using machine-learning algorithms has made progress in theory and application. Frequency prediction using neural nets has been implemented and the method of feature election analyzed [18]. A network framework of power-system frequency prediction has been established to acquire real-time data [19]. Approximate prediction by curve fitting has also been investigated and shown to support under-frequency load shedding [20], while there are still ideal assumptions, such as adequate training samples and the simplification of frequency curves.

Power grid failures are small-probability events. Therefore, the number of failure samples is inadequate for machine-learning training [21]. Consequently, the potential under-fitting problem restricts the application of machine learning to post-contingency frequency prediction. To solve the problem caused by the inadequacy of severe disturbance samples, an ensemble algorithm based on cross-entropy is proposed to predict the minimum frequency after a disturbance. Cross-entropy can be used as a mathematical evaluation of the fitting problem by calculating the generalized distance between two datasets to characterize the similarity between the two data distributions [22]. In the ensemble algorithm, cross-entropy is applied to evaluate base learners as the basis for quantifying the weight index of each base learner. The outputs of various base learners are merged to fully excavate the information contained in a sample and realize fast, reliable prediction of power system frequency in the event of sample scarcity. Normally, base learners, such as artificial neural networks (ANNs) and support vector regression (SVR), are applied to the fitting problem in several steps, including data acquisition, training and verification [23].

The primary contribution of this paper consists of three aspects: the problem we investigate, the method used and the conclusion. In this paper, the scarcity of severe disturbances is first considered when predicting frequency drift. The ensemble algorithm using cross-entropy is used to enhance accuracy. Finally, we demonstrate that the proposed algorithm improves prediction accuracy compared to base learners. The remainder of the paper is organized as follows. In Section 2, the trade-offs in frequency prediction are discussed. The ensemble algorithm based on cross-entropy is presented in Section 3. In Section 4, a frequency prediction model is established and then verified in Section 5 using a case study. The conclusions are presented in Section 6.

2. Trade-Offs in Frequency Prediction Methods

The lowest frequency after a power-system disturbance is an important indicator of the system's security and stability. Therefore, the prediction of the lowest frequency is of substantial value to auxiliary control. Considering trade-offs between calculation speed and accuracy as well as between prediction accuracy and value in the existing prediction method, this paper adopts a machine-learning algorithm termed ensemble learning to predict frequency (Figure 1).

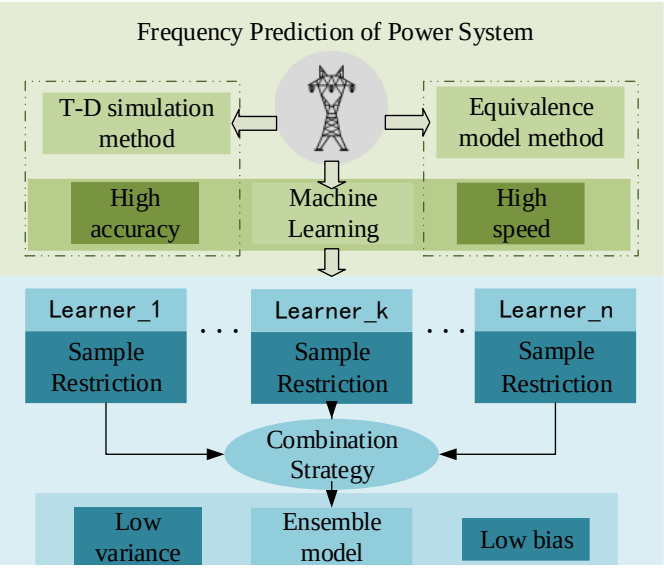

Figure 1. Motivation of ensemble learning in frequency prediction.

2.1. Trade-Off: Calculation Speed versus Prediction Accuracy

A power system requires a high-dimensional nonlinear differential equation to accurately describe its dynamic characteristics, which results in the computational complexity of T-D simulation. Therefore, it is difficult to precisely predict system frequency in a short time after contingencies.

At present, the usual means to reduce the complexity of the power system is primarily to simplify the power system model to quickly predict the lowest frequency of the system. This method ignores a large number of factors that influence frequency, and, as a result, the equivalent model method is unreliable. In addition, the decision-maker's risk of being misled by inaccurate predictions is increased, which affects the security and stability of the system. Based on a historical dataset, machine learning maps system state variables into the lowest frequency. Instead of simplifying influential indicators, machine learning introduces as many attributes as possible into a training period. Thus, a physical model's advantage of high accuracy is inherited. Once the prediction model is well trained, the model's calculation speed is extremely fast (i.e., a matter of milliseconds), which gives a machine-learning model an advantage over a physical model.

2.2. Trade-Off: Calculation Speed versus Prediction Value

Training a typical machine-learning model relies on having an adequate number of training samples. Otherwise, the fitting error of the physical system is largely biased.

The operating point (OP) of a power system is highly centralized. That is, OPs surround a feasible area. This feature explains why severe disturbance samples are rare but small disturbance samples are numerous. Based on a highly imbalanced dataset, the machine-learning algorithm fails to produce credibly predict severe post-contingency deviations of power system frequency. Owing to the adequacy of small disturbance samples, reliable predictions are easily made by normal learners. However, a biased OP indicates a severe fault in the power grid, where the lowest frequency prediction should play an important role in post-contingency control and restoration. That is, the prediction value at biased OPs is much higher than that at normal OPs, while prediction accuracy is exactly the reverse.

In this paper, an ensemble model is adopted as a substitute for a single machine-learning algorithm to improve prediction accuracy when data are inadequate. The ensemble model strategically combines the output of multiple base learners to obtain a better fit result, which sacrifices system simplicity.

3. Cross-Entropy Based Ensemble Algorithm

Theoretically, ensemble learning obtains a better prediction result by the integration of base learners, which means that base learners are fundamental to the model. Therefore, an objective evaluation of the performance of the base learners is a prerequisite for this research, and a proper strategy to combine base learners is the core of the ensemble model.

3.1. Cross-Entropy Theory

Machine-learning regression algorithms are normally evaluated by the overall error of the test samples, whereby all samples are treated as possessing equal status and the particularity of each sample is ignored. However, the prediction value in under different circumstances varies considerably, with the result that each case is unique. Therefore, an evaluation index that considers the importance of individual samples is urgently needed. Power-system development is becoming increasingly related to informatics. Cross-entropy is an important concept in Shannon entropy that can be used to quantify the contribution of an individual sample in an entire dataset and which is suitable for the evaluation of machine-learning algorithms in this case.

Cross-entropy is an index used to measure the similarity of two distributions and can be regarded as the "distance" of two distributions. It should be emphasized that the cross-entropy method has been successfully applied to both combinatorial optimization and rare-event simulation.

Let p and q be two distributions in some space $\boldsymbol{X}$. Let $\{f(*)\}$ be a family of probability density functions (PDFs) on $\boldsymbol{X}$. Thus, cross-entropy is defined as follows:

$$D(p||q) = \int_{\mathrm{X}} f(p) \log \frac{f(p)}{f(q)} dx. \tag{1}$$

In a discrete case,

$$D(p||q) = \sum_{i=1}^{n} f(p_i) \log \frac{f(p_i)}{f(q_i)}. \tag{2}$$

It can be easily derived from Equations (1) and (2) that the more similar that p and q are, the smaller the cross-entropy. An extreme example is when p and q are exactly the same. Then, cross-entropy equals 0. It should also be noted that D is not a "distance" in the formal sense; for example, it is asymmetric. Thus,

$$D(p||q) \neq D(q||p). \tag{3}$$

Generally, p is the real distribution, and q is the approximation distribution. In optimization, the iterative method is used to modify the parameters of the approximate distribution q to reduce the cross-entropy approaching the real distribution p.

Cross-entropy is modified in this research to suit the features of frequency prediction. It is obvious that frequency prediction is a regression problem in that samples are discrete and the probability of each sample is hard to precisely calculate. The modified cross-entropy is defined as follows:

$$D(p,q) = \sum_{i=1}^{n} w_i \left| \log \frac{f(p_i)}{f(q_i)} \right|, \tag{4}$$

where $f(p_i)$ is the real frequency of the i-th sample, $f(q_i)$ is the prediction output of the i-th sample and w_i is the weight of the i-th sample.

As a global evaluation index, cross-entropy measures machine-learning algorithm accuracy while considering the distinction of an individual sample. Consequently, by setting the weight for samples, the algorithm improves the utilization efficiency of rare samples.

3.2. Combination Strategy of Ensemble Algorithm

Rather than finding the best hypothesis to fit samples, the ensemble learning algorithm requires several hypotheses to be integrated so that three improvements are achieved: decreased risk of over-fitting, avoidance of falling into the local optimum point and enlargement of the space of possible fitting hypotheses [24].

In this paper, weighted averaging is applied as the combination strategy, whereby the weight of base learners is determined by cross-entropy, which is sensitive to severe disturbance data. Compared with two other common combination strategies (i.e., simple averaging and meta learning), weighted averaging ensures better performance in frequency prediction. The reasons are listed as follows.

Simple averaging treats base learners equally without considering the different performance of each algorithm [25]. In doing so, prediction accuracy may be dragged down to an unacceptable level by poor learners. Meta learning is a two-layer structure in which primary learners are combined with a meta learner [26]. Naturally, the meta learner's training process requires many more samples than that of weighted averaging. In addition, in the frequency-prediction problem, the number of samples cannot meet the requirement of meta learning, and under-fitting may occur. In summary, weighted averaging should be selected as the combination strategy in this paper's learning task, and the weight factors are calculated by the modified cross-entropy method.

4. Proposed Prediction Model

In this section, a frequency prediction model is proposed in which the core algorithm is the weighted averaging ensemble learning based on modified cross-entropy. First, base learners are selected according to their characteristics and the diversity theory. Then, the ensemble algorithm is explained in detail. Finally, a robust online prediction model is provided. The model includes modules for offline training, real-time prediction and sample collection.

4.1. Selection of Base Learners

To guarantee prediction accuracy, five regression algorithms are introduced as base learner candidates: decision tree (DT), multivariable linear regression (MLR), artificial neural networks (ANN), least square support vector machine (LSSVM) and extreme learning machine (ELM) (Table 1).

Table 1. Comparison among regression algorithms.

Algorithm	Advantage	Disadvantage
Decision tree (DT)	Easy to understand and implement	Discrete output
Multivariable linear regression (MLR)	Easy optimization	Linear problem only
Artificial neural networks (ANN)	Strong generalization ability	Slow training rate
Least square support vector machine (LSSVM)	High accuracy	Parameter optimization
Extreme learning machine (ELM)	Quickness	Simple network structure

It can be concluded that there are substantial differences among the characteristics of each algorithm, which conforms to the diversity theory of ensemble learning. In theory, the number of base learners should be as large as possible to improve the diversity of the ensemble model. However, computational resources typically cannot satisfy the parallel operation of a large number of base learners. Considering accuracy and "ambiguity", which is used to describe diversity among regression algorithms, base learners are selected from the five candidates.

Let f be a regression task; $h_1, h_2, \ldots, h_n$ be candidate base learners; and H be the simple averaging model consisting of $h_1, h_2, \ldots, h_n$. For test dataset $\boldsymbol{x}$, the ambiguity of h_i is defined as follows:

$$A(h_i|\boldsymbol{x}) = (h_i(\boldsymbol{x}) - H(\boldsymbol{x}))^2. \tag{5}$$

Similarly, the ambiguity of H is written as follows:

$$\begin{aligned} \overline{A}(h|\boldsymbol{x}) &= \frac{1}{n}\sum_{i=1}^{n} A(h_i|\boldsymbol{x}) \\ &= \frac{1}{n}\sum_{i=1}^{n} (h_i(\boldsymbol{x}) - H(\boldsymbol{x}))^2 \end{aligned}. \tag{6}$$

The error formulas of the base learners and the ensemble model are as follows:

$$E(h_i|\boldsymbol{x}) = (f(\boldsymbol{x}) - h_i(\boldsymbol{x}))^2, \tag{7}$$

$$E(H|\boldsymbol{x}) = (f(\boldsymbol{x}) - H(\boldsymbol{x}))^2. \tag{8}$$

The average error of the base learners is derived as follows:

$$\overline{E}(h|\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} E(h_i|\boldsymbol{x}). \tag{9}$$

Thus,

$$\begin{aligned} \overline{A}(h|\boldsymbol{x}) &= \frac{1}{n}\sum_{i=1}^{n} E(h_i|\boldsymbol{x}) - E(H|\boldsymbol{x}) \\ &= \overline{E}(h|\boldsymbol{x}) - E(H|\boldsymbol{x}) \end{aligned}, \tag{10}$$

which can be rewritten as follows:

$$E(H|\boldsymbol{x}) = \overline{E}(h|\boldsymbol{x}) - \overline{A}(h|\boldsymbol{x}), \tag{11}$$

indicating that the error of the ensemble model is related to both the accuracy and the ambiguity of the base learners. More specifically, the ensemble model performs better when the accuracy or ambiguity of the base learners becomes higher.

4.1.1. Decision Tree (DT)

DT contains a classification tree and a regression tree. The latter is adopted in this learning task. The Gini index is used to determine the order of the internal nodes and defined as follows:

$$\mathrm{Gini}(D) = \sum_{i=1}^{n}\sum_{i\prime\neq i} p_i p_{i\prime} = 1 - \sum_{i=1}^{n} p_i^2, \tag{12}$$

where D is the dataset and p_i is the proportion of the i-th class. To feature a in D, the Gini index is written as follows:

$$\mathrm{Gini}(D,a) = \sum_{k=1}^{K} \frac{\left|D^k\right|}{|D|}\mathrm{Gini}(D^k), \tag{13}$$

where K is the total number of classes, $|D^k|$ is the number of the k-th class, $|D|$ is the number of all samples and $\mathrm{Gini}(D^k)$ is the Gini index of the k-th class in D. Once all features are sorted by the Gini index, it is convenient to establish a complete regression tree by arranging all the features sequentially.

4.1.2. Multivariable Linear Regression (MLR)

The MLR model is primarily used for multivariable problems, whereby the algebra method is employed for parameter optimization. Let y be the target variable and $x_1, x_2, \ldots, x_n$ be input variables. Then, a typical MLR model is as follows:

$$y = \sum_{i=1}^{n} k_i x_i + b, \tag{14}$$

where k_i and b are the parameters to be optimized. Regular optimization involves solving a set of normalized equations using the least squares method. Taking a two-variable regression problem as an example, the normalized equations are as follows:

$$\begin{cases} \sum y = nb_0 + b_1 \sum x_1 + b_2 \sum x_2 \\ \sum x_1 y = b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 \\ \sum x_2 y = b_0 \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2 \end{cases} . \tag{15}$$

All of the parameters can be easily derived by solving the equations.

4.1.3. Artificial Neural Networks (ANN)

In this paper, ANN is a two-layer feed-forward network that contains a hidden layer and an output layer [27]. Its structure is shown in Figure 2.

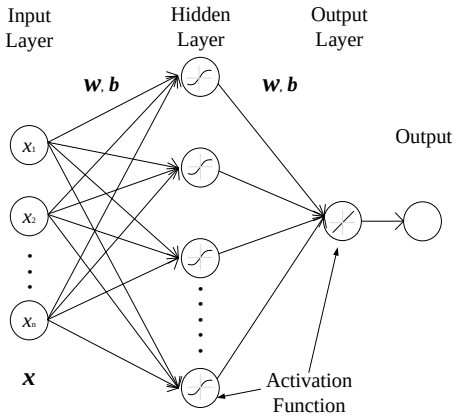

Figure 2. Two layer feed-forward network.

It has been demonstrated that ANN can fit the input–output relationship of arbitrary complexity and has excellent learning performance. Therefore, it is regarded as the most promising algorithm in the big-data era. Suppose an ANN output is expressed as follows:

$$y = f_2\left(\sum_{i=1}^{m} w_{2i} f_1\left(\sum_{i=1}^{n} w_{1i} x_i + b_{1i}\right) + b_{2i}\right). \tag{16}$$

In this expression, n is the number of hidden nodes, w_i is the weight of the i-th hidden node, b_i is the biased parameter of the i-th hidden node, x_i is the i-th input feature and f is an activation function.

The training process is essentially a matter of solving the optimal weight and biased parameter iteratively, minimizing the error between output and target by the gradient descent method.

4.1.4. Least Square Support Vector Machine (LSSVM)

LSSVM was proposed by Suykens et al. and is based on the support vector machine (SVM) to improve the applicability for a large dataset when computational complexity increases [28]. The cost function is represented by 2-norm in LSSVM, and least squares error fitting is adopted to fit the sample space, which is shown in Figure 3. In this paper, the regression form of LSSVM is implemented.

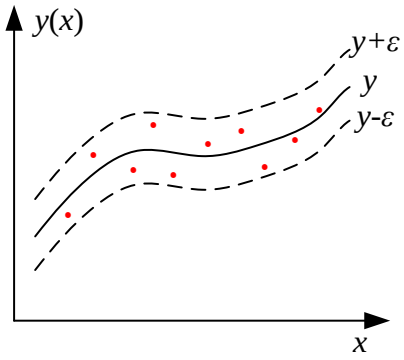

Figure 3. Regression of least square support vector machine (LSSVM).

For a specific learning task, assume the mathematical expression of LSSVM is as follows:

$$y(x) = \omega^T \varphi(x) + b. \tag{17}$$

Thus, the objective function is as follows:

$$\min_{\omega,e} \mathrm{J}(\omega, e) = \frac{1}{2}\omega^T\omega + \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2, \tag{18}$$

and the restriction is as follows:

$$y_k = \omega^T \varphi(x_k) + b + e_k, k = 1, \ldots, N. \tag{19}$$

All parameters can be solved by a Lagrange multiplier.

4.1.5. Extreme Learning Machine (ELM)

ELM is a single-layer neural network that has been studied and applied widely [29]. Compared to previous ANNs, the most significant advantage of ELM is that it generates parameters randomly instead of optimizing iteratively. Therefore, the training speed is extremely fast with little sacrifice of accuracy [30]. The structure of ELM is shown in Figure 4.

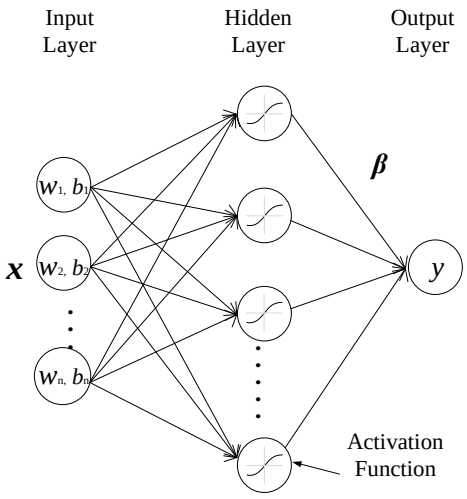

Figure 4. Regression of the structure of extreme learning machine (ELM).

The mathematical equation in the network can be written as follows:

$$\mathbf{H}\beta = \mathbf{T}, \tag{20}$$

where $\mathbf{H}$ is the hidden layer output matrix, β is the output weight vector connecting the hidden layer to the output layer and $\mathbf{T}$ is the target output matrix. It is reasonable to regard the ELM training process as a solution to linear equations. That is, output weight vector β is solved by

$$\hat{\beta} = \mathbf{H}^{\dagger}\mathbf{T}. \tag{21}$$

In Equation (21), $\mathbf{H}^{\dagger}$ is the Moor–Penrose generalized inverse of $\mathbf{H}$. In addition, the problem is transformed into solving the Moor–Penrose generalized inverse of $\mathbf{H}$.

4.2. Calculation of Weights

The ensemble learning algorithm proposed in this paper is based on the idea of the weighted average. The weight of each base learner is calculated by using modified cross-entropy. The final result of the ensemble learning is the weighted average of all base learners, thus reducing the prediction error.

First, we randomly select a training sample set for a base learner. According to the test results of each learner, we calculate the modified cross-entropy between the predicted result and the target using the following formula:

$$D(p,q) = \sum_{i=1}^{n} \Delta f \left| \log \frac{f(p_i)}{f(q_i)} \right|, \tag{22}$$

where f_p is the predicted frequency. The transformation from cross-entropy to ensemble weight requires the following formula:

$$W_i = \frac{1/D_i}{\sum\limits_{j=1}^{n} 1/D_j}. \tag{23}$$

In Equation (23), W_i is defined as the weight of the i-th base learner. According to the weight obtained and the output of the base learners, the weighted average is calculated, which is the result of the ensemble learning:

$$f_{ensemble} = \sum_{i=1}^{n} W_i \cdot f_i. \tag{24}$$

4.3. Online Frequency Prediction Model

An online frequency prediction model based on the cross-entropy ensemble learning algorithm is established, including offline training, online forecasting and real-time dataset updating (Figure 5).

The data source for offline training is the historical data of the power system operation, and data preprocessing is conducted to form the historical sample database. The base learners are trained with n groups of training samples. The training phase includes a verification and test process to ensure the generalization of the machine-learning algorithms. The cross-entropy of the trained learners is calculated to determine the base learners' weight. The n base learners form the final ensemble model by weighted averaging.

Online prediction is based on the acquisition of wide-area measurement information, which is processed into a formatted input vector. After data filtering and other procedures, the vector is introduced into the ensemble model. Then, the lowest frequency in the next time period is calculated and sent to the control center to assist dispatchers with decision-making. The frequency prediction result is updated in real time. The updating interval is determined by the speed of the prediction algorithm and the sampling frequency of the wide-area measurement information.

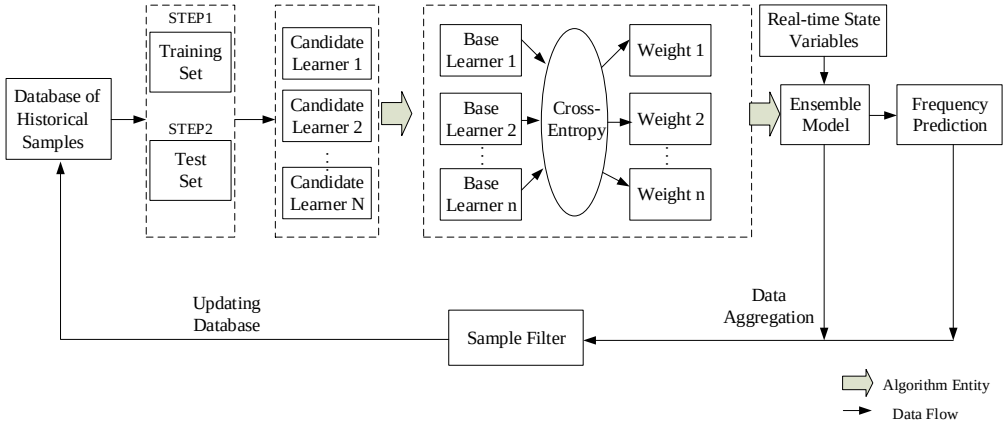

Figure 5. Structure of online frequency prediction model.

In Figure 6, sample-screening process is described, whereby the predicted frequency is expressed as f_p, real frequency is expressed as f_r and ε and δ are the judgment indicators. The sample-screening procedure updates unreliable prediction samples in the historical database, thereby improving the ability of the algorithm to correct error. In addition, the system can judge severe disturbance events. This type of event is processed into the database. Therefore, the forecasting model gradually improves the prediction accuracy of rare events. The updating frequency should not be too high because the training process of the ensemble algorithm is time-consuming.

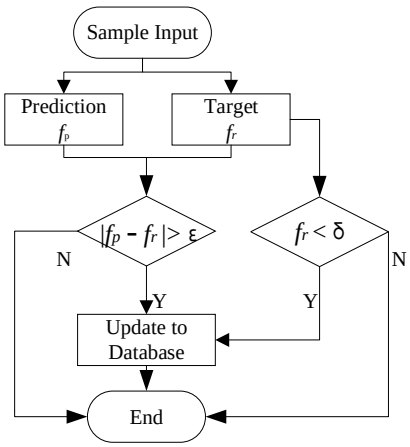

Figure 6. Screening procedures in sample filter.

Based on the three modules, the frequency prediction model is equipped with fast forecasting, self-correction and real-time updating.

5. Case Study

This section aims to predict the lowest frequency after a disturbance. A Western System Coordinating Council 9-bus (WSCC 9) system is used to compare the proposed algorithm with individual base learners under the condition of inadequate training samples, while an Institute of Electrical and Electronics Engineers 39-bus (IEEE 39) system is used as a test case for the relation between the number of samples and prediction accuracy. WSCC 9 system and IEEE 39 system are shown in Figure 7.

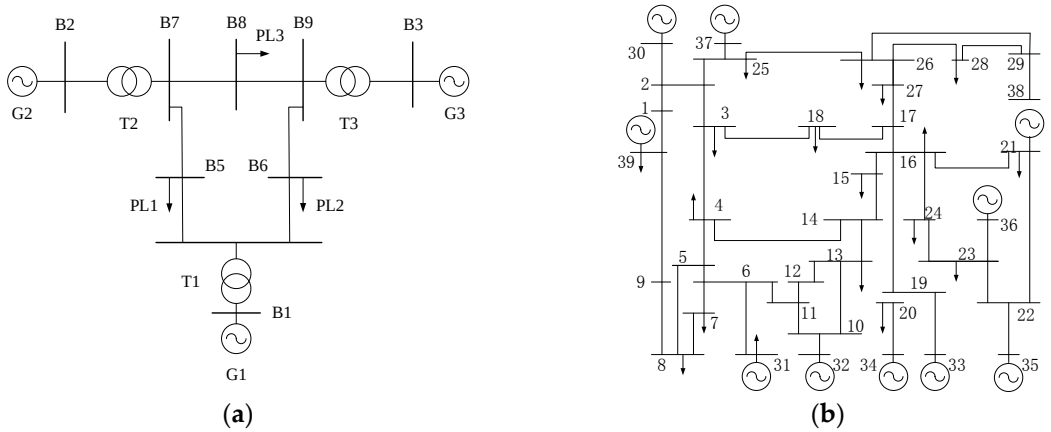

Figure 7. Test systems (**a**) Western System Coordinating Council 9-bus (WSCC 9) system; (**b**) Institute of Electrical and Electronics Engineers 39-bus (IEEE 39) system.

In both cases, power system toolbox in Matlab is used to generate training and test samples. For simplicity, all parameters are per-unit values. First, the WSCC 9 and IEEE 39 system models are built in software to simulate various operation states. The overall load is uniformly distributed from 0.8 to 1.2. The node injection power is a normal distribution in which variance is 0.1 and expectation is 1. The disturbance power is uniformly distributed from 0.1 to 1.2. After a complete simulation of the transient process, the lowest frequency is obtained by the software.

All the computations are performed on a PC with 2.2 GHz CPU and 8 G RAM. The error evaluation indicators are relative error e_r, mean of absolute error e_{MAE} and root mean square error e_{RMSE}:

$$e_r = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{f_p(x_i) - f_t(x_i)}{f_e - f_t(x_i)}\right| \times 100\%, \tag{25}$$

$$e_{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|f_p(x_i) - f_t(x_i)\right|, \tag{26}$$

$$e_{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f_p(x_i) - f_t(x_i)\right)^2}. \tag{27}$$

In the preceding equations, x_i represents the i-th input feature, f_p is the prediction model when f_t is the real relation between input and target, f_e is the rated frequency in the power system and n is the number of test samples.

5.1. Selection and Configuration of Base Learners

According to the evaluation method in Section 3, five regression algorithms are combined with a simple averaging method in both test systems. Each of the algorithms are configured as follows:

(1) Classification and regression tree (CART) is used to train the regression tree, which means the Gini index is calculated to determine the sequence of features. After a complete tree is formed, pruning is iteratively conducted to increase the generalization ability according to the number of leaves and errors.

(2) The significance level of MLR is set at 0.05, and the error is optimized using the least squares method.

(3) In ANN, the number of hidden nodes is set to be the average number of inputs and outputs. The maximum number of iterations is 1000, and training stops when 20 straight iterations do not improve the result by 5% (relative error). Tenfold cross-validation is applied to prevent over-fitting.

(4) *Gam* and *sig*2 are two parameters in LSSVM training. Leave-one-out cross-validation is commonly adopted as the optimization strategy.

(5) In the ELM configuration, the sigmoid function is used as the activation function because of its versatility. The number of hidden neurons is determined by dichotomy to avoid under- or over-fitting.

Finally, a simple averaging model is constructed based on the well-trained base learners. The evaluation results for WSCC 9 and IEEE 39 are shown in Tables 2 and 3, respectively.

Table 2. Evaluation of candidate learners in WSCC 9.

Base Learners	DT	MLR	ANN	LSSVM	ELM
Ambiguity	0.31	0.51	0.41	0.44	1.29
Error (%)	12.52	17.21	10.75	9.77	11.98

Table 3. Evaluation of candidate learners in IEEE 39.

Base Learners	DT	MLR	ANN	LSSVM	ELM
Ambiguity	0.23	0.37	0.52	0.63	0.97
Error (%)	16.52	21.09	15.10	16.32	15.94

In the WSCC 9 system, the ambiguity of DT is the smallest of the five algorithms, which has a negative influence on diversity. MLR has the largest error among the algorithms, decreasing overall accuracy. The situation is the same in the IEEE 39 system. Therefore, ANN, LSSVM and ELM are selected as the base learners of the proposed ensemble algorithm.

5.2. WSCC 9 System with Inadequate Samples

The WSCC 9 system contains nine nodes: three generator nodes and six load nodes. The number of training and test samples is 90 and 60, respectively. The trained base learners ANN, ELM and LSSVM perform the modified cross-entropy calculation. The results are shown in Figure 8.

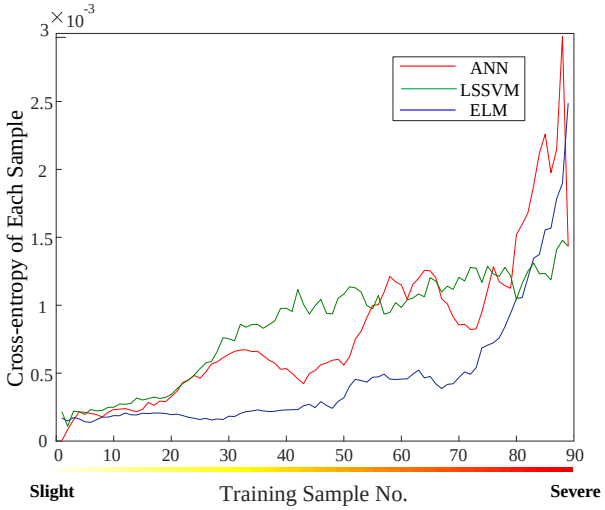

Figure 8. Calculation of modified cross-entropy in the WSCC system.

Samples are sorted by the severity of disturbance in Figure 8 such that the number of samples increases as the disturbance becomes more severe. It is easily observed in Figure 8 that the accuracy

of ELM, LSSVM and ANN varies in different samples. This phenomenon is a reflection of algorithm diversity. In Table 4, the weights of the base learners are derived from modified cross-entropy.

Table 4. Weights of base learners.

Base Learner	ANN	SVM	ELM
Weight	0.2047	0.4040	0.3913

Prediction error among the different algorithms is shown in Figure 9. A detailed error index is provided in Table 5. It is obvious in Figure 9 that the relative error of the ensemble algorithm is distributed more closely to zero than the others. According to Table 5, the prediction error of the three base learners is low. For example, relative error is between 9.77% and 11.98%. Analysis of the WSCC 9 system shows that the number of state variables is small and the power flow is relatively simple. Sufficient samples also contribute to a good prediction. Cross-entropy ensemble learning achieves a better result whose relative prediction error is only 6.67%. Considering the theory of ensemble learning, the algorithm reduces the prediction error by weighted averaging.

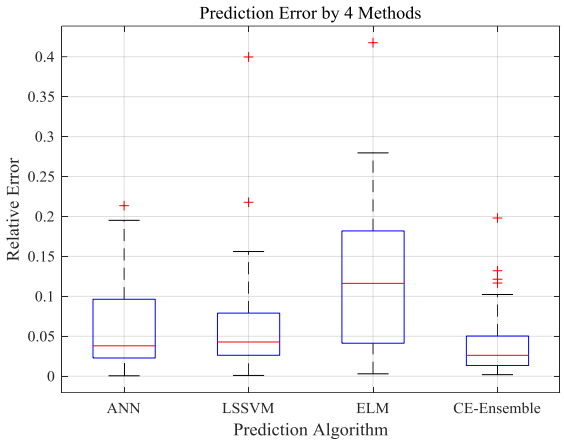

Figure 9. Prediction errors among different algorithms.

Table 5. Prediction error index in WSCC 9.

Algorithm	e_r	e_{MAE}	e_{RMSE}
ANN	10.75%	0.0626	0.0829
LSSVM	9.77%	0.0606	0.0859
ELM	11.98%	0.1198	0.1469
Ensemble	6.67%	0.0377	0.0523

5.3. IEEE 39 with Inadequate Samples

In the IEEE 39 system, a situation close to reality is simulated in which the number of training samples is inadequate for individual learner training, particularly samples of severe disturbance. Therefore, the performance of the ensemble algorithm can be examined to determine if it can improve prediction accuracy under such conditions. In this case, 200 samples are used to train learners, and 108 samples are used as a test set. The calculation results of modified cross-entropy for three base learners are shown in Figure 10.

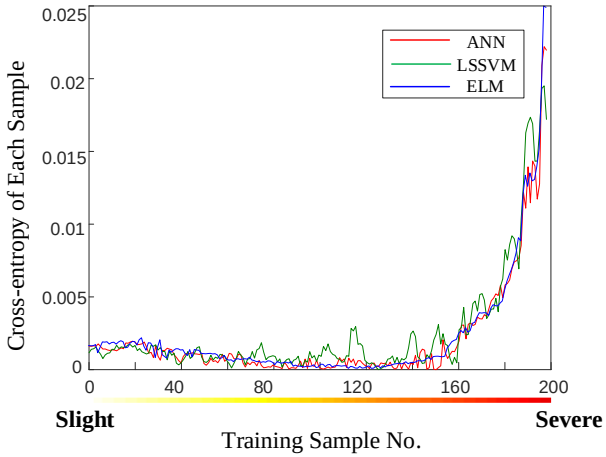

Figure 10. Calculation of modified cross-entropy in the IEEE 39-bus system.

In Figure 10, samples are sorted in ascending order of disturbance severity. Three base learners exhibit a similar trend with respect to cross-entropy, and LSSVM fluctuates slightly more. Consequently, the weights of each base learner are highly approximate in Table 6. A comparison of Figures 8 and 10 reveals the interesting phenomenon that the two systems show considerable difference in the same machine-learning task: lowest frequency prediction. This outcome can be explained by noting that machine learning is a data-sensitive tool and independent of all forms of theory based on physical systems.

Table 6. Weights of base learners in IEEE 39.

Base Learners	ANN	LSSVM	ELM
Weight	0.3139	0.3570	0.3291

Taking relative error in Table 7 as an example, the relative error of the predictions is 17.1–17.82% using individual base learners because the training sample is insufficient and the algorithm is under-fitted. The prediction error of the ensemble algorithm is 13.62%, which indicates better prediction accuracy. In addition, as shown in Figure 11, the distribution of relative error is improved by dragging the boxplot close to the horizontal axis. This improvement indicates that the cross-entropy ensemble learning algorithm has an advantage over single machine-learning algorithms when the number of training samples is inadequate. To further elaborate the relationship between training sample size and prediction accuracy, the number of training samples is adjusted several times. The test results are in Figure 12.

Table 7. Prediction error index in IEEE 39.

Algorithm	e_r	e_{MAE}	e_{RMSE}
ANN	17.10%	0.1003	0.1418
LSSVM	17.82%	0.0985	0.1514
ELM	17.58%	0.1088	0.1752
CE-Ensemble	13.58%	0.0860	0.1411

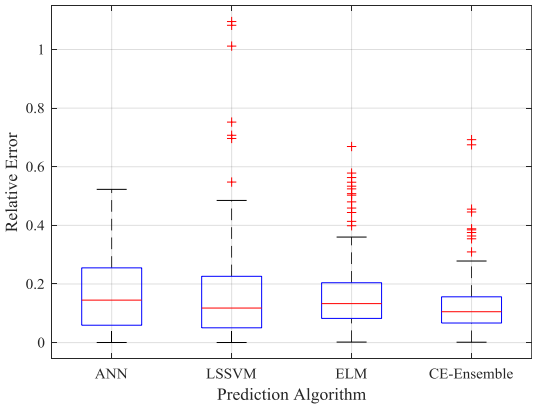

Figure 11. Prediction error among different algorithms.

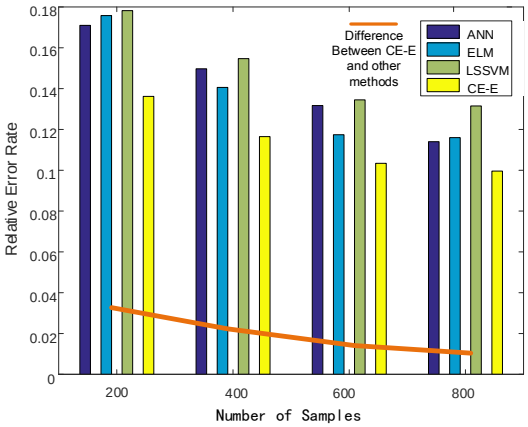

Figure 12. Relation between accuracy and sample size.

It can be surmised from Figure 12 as follows:

(1) With an increase in the sample size, the base learners' prediction error decreases, converging to a limit of 12–14%. This phenomenon reveals that newly added samples provide little information about this test system when the number of samples increases to 800.

(2) The prediction accuracy of the three base learners for different sample sizes varies and LSSVM show the best overall performance. Generally, the performance of machine-learning algorithms may exhibit a fluctuation under different training and test conditions, which is inevitable.

(3) The error difference curve shows that, as the sample size becomes smaller, the ensemble method has a larger advantage over the base learners. The primary reason is that the sample size is too small for the base learners to fit the real quantity relation, while ensemble learning reduces both deviation and variance through the proposed ensemble algorithm.

5.4. Summary

In the WSCC 9 system, a sample-rich system, the prediction accuracy of the proposed algorithm is higher than that of the individual base learners. This outcome verifies the effectiveness of the described ensemble algorithm in normal learning tasks.

More significantly, in the IEEE 39 system, which is a widely acknowledged benchmark power system for security analysis, the sample inadequacy situation is simulated. The outcome demonstrates that the accuracy advantage of the proposed algorithm is more obvious when the number of samples decreases in a certain interval. That is, the proposed algorithm solves the prediction difficulty despite the scarcity of severe disturbance samples.

6. Conclusions

In this paper, the trade-off between accuracy and calculation speed in the power system frequency prediction is resolved by the described ensemble algorithm. More importantly, when the frequency prediction, in the case of inadequate samples, was investigated, it was shown that the proposed algorithm tended to be more accurate under those conditions. Specifically, an ensemble algorithm based on modified cross-entropy was proposed to predict post-disturbance deviations of the power system frequency and to compensate for the errors of a single machine-learning algorithm. An online prediction system using the algorithm was established to perform real-time prediction while simultaneously updating the dataset. For WSCC 9 and IEEE 39 systems, the performance of the proposed method was verified for inadequate samples. By strategically increasing the complexity of the prediction model, the proposed algorithm showed higher accuracy for both systems. Compared to other frequency-prediction methods, the proposed cross-entropy ensemble algorithm was able to make a quick prediction that was relatively accurate with limited training samples. In future research, the performance of the proposed algorithm will be analyzed more comprehensively from the perspective of machine learning theory, and additional applications will be considered.

Acknowledgments: This work is supported by the National Natural Science Foundation of China (Grant No. 51577030) and the National Key Research and Development Program of China (Grant No. 2017YFB0903000).

Author Contributions: Yi Tang, Han Cui and Qi Wang contributed to developing the ideas related to this research. Han Cui performed this research. Yi Tang and Qi Wang proposed the algorithm structure and further optimization. All of the authors read and approved the final manuscript.

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

References

1. Zhou, H.; Su, Y.; Chen, Y.; Ma, Q. The China southern power grid: Solutions to operation risks and planning challenges. *IEEE Power Energy Mag.* **2016**, *14*, 72–78. [CrossRef]
2. Gomez, F.R.; Rajapakse, A.D.; Annakkage, U.D.; Fernando, I.T. Support vector machine-based algorithm for post-fault transient stability status prediction using synchronized measurements. *IEEE Trans. Power Syst.* **2011**, *26*, 1474–1483. [CrossRef]
3. Polajžer, B.; Drago, D.; Jožef, R. Estimation of area's frequency response characteristic during large frequency changes using local correlation. *IEEE Trans. Power Syst.* **2016**, *31*, 3160–3168. [CrossRef]
4. Sun, Y.; Yu, X.; Tan, Z.; Xu, X.; Yan, Q. Efficiency evaluation of operation analysis systems based on dynamic data envelope analysis models from a big data perspective. *Appl. Sci.* **2017**, *7*, 624. [CrossRef]
5. Esfahanian, M.; Erdol, N.; Gerstein, E.; Zhuang, H. Two-stage detection of north atlantic right whale upcalls using local binary patterns and machine learning algorithms. *Appl. Acoust.* **2017**, *120*, 158–166. [CrossRef]
6. Esfahanian, M.; Zhuang, H.; Erdol, N.; Gerstein, E. Comparison of two methods for detection of North Atlantic Right Whale upcalls. In Proceedings of the 2015 23rd European Signal Processing Conference (EUSIPCO), Nice, France, 31 August–4 September 2015; pp. 559–563.
7. Chatzakou, D.; Vakali, A.; Kafetsios, K. Detecting variation of emotions in online activities. *Expert Syst. Appl.* **2017**, *89*, 318–332. [CrossRef]
8. Zhan, S.; Tao, Q.Q.; Li, X.H. Face detection using representation learning. *Neurocomputing* **2016**, *187*, 19–26. [CrossRef]
9. Gao, B.; Wu, C.; Wu, Y.; Tang, Y. Expected Utility and Entropy-Based Decision-Making Model for Large Consumers in the Smart Grid. *Entropy* **2015**, *17*, 6560–6575. [CrossRef]
10. Moradi, H.; Esfahanian, M.; Abtahi, A.; Zilouchian, A. Modeling a Hybrid Microgrid Using Probabilistic Reconfiguration under System Uncertainties. *Energies* **2017**, *10*, 1430. [CrossRef]
11. Angel, A.D.; Geurts, P.; Ernst, D.; Glavic, M.; Wehenkel, L. Estimation of rotor angles of synchronous machines using artificial neural networks and local PMU-based quantities. *Neurocomputing* **2007**, *70*, 2668–2678. [CrossRef]

12. AL-Masri, A.N.; Ab Kadir, M.Z.A.; Hizam, H.; Maruin, N. A novel implementation for generator rotor angle stability prediction using an adaptive artificial neural network application for dynamic security assessment. *IEEE Trans. Power Syst.* **2013**, *28*, 2516–2525. [CrossRef]
13. Malbasa, V.; Zheng, C.; Chen, P.C.; Popovic, T.; Kezunovic, M. Voltage stability prediction using active machine learning. *IEEE Trans. Smart Grid* **2017**. [CrossRef]
14. Diao, R.; Sun, K.; Vittal, V.; O'Keefe, R.J.; Richardson, M.R.; Bhatt, N. Decision tree-based online voltage security assessment using PMU measurements. *IEEE Trans. Power Syst.* **2009**, *24*, 832–839. [CrossRef]
15. Xu, Y.; Dai, Y.; Dong, Z.; Zhang, R.; Meng, K. Extreme learning machine-based predictor for real-time frequency stability assessment of electric power systems. *Neural Comput. Appl.* **2013**, *22*, 501–508. [CrossRef]
16. Zhao, C.; Topcu, U.; Li, N.; Low, S. Design and stability of load-side primary frequency control in power systems. *IEEE Trans. Autom. Control* **2014**, *59*, 1177–1189. [CrossRef]
17. Zhang, Y.; Xu, Y.; Dong, Z.; Xu, Z.; Wong, K.P. Intelligent early-warning of power system dynamic insecurity risk: Towards optimal accuracy-earliness tradeoff. *IEEE Trans. Ind. Inform.* **2017**, *13*, 2544–2554. [CrossRef]
18. Djukanovic, M.B.; Popovic, D.P.; Sobajic, D.J.; Pao, Y.H. Prediction of power system frequency response after generator outages using neural nets. *IEEE Proc. C Gener. Transm. Distrib.* **2002**, *140*, 389–398. [CrossRef]
19. Dong, J.; Ma, X.; Djouadi, S.M.; Li, H. Real-time prediction of power system frequency in FNET: A state space approach. *IEEE Int. Conf. Smart Grid Commun.* **2013**, *143*, 109–114.
20. Rudez, U.; Mihalic, R. Wams-based underfrequency load shedding with short-term frequency prediction. *IEEE Trans. Power Deliv.* **2016**, *31*, 1912–1920. [CrossRef]
21. Zhang, J.; Zheng, X.; Wang, Z.; Guan, L.; Chung, C.Y. Power system sensitivity identification—Inherent system properties and data quality. *IEEE Trans. Power Syst.* **2017**, *32*, 2756–2766. [CrossRef]
22. Guan, X.; Wang, Y.; He, J. A Probabilistic Damage Identification Method for Shear Structure Components Based on Cross-Entropy Optimizations. *Entropy* **2017**, *19*, 27. [CrossRef]
23. Alizadeh, M.; Amraee, T. Adaptive scheme for local prediction of post-contingency power system frequency. *Electr. Power Syst. Res.* **2014**, *107*, 240–249. [CrossRef]
24. Dietterich, T.G. *Ensemble Methods in Machine Learning*; Springer: Berlin, Germany, 2000; pp. 1–15. ISBN 978-3-540-67704-8.
25. Breiman, L. *Machine Learning*, 1st ed.; Kluwer Academic Publishers: Philadelphia, PA, USA, 1996; Volume 24, pp. 49–64.
26. Mao, S.; Jiao, L.; Xiong, L.; Gou, S.; Chen, B.; Yeung, S.K. Weighted classifier ensemble based on quadratic form. *Pattern Recogn.* **2015**, *48*, 1688–1706. [CrossRef]
27. Agatonovic-Kustrin, S.; Beresford, R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J. Pharm. Biomed. Anal.* **2000**, *22*, 717–727. [CrossRef]
28. Kruif, B.J.D.; Vries, T.J.A.D. Pruning error minimization in least squares support vector machines. *IEEE Trans. Neural Netw.* **2003**, *14*, 696–702. [CrossRef] [PubMed]
29. Huang, G.B.; Wang, D.H.; Lan, Y. Extreme learning machines: A survey. *Int. J. Mach. Learn. Cybern.* **2011**, *2*, 107–122. [CrossRef]
30. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme learning machine: Theory and applications. *Neurocomputing* **2006**, *70*, 489–501. [CrossRef]