# Short-Term Load Forecasting Based on Big Data Technologies

Pei Zhang, *Member, CSEE, Senior Member, IEEE*, Xiaoyu Wu, *Student Member, IEEE,* Xiaojun Wang, and Sheng Bi

***Abstract*—With the construction of smart grid, lots of renewable energy resources such as wind and solar are deployed in power system. It might make the power system load varied complex than before which will bring difficulties in short-term load forecasting area. To overcome this issue, this paper proposes a new short-term load forecasting framework based on big data technologies. First, a cluster analysis is performed to classify daily load patterns for individual loads using smart meter data. Next, an association analysis is used to determine critical influential factors. This is followed by the application of a decision tree to establish classification rules. Then, appropriate forecasting models are chosen for different load patterns. Finally, the forecasted total system load is obtained through an aggregation of an individual load's forecasting results. Case studies using real load data show that the proposed new framework can guarantee the accuracy of short-term load forecasting within required limits.**

***Index Terms*—Association analysis, big data, cluster analysis, decision tree, short-term load forecasting.**

## I. Introduction

SHORT-TERM load forecasting, a key parameter that helps electric utility operators make decisions such as purchasing and selling electric power, load switching, and maintenance planning, serves an important function in system operations. Researchers have proposed different approaches to short-term load forecasting, which can be generally classified into two categories: time series analysis methods, including ARIMA models [1]–[3], exponential smoothing [4], [5], and machine learning approaches, such as neural networks (NN) [6]–[8], support vector machine (SVM) [9]–[11], and fuzzy system (FS) [12], [13]. To enhance forecasting accuracy of these two categories of methods, researchers have used multi-wavelet analysis [14], [15] and chaos theory [16], [17] for improved outcomes.

In general, short-term forecasting methods perform direct forecasting of the total system load using historical load data and weather data as inputs. However, since the grid consists of thousands of individual users and many time-varying characteristics, a single forecasting method, such as those mentioned earlier, cannot adequately forecast individual loads, as well as the accompanying factors that influence the variations in these loads. Therefore, current approaches, which treat all users as a single entity, sometimes may not be able to meet accuracy requirements under all circumstances. Another issue is that the load needs to be forecasted at the substation or bus level for calculation of the power flow. Most utilities do not process load forecasting at the substation or bus level because of the complexities involved in capturing the necessary information or because there is very little data available.

The construction of the smart gird has led to wide deployment of smart meters [18]. Smart meters provide an opportunity to analyze the time-varying characteristics of individual loads. In China, smart meters measure electricity usage of industrial loads every 15 min. For commercial and residential loads, smart meters measure electricity usage every hour. The huge amount of data from millions of smart meters has reached the tera byte (TB) or peta byte (PB) scale, offering valuable information that can be used for short-term load forecasting. To handle large quantities of data, however, presents a challenge, and in recent years, a few industries [19]–[22] have utilized big data tools, leading to significant results. Yahoo's R&D team, for example, uses the Hadoop platform for sorting through 1 TB of data in only 62 seconds.

A new short-term load forecasting framework based on big data technologies is proposed in this paper. In Section II, the framework and relevant techniques of the short-term load analysis and forecasting method are presented in detail. Section III introduces a technical framework of the proposed method using big data technologies. Section IV provides case study results. Section V concludes this paper.

## II. Forecasting Framework and Core Techniques

The proposed forecasting framework includes five steps, as shown in Fig. 1. Steps 1 through 3 consist of machine learning techniques that aim to discover typical load patterns of historical load data and then find their critical influential factors to establish classification rules. Step 4 is a model training process, where parameter combinations for corresponding load patterns are chosen to build forecasting models. In Step 5, individual load forecasting results are added to arrive at the final system load.

Manuscript received March 27, 2015; revised June 19, 2015; accepted August 9, 2015. Date of publication September 30, 2015; date of current version August 19, 2015.

P. Zhang is with NARI Accenture Information Technology Center, Beijing 100044, China (e-mail: peizhang166@gmail.com).

X. Y. Wu, X. J. Wang, and S. Bi are with School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: 13117375@bjtu.edu.cn; xjwang1@bjtu.edu.cn; 13121385@bjtu.edu.cn).

Digital Object Identifier 10.17775/CSEEJPES.2015.00036

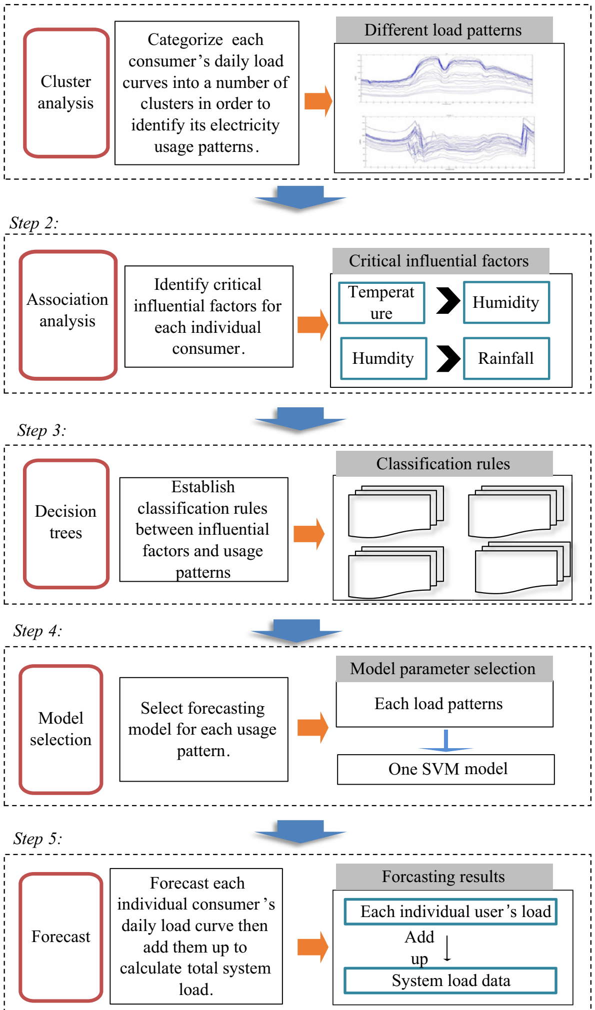



Fig. 1. Framework of proposed approach.

## A. Improved Hierarchical Clustering Technique

Since each individual load can consist of different types of load curves due to factors such as weather, dates, etc., a hierarchical clustering technique is used to classify historical load curves into several patterns.

Using a hierarchical clustering algorithm represents a bottom-up approach, as shown in Fig. 2. First, take each sample as a separate class. Second, calculate the distance between each sample. Third, merge classes that meet the distance requirements.

Finally, repeat the above three steps until the number of classes meet the requirements. Proximity matrix $\boldsymbol{A}$ stores the distances between every two classes [23].

Selection of a proper distance algorithm is the key for the clustering technique. In this paper, a new distance named "Normalized Euclidean Distance" is proposed as follows:

$$d_{12} = \sqrt{\sum_{k=1}^{n} \left(\frac{x_{1k} - x_{2k}}{x_{\max}}\right)^2} \tag{1}$$

where $d_{12}$ represents Normalized Euclidean Distance between Sample 1 and Sample 2, $n$ means $n$ dimensions. $x_{\max}$ means the maximum load value of each dimension. Normalized Euclidean Distance helps adopt a standard accuracy limit to evaluate all the loads with different peak values.

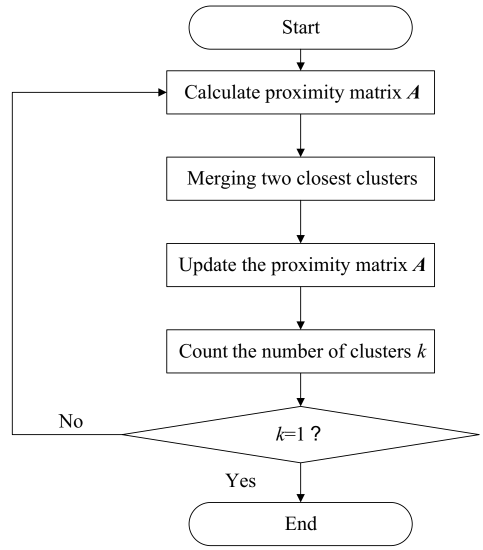



Fig. 2. Flow chart of hierarchical clustering algorithm.

## B. Identification of Critical Influencing Factors

Many factors, including temperature, humidity, day type, etc. have impact on loads. For different loads, critical influential factors are not the same. For example, temperature could be a major influential factor for residential loads. But temperature may have little impact on some temperature non-sensitive industrial loads. Grey correlation analysis is applied to determine the critical influential factors of each individual load.

The degree of correlation is determined by analyzing the correlation between the object series and the reference series according to the principle of grey correlation theory [24], [25]. The reference series is shown as $\boldsymbol{X}_0 = \{x_0(1), x_0(2), \ldots, x_0(k), \ldots, x_0(n)\}$, and the object data series are $\boldsymbol{X}_i = \{x_i(1), x_i(2), \ldots, x_i(k), \ldots, x_i(n)\}$, where $x_0(k)$, $x_i(k)$ are the values of $\boldsymbol{X}_0$ and $\boldsymbol{X}_i$ in the $k$ moment. After the non-dimensional transformation, the correlation coefficient between two series at $k$ moment is shown as follows:

$$\xi_{0i}(k) = \frac{\min\limits_i \min\limits_k |x_0(k) - x_i(k)| + \rho \max\limits_i \max\limits_k |x_0(k) - x_i(k)|}{|x_0(k) - x_i(k)| + \rho \max\limits_i \max\limits_k |x_0(k) - x_i(k)|} \tag{2}$$

where $\xi_{0i}(k)$ is the correlation coefficient between two series at $k$ moment; $\rho$ is the distinguishing coefficient.

The correlation between two series is the average value of all the coefficients, and is given as follows:

$$\gamma_{0i} = \frac{1}{n} \sum_{t=1}^{n} \xi_{0i}(t). \tag{3}$$

Object series are load series, and reference series are several influential factors. The less the two series differ, the more correlations they have. High grey association coefficient implies that the corresponding factor has strong correlation with the load. These factors are critically influential. The results from association analysis provide inputs for the next step, which is the decision tree analysis.

### C. Establishment of Classification Rules

Based on cluster analysis and association analysis results, a decision tree, which is based on CART algorithm, is developed to establish relationships between clustering results and critical factors.

Proposed by Breiman in 1984, CART algorithm [26], makes a decision in each node based on a split of a variable, and then makes its way down until it reaches a leaf node. Another feature of CART is the splitting of each node into 2 sub-nodes by finding a best split variable according to the *Gini* index.

$$Gini(D) = 1 - \sum_{i=1}^{m} {p_i}^2 \tag{4}$$

where $D$ represents the original data set, $p_i$ is the sample classification probability of each cluster $C_i$, and $m$ means this original data set $D$ includes $m$ categories.

The historical data set consists of load influential factors, and the leaf node labels of CART are load patterns determined by hierarchical clustering analysis. CART is then able to make a decision in each node based on a split of an influential factor until it reaches a leaf node, which is determined by cluster analysis. Fig. 3 shows the process of using a decision tree to establish the classification process. Given the critical influential factors of the forecasting day, the cluster, which the forecasting day belongs to, can be determined according to the classification rules in the decision tree, as shown in Table IV at the end of this paper.

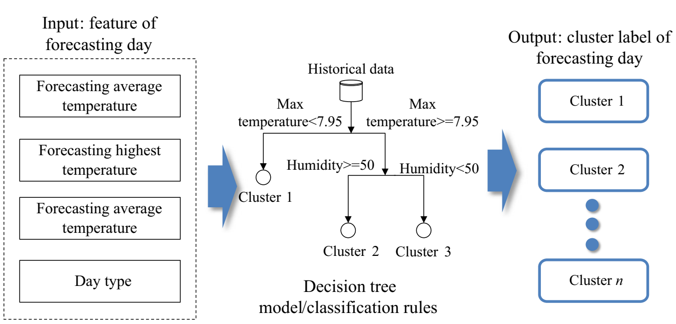



Fig. 3. The process of using a decision tree to establish classification rules.

### D. Selection of Appropriate Forecasting Models

For different load patterns, different support vector machine (SVM) models and parameters are developed to ensure the forecasting accuracy within the required limits.

First proposed by Vapnik [27], SVM is an effective technique for classification and regression problems. SVMs are supervised learning models with associated learning algorithms that analyze data and recognize patterns used for classification and regression analysis. SVM can efficiently perform a nonlinear classification using the kernel trick, implicitly mapping the inputs into high-dimensional feature spaces.

The support vector regression problem is shown below:

$$\min_{\omega,b,\xi,\xi^*} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{n}(\xi_i + {\xi_i}^*) \tag{5}$$

subject to:

$$\begin{aligned} & y_i(\omega^T\phi(x_i) + b) \le \varepsilon + \xi_i, \\ & (\omega^T\phi(x_i) + b)y_i \le \varepsilon + {\xi_i}^*, \\ & \xi_i, {\xi_i}^* \ge 0,\ i = 1, \ldots, n \end{aligned}$$

where $(x_1, y_1), \ldots, (x_n, y_n)$ are a pair of input and output vectors, $n$ is the number of samples, $\omega$ is weight factor, $b$ is the threshold value, and $C$ is error cost.

Input samples are mapped to higher dimensional space by using kernel function $\phi$; $\xi_i$ is the upper training error; and $\xi_i^*$ is the lower training error subject to $\varepsilon$-insensitive tube $|y - (\omega^T\phi(x_i) + b)| \le \varepsilon$. Kernel function is responsible for nonlinear mapping between input and feature space. There are three typical kernel functions: gaussian (RBF), polynomial, and linear.

For each type of load pattern, a corresponding data set is chosen to train the SVM model in order to determine the values of $\varepsilon$, $C$, and kernel parameter $\delta^2$. Therefore, a set of SVM models is developed for each load. Once the load pattern is determined according to the decision tree, the corresponding SVM forecasting model, with appropriate parameters, is used to ensure the forecasting accuracy.

### E. Forecasting Total Load

The total system load is forecasted based on aggregation of an individual load's forecasting results, as shown in Fig. 4. After the forecasting result of each user's load is obtained, the forecasted total load $L_{\text{total}}$ can be calculated by adding all the forecasted individual loads, together with line loss $L_{\text{loss}}$.

$$L_{\text{total}} = L_{\text{loss}} + \sum_{i=1}^{n} l_{\text{user}(i)}. \tag{6}$$

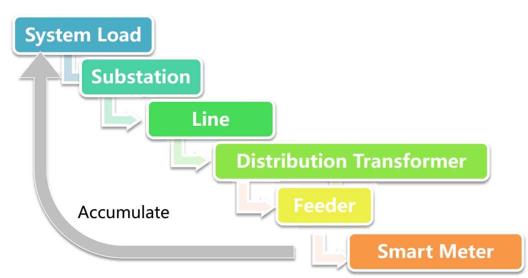



Fig. 4. System topology.

## III. Technical Architecture Based on Big Data Technologies

Traditional forecasting approaches forecast the total system load directly. The proposed approach in this paper forecasts each individual load. For example, a target system consists of data from 1 million smart meters with critical weather data reaching approximately 15 TB per year. The Hadoop platform is applied to handle the vast amounts of smart meter and weather data. Fig. 5 illustrates the technical architecture using big data technologies.

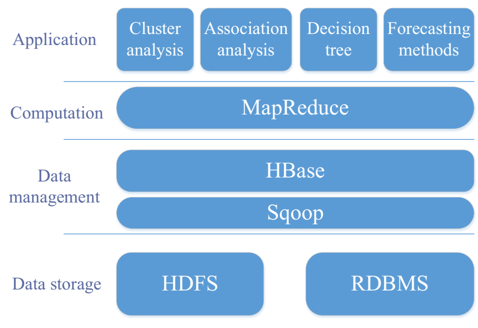



Fig. 5. Big data processing platform.

The technical architecture in Fig. 5 contains four layers: data storage, data management, computation, and application.

### A. *Data Storage Layer*

The data storage layer applies two technologies: Hadoop distributed file system (HDFS) and relationship database management system (RDBMS). HDFS is applied to store smart meter data and weather data. RDBMS is used to store consumer data and system topology.

### B. *Data Management Layer*

The data management layer applies Sqoop technology to convert data from weather, smart meter, and consumer usage into a Hadoop database (HBase). In HBase, data from each consumer's load is stored in a single table. HBase applies a key/value technique, as shown in Fig. 6 for quick and efficient access to target data. The row key includes user ID and date. The column key is time. The value consists of measured electricity usage, temperature, humidity, wind velocity, and day type.

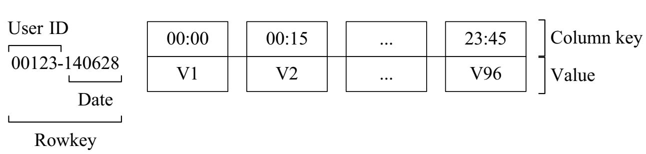



Fig. 6. Key/value in HBase.

### C. *Computation Layer*

MapReduce in Hadoop processes parallelizable problems across huge datasets by using a large number of computers (nodes). The process includes three steps: map, shuffle, and reduce, as shown in Fig. 7.

For data importing processes, corresponding steps are shown below:

*1) Map:* Historical data sets, including load data and weather data are stored in HDFS in plain text style. Then, a map function is carried out by each worker node (computer server), which reads and analyzes the data and forms them into a key/value.

*2) Shuffle:* Worker nodes (computer servers) redistribute data based on the output keys (produced by the "map()" function), such that all data belonging to one key is located on the same worker node. In this process, load data of different collection times with corresponding users will be stored in the same group.

*3) Reduce:* Worker nodes (computers or servers) now process each group of output data per key, and in parallel. This step aims at merging similar load data and eliminating data error.

For forecasting processes, only the Map step is applied. The logical sequence of the forecasting task has already been presented in Section II. For each user, these aforementioned four steps of clustering, association analysis, decision tree, and SVM forecasting are carried out in one map. Forecasting results for each load is then added up in an extra map threading.

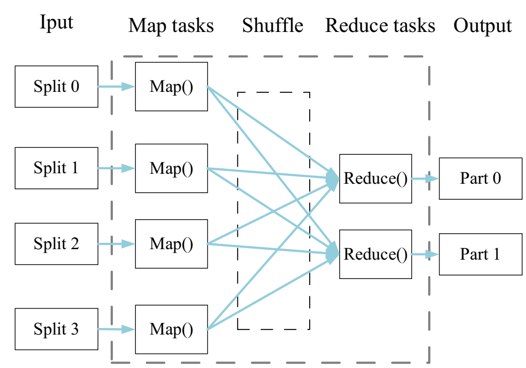



Fig. 7. MapReduce working framework.

### D. *Application Layer*

The application layer includes all the technologies described in Section II, i.e., cluster analysis, association analysis, decision tree, and forecasting methods.

## IV. Case Study

### A. *Description of Data Set*

The proposed short-term load forecasting framework is verified using smart meter data collected from a real distribution system consisting of 1.2 million consumers. The user types are industrial, commercial, residential, and municipal, as shown in Table I. The total amount of historical data is approximately 1.5 TB. The historical load data was captured at every hour from Jan. 1, 2012 to Dec. 31, 2012, and is used as training data. The load data and corresponding weather data from 25/4/2013 to 1/5/2013 are used for testing data.

### B. *Identification of Load Patterns Using Hierarchical Clustering Algorithm*

Fig. 8 shows the daily load curves of customer 981, an industrial client from 1/1/2012 to 31/12/2012. The improved hierarchical clustering algorithm with normalized Euclidean distance is applied. Six clusters are identified, as shown in Fig. 9(a)–(g).

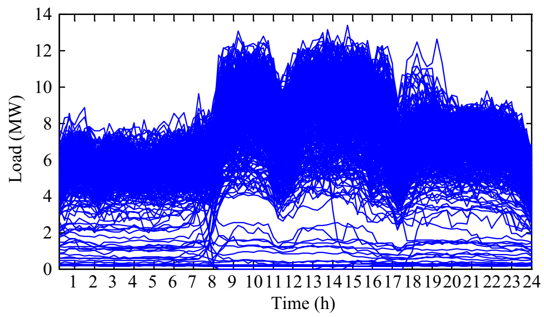



Fig. 8. Daily load curve of customer 981 from 1/1/2012 to 31/12/2012.

TABLE I
TYPES OF CUSTOMERS IN TARGET DISTRIBUTION SYSTEM

| Types | Amount of Smart Meters | Types | Amount of Smart Meters |
|---|---|---|---|
| Industrial | 231,810 | Residential | 538,458 |
| Commercial | 330,271 | Municipal | 117,613 |

Table II summarizes the number of days within each cluster. Cluster 2 ($C_2$) contains most weekends. $C_3$ and $C_4$ contain most weekdays. Cluster 5 ($C_5$) is mainly composed of the first day after weekends.

A further analysis indicates that the sixth cluster ($C_6$) includes all holidays: New Year (30/12/2012, 1/1/2012), Spring Festival (21/1/2012–28/1/2012), Qingming Festival (3/4/2012–4/4/2012), Labor Day (1/5/2012, 2/5/2012), and Chinese National Holiday (30/9/2012–7/10/2012). The first cluster con-

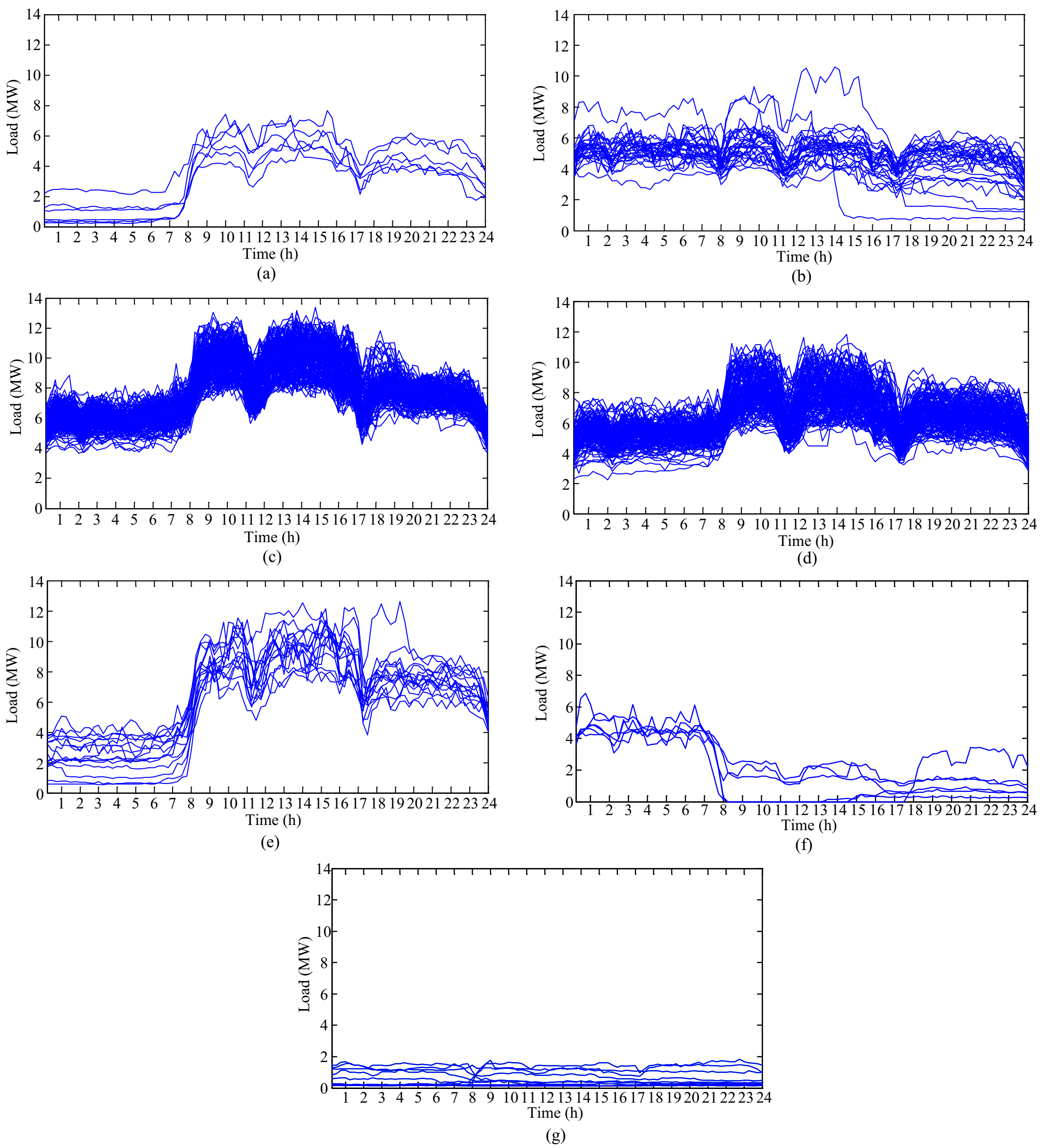



Fig. 9. Cluster results of customer 981 from 1/1/2012 to 31/12/2012. (a) Cluster 1. (b) Cluster 2. (c) Cluster 3. (d) Cluster 4. (e) Cluster 5. (f) Cluster 6. (g) Cluster 7.

TABLE II
NUMBER OF DAYS ASSOCIATED WITH SIX CLUSTERS

| Day Type | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ |
|---|---|---|---|---|---|---|---|
| Monday | 4 | 0 | 11 | 27 | 8 | 3 | 0 |
| Tuesday | 0 | 0 | 22 | 26 | 1 | 2 | 1 |
| Wednesday | 2 | 1 | 27 | 18 | 2 | 1 | 1 |
| Thursday | 2 | 1 | 26 | 19 | 2 | 2 | 0 |
| Friday | 0 | 1 | 32 | 18 | 0 | 1 | 0 |
| Saturday | 2 | 24 | 6 | 8 | 9 | 1 | 2 |
| Sunday | 2 | 22 | 12 | 9 | 0 | 6 | 2 |

tains most of first days following a holiday.

### C. Association Analysis to Identify Influential Factors

The grey association analysis method described in Section II is applied to determine influential factors. Table III shows association coefficients corresponding to each factor. The maximum temperature, average humidity, average temperature, and day type are more relevant than wind velocity and average precipitation. These factors are used as inputs for decision tree analysis in Section IV D and the SVM forecasting model in Section IV E.

### D. Establishment of Decision Tree

Once the forecasting temperature, the forecasting humidity and the day type are known, the cluster of the forecasted load can be determined using the decision tree. Fig. 10 shows the decision tree built using CART algorithm, which is a sketch map (the entire CART diagram can be seen at the end of paper).

For example, on April 26, 2013, which was Saturday, the average temperature was 12.75°C, the maximum temperature was 21°C, and the average humidity was 54.3%. According to the decision tree, which is attached at the end of this paper, it belongs to Cluster 4. Table IV shows the corresponding cluster for each day to be forecasted based on the decision tree analysis. Once the cluster of the forecasting day is determined, a proper forecasting model can then be applied.

TABLE III
GREY CORRELATION COEFFICIENT OF CUSTOMER 981

| MT | AT | AH | DT | WV | AP |
|---|---|---|---|---|---|
| 0.7512 | 0.53459 | 0.6588 | 0.4472 | 0.2214 | 0.2137 |

MT = maximum temperature, AT = average temperature, AH = average humidity, DT = day type, WV = wind velocity, AP = average precipitation.

TABLE IV
CLASSIFICATION RESULTS FROM 25/4/2013 TO 1/5/2013

| Day Number | Results | Day Number | Results |
|---|---|---|---|
| 25/4/2013 | Cluster 3 | 29/4/2013 | Cluster 5 |
| 26/4/2013 | Cluster 4 | 30/4/2013 | Cluster 1 |
| 27/4/2013 | Cluster 2 | 1/5/2013 | Cluster 7 |
| 28/4/2013 | Cluster 1 | | |

### E. Application of Appropriate Forecasting Model

Using the training data, SVM parameters including kernel parameter $\delta^2$, error cost $c$ and insensitive value $\varepsilon$ are determined for the six types of clusters, as shown in Table V. The traditional parameter combination is shown in Table VI.

TABLE V
SVM PARAMETER COMBINATION OF EACH CLUSTER

| Cluster Label | $\varepsilon$ | $\delta^2$ | $c$ |
|---|---|---|---|
| Cluster 1 | 0.0010 | 20 | 110 |
| Cluster 2 | 0.0015 | 37 | 601 |
| Cluster 3 | 0.0071 | 87 | 115 |
| Cluster 4 | 0.0032 | 20 | 341 |
| Cluster 5 | 0.0016 | 28 | 200 |
| Cluster 6 | 0.0092 | 32 | 123 |
| Cluster 7 | 0.0024 | 25 | 111 |

TABLE VI
SVM PARAMETER COMBINATION OF TRADITIONAL APPROACH

| Parameter | $\varepsilon$ | $\delta^2$ | $c$ |
|---|---|---|---|
| Value | 0.002 | 20 | 100 |

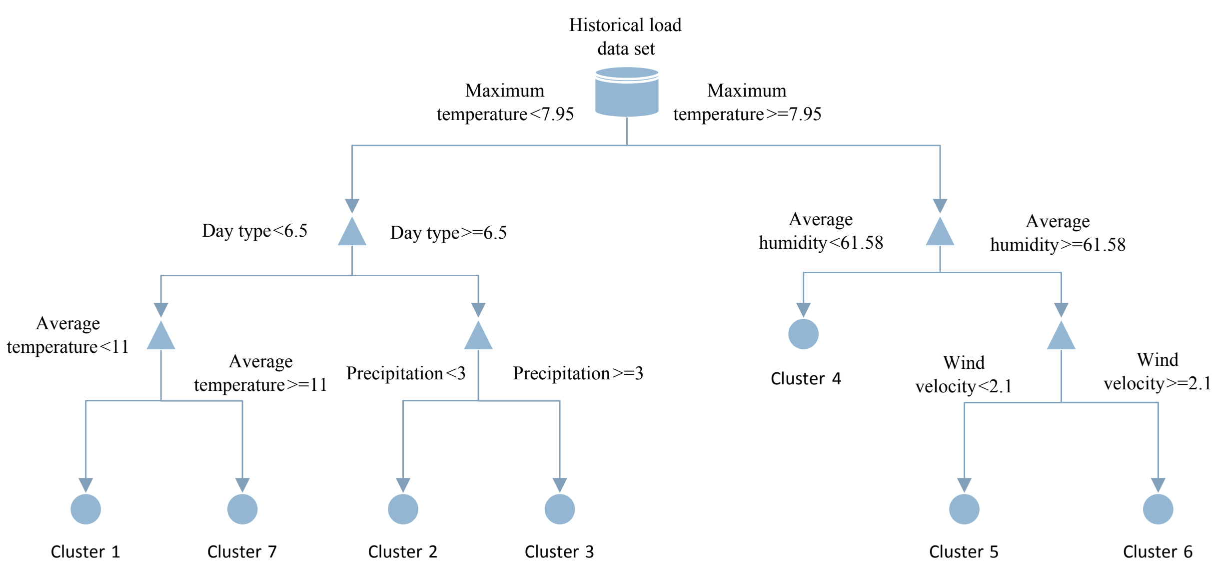



Fig. 10. Establishment of decision tree using CART algorithm (sketch map).

The forecasting results of customer 981 on 29/4/2013 are shown in Fig. 11. The proposed approach totally meets 3% accuracy requirements set by the utility company. The new approach reflects better performance than traditional SVM and ANN approaches.

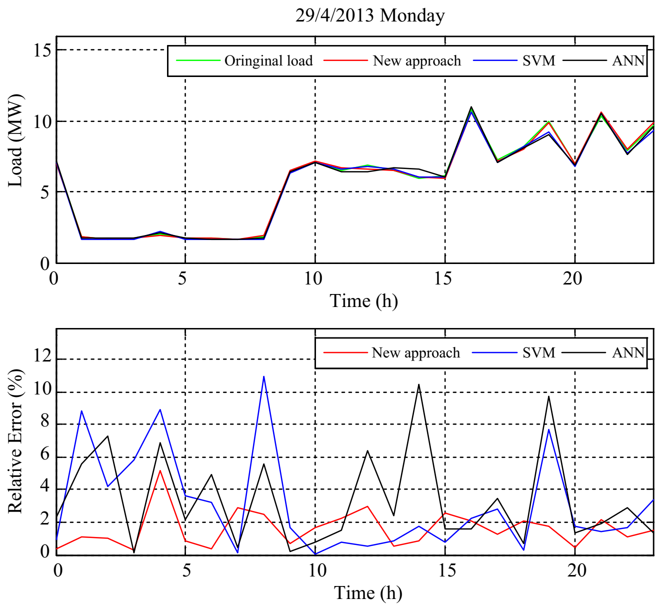



Fig. 11. Forecasting results of customer 981, 29/4/2013.

## F. Forecasting the Total Load

The short-term forecasting framework in this paper was implemented on Hadoop platform using 5 PC servers. Each server has two E5-2403 CPUs and 8 GB memory.

Table VII shows the computation time, which meets online forecasting requirements. The computation time can be reduced as the number of servers increases. Table VIII shows the computation time of traditional ANN and SVM approaches. Owing to these two methods of forecasting system load directly, they are implemented on a single PC server with two E5-2403 CPUs and 8 GB memory. Fig. 12 shows the forecasting results of the total system load.

The forecasting error using a traditional approach, which forecasts the total system load directly without analyzing each consumer's electricity usage pattern, may exceed 3% sometimes. The forecasting error using the proposed forecasting framework here is within 1% of real system load. This clearly indicates that the proposed approach produces improved accuracy than the traditional approach.

TABLE VII
TASKS AND COMPUTATION TIME OF PROPOSED APPROACH

| Tasks | Size | Computation Time |
|---|---|---|
| Cluster (offline) | 1.2 million consumers (3 years) | 24 min |
| Forecasting (online) | 1.2 million consumers (daily) | 110 s |

Table IX shows relative error of consumer 981, and the total system load. It also shows relative error of the traditional SVM method (SVM in Fig. 11, Fig. 12, and Table IX), the proposed method (new approach) and artificial neural network method (ANN). This clearly shows that regardless of consumer or system levels, our forecasting framework produces improved accuracy over traditional methods.

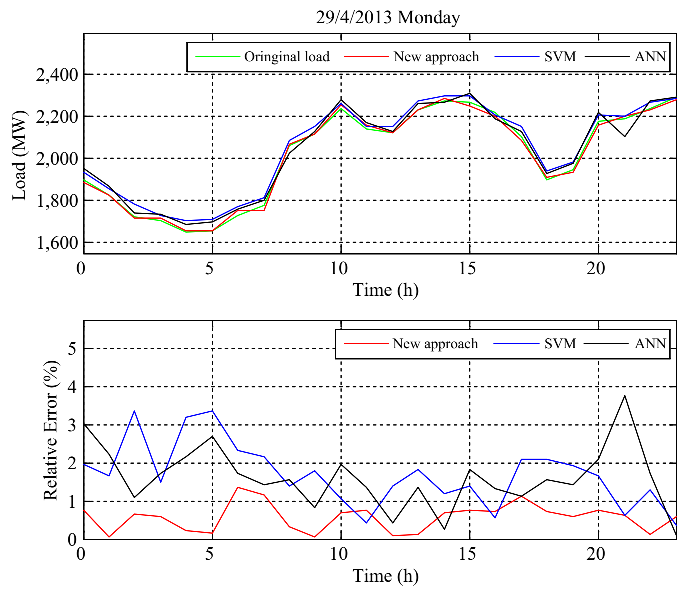



Fig. 12. Forecasting results of system load, 29/4/2013.

TABLE VIII
COMPUTATION TIME OF SVM AND ANN

| Approaches | Computation Time |
|---|---|
| SVM | 8.03 s |
| ANN | 10.12 s |

TABLE IX
RELATIVE ERROR OF CONSUMER 981 AND SYSTEM LOAD ON 29/4/2013

| Time (h) | Relative Error of Customer 981 (%) | | | Relative Error of System Load (%) | | |
|---|---|---|---|---|---|---|
| | New | SVM | ANN | New | SVM | ANN |
| 0 | 0.39 | 0.92 | 2.35 | 0.74 | 1.95 | 3.01 |
| 1 | 1.11 | 8.86 | 5.58 | 0.07 | 1.66 | 2.21 |
| 2 | 1.05 | 4.16 | 7.29 | 0.64 | 3.36 | 1.09 |
| 3 | 0.29 | 5.79 | 0.14 | 0.58 | 1.50 | 1.73 |
| 4 | 5.20 | 8.92 | 6.93 | 0.23 | 3.19 | 2.16 |
| 5 | 0.82 | 3.65 | 2.13 | 0.16 | 3.35 | 2.68 |
| 6 | 0.33 | 3.20 | 4.96 | 1.35 | 2.30 | 1.73 |
| 7 | 2.91 | 0.12 | 0.48 | 1.16 | 2.14 | 1.41 |
| 8 | 2.45 | 10.94 | 5.56 | 0.31 | 1.37 | 1.54 |
| 9 | 0.72 | 1.67 | 0.20 | 0.06 | 1.77 | 0.83 |
| 10 | 1.64 | 0.07 | 0.78 | 0.68 | 1.05 | 1.94 |
| 11 | 2.28 | 0.76 | 1.53 | 0.74 | 0.41 | 1.34 |
| 12 | 2.95 | 0.53 | 6.37 | 0.09 | 1.37 | 0.43 |
| 13 | 0.51 | 0.85 | 2.39 | 0.11 | 1.80 | 1.35 |
| 14 | 0.85 | 1.73 | 10.44 | 0.67 | 1.19 | 0.26 |
| 15 | 2.58 | 0.74 | 1.56 | 0.75 | 1.37 | 1.82 |
| 16 | 2.11 | 2.23 | 1.56 | 0.72 | 0.56 | 1.33 |
| 17 | 1.29 | 2.82 | 3.50 | 1.13 | 2.07 | 1.12 |
| 18 | 2.07 | 0.30 | 0.66 | 0.71 | 2.07 | 1.54 |
| 19 | 1.73 | 7.71 | 9.71 | 0.59 | 1.92 | 1.41 |
| 20 | 0.48 | 1.79 | 1.36 | 0.75 | 1.64 | 2.10 |
| 21 | 2.12 | 1.41 | 1.89 | 0.63 | 0.60 | 3.73 |
| 22 | 1.11 | 1.63 | 2.91 | 0.12 | 1.27 | 1.72 |
| 23 | 1.52 | 3.41 | 1.34 | 0.59 | 0.34 | 0.08 |
| Average | 1.61 | 3.09 | 3.40 | 0.57 | 1.68 | 1.61 |

## V. Conclusion

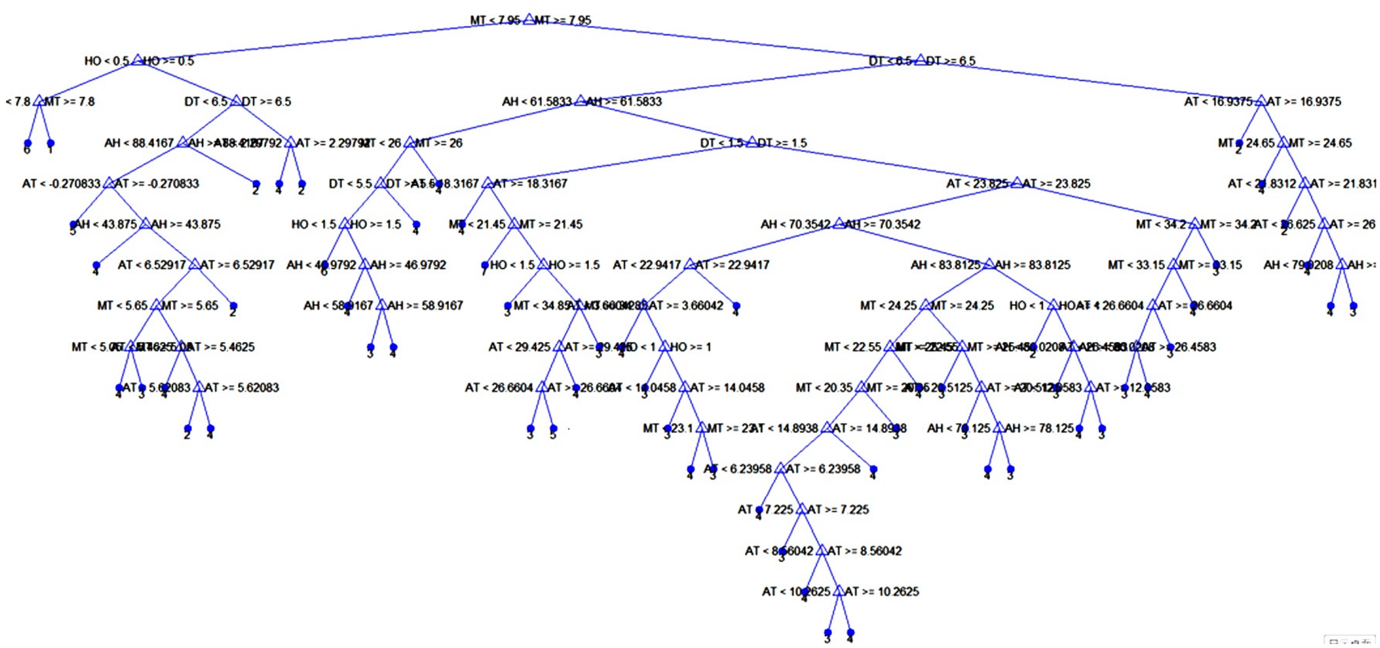
This paper proposes a new framework for performing short-term load forecasting using smart meter data based on big data technologies. Our proposed solution includes five steps:

1) Load identification of load patterns using clustering techniques,
2) determination of critical influencing factors using association analysis,
3) establishment of classification criteria using decision tree,
4) choosing appropriate load forecasting model for each load pattern and associated critical factors,
5) forecasting each individual load and calculating the forecasted system load by adding individual load's forecasting results and line losses.

These forecasting processes are implemented on our proposed processing architecture based on big data technology. In distribution EMS real-time application, once the date and forecasting weather data are available, the decision tree can identify which cluster the individual load belongs to. Then the appropriate forecasting model is selected to forecast individual load. Finally, the system load is forecasted. Case studies using a practical distribution system smart meter data indicate that the proposed forecasting method not only guarantees prediction accuracy within the predefined range, but also processes the large quantity of data within real-time operation requirements.

## Appendix

### Complete CART Decision Tree of Fig. 10




## References

[1] S.-J., Huang and K.-R. Shih, "Short-term load forecasting via ARMA model identification including non-Guassian process considerations," *IEEE Transactions on Power Systems*, vol. 18, no. 2, pp. 673–679, May 2003.

[2] H.-T. Yang, C.-M. Huang, and C.-L. Huang, "Identification of ARMA model for short term load forecasting: an evolutionary programming approach," *IEEE Transactions on Power Systems*, vol. 11, no. 7, pp. 403–408, Aug. 1996.

[3] K. Moshkbar-Bakhshayesh and M. B. Ghofrani, "Development of a robust identifier for NPPs transients combining ARIMA model and EBP algorithm," *IEEE Transactions on Nuclear Science*, vol. 61, no. 4, pp. 2383–2391, Jul. 2014.

[4] P. R. Ji, D. Xiong, P. Wang, and J. Chen, "A study on exponential smoothing model for load forecasting," in *Proceedings of 2012 Power and Energy Engineering Conference (APPEEC),* Shanghai, China, 2012, pp. 1– 4.

[5] R. Gob, K. Lurz, and A. Pievatolo, "Electrical load forecasting by exponential smoothing with covariates," *Applied Stochastic Models in Business and Industry*, vol. 29, no. 6, pp. 629–645, Dec. 2013.

[6] M. R. Cogollo and J. D. Velasquez, "Methodological advances in artificial neural networks for time series forecasting," *Latin America Transactions, IEEE (Revista IEEE America Latina)*, vol. 12, no. 4, pp. 764–771, Jun. 2014.

[7] H. S. Hippert, C. E. Pefreira, and R. C. Souza, "Neural network for short-term load forecasting: a review and evaluation," *IEEE Transactions on Power Systems*, vol. 16, no. 1, pp. 44–55, Feb. 2001.

[8] C.-I. Kim, I.-K. Yu, and Y. H. Song, "Kohonen neural network and wavelet transform based approach to short-term load forecasting," *Electric Power System Research*, vol. 63, no. 3, pp. 169–176, May 2002.

[9] L. Zhu, Q. H. Wu, M. S. Li, L. Jiang, and J. S. Smith, "Support vector regression-based short-term wind power prediction with false neighbours filtered," in *2013 International Conference on Renewable Energy Research and Applications (ICRERA)*, Madrid, Spain, 2013, pp. 740–744.

[10] B.-J. Chen, M.-W. Chang, and C.-J. Lin, "Load forecasting using support vector machines: A study on EUNITE competition 2001," *IEEE Transactions on Power Systems*, vol. 19, no. 4, pp. 1821–1830, Nov. 2004.

[11] A. Selakov, D. Cvijetinovic, L. Milovic, "Hybrid PSO-SVM method for short-term load forecasting during periods with significant temperature variations in city of Burbank," *Applied Soft Computing*, vol. 16, no. 3,

pp. 80–88, Mar. 2013.

[12] R.-H. Liang and C.-C. Cheng, "Combined regression-fuzzy approach for short-term load forecasting," *IEE Proceedings of Generation, Transmission and Distribution,* vol. 147, no. 4, pp. 261–266, Jul. 2000.

[13] P. A. Mastorocostas, J. B. Theocharis, and A. G. Bakirtzis, "Fuzzy modeling for short term load forecasting using the orthogonal least squares method," *IEEE Transactions on Power Systems*, vol. 14, no. 1, pp. 29–36, Feb. 1999.

[14] Y. G. Zhao, P. B. Luh, C. Bomgardner, and G. H. Beerel, "Short-term load forecasting: multi-level wavelet neural networks with holiday corrections," in *Proceedings of 2009 General Meeting of the IEEE Power and Energy Society*, Calgary, Canada, 2009, pp. 1–7.

[15] A. Deihimi, O. Orang, and H. Showkati, "Short-term electric load and temperature forecasting using wavelet echo state networks with neural reconstruction," *Energy*, vol. 57, no. 4, pp. 382–401, Aug. 2013.

[16] M. Hiroyuki and U. Shouichi, "Short term load forecasting with chaos time series analysis," in *Proceedings of 1996 International Conference on Intelligent Systems Application to Power Systems*, Orlando, FL, 1996, pp. 133–137.

[17] L. Li and C. X. Liu. "Application of chaos and neural network in power load forecasting," *Discrete Dynamics in Nature and Society*, vol. 57, pp. 82–89, Aug. 2011.

[18] Y. Simmhan, S. Aman, and A. Kumbhare, "Cloud-based software platform for big data analytics in smart grids," *Computing in Science & Engineering*, vol. 15, no. 4, pp. 38–47, Aug. 2013.

[19] M. Peter, S. Jan, and Z. Iveta, "Concept definition for big data architecture in the education system," in *Proceedings of IEEE 12th International Symposium Conference*, Herl'any, Slovakia, 2014, pp. 331–334.

[20] H. W. Park, Y. Y. Lee, J. R. Jiang, "Study on big data center traffic management based on the zeparation of large-scale data stream," in *Proceedings of Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*, Taichung, China, 2013, pp. 591–594.

[21] T. Armes and M. Refern, "Using big data and predictive machine learning in aerospace test environments," in *Proceedings of 2013 IEEE AUTOTESTCON Conference,* Schaumburg, IL, 2013, pp. 1–5.

[22] M. Herland, T. M. Khoshgoftaar, and R. Wald, "Survey of clinical data mining applications on big data in health informatics," in *Proceedings of 12th International Conference on Machine Learning and Applications (ICMLA)*, Miami, FL, 2013, pp. 465–472.

[23] M. Arguelles, C. Benavides, and I. Fernandez, "A new approach to the identification of regional clusters: hierarchical clustering on principal components," *Applied Economics,* vol. 46, no. 21, pp. 2511–2519, Feb. 2014.

[24] J. L. Deng, "Essentials of grey resources theory (GRT)," *Journal of Grey System,* vol. 19, no. 1, pp. 48–48, Aug. 2007.

[25] J. L. Deng, "Grey modeling in energy-gathering space," *Journal of Grey System,* vol. 19, no. 4, pp. 301–308, Dec. 2007.

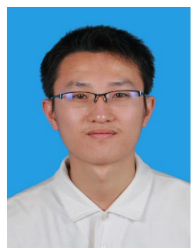
[26] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, vol. 3. San Diego, CA: Chapman & Hall/CRC, CRC Press, 1984, pp. 1–358.

[27] C.-H. Wua, G.-H. Tzeng, R.-H. Lin, "A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression," *Expert Systems with Applications,* vol. 36, no. 3, pp. 4725–4735, Jul. 2009.

**Pei Zhang** (M'00, SM'05) is the CTO of NARI Accenture Information Technology Center. He served as the Technical Director, leading smart grid activities in Accenture, and a Program Manager responsible for grid operations and planning at Electric Power Research Institute (EPRI). Before joining EPRI, he worked at National Grid Company in United Kingdom (UK). He received his Ph.D. degree from Imperial College of Science, Technology and Medicine, University of London, U.K. His current research interests include power system stability and control, reliability and security assessment, and application of advanced IT technologies to power systems.

**Xiaoyu Wu** (S'15) received his B.S. degree from Beijing Jiaotong University, Beijing, China, in 2013. He is currently pursuing his Ph.D. degree at Beijing Jiaotong University, Beijing, China. His research interests include power system stability and control, power system load forecasting, power system stability and application of advanced IT technologies to power system.

**Xiaojun Wang** received his B.S., M.S., and Ph.D. degrees from North China Electric Power University, Beijing, China. He is currently on the faculty at Beijing Jiaotong University. His research interests include power system stability and control, power quality problem analysis, and vehicle-to-grid technology.



**Sheng Bi** received his B.S. degree from South China University of Technology, Guangzhou, China, in 2013. He is currently pursuing a master's degree at Beijing Jiaotong University, Beijing, China. His research interests include power system load forecasting, and application of advanced IT technologies to power systems.