Computing and Informatics, Vol. 36, 2017, 470–492, doi: 10.4149/cai_2017_2_470

A COMPREHENSIVE LEARNING-BASED MODEL FOR POWER LOAD FORECASTING IN SMART GRID

Huifang Li, Yidong Li, Hairong Dong

School of Computer and Information Technology
State Key Laboratory of Rail Traffic Control and Safety
Beijing Jiaotong University
100044 Beijing, China
&
Guangdong Key Laboratory of Popular High Performance Computers
Shenzhen Key Laboratory of Service Computing and Applications
Shenzhen University
e-mail: {14120398, ydli, hrdong}@bjtu.edu.cn

Abstract. In the big data era, learning-based techniques have attracted more and more attention in many industry areas such as smart grid, intelligent transportation. The power load forecasting is one of the most critical issues in data analysis of smart grid. However, learning-based methods have not been widely used due to the poor data quality and computational capacity. In this paper, we propose a comprehensive learning-based model to forecast heavy and over load (HOL) accidents according to the data from various information systems. At first, we present a combined random under- and over-sampling technique for imbalanced electric data, and choose an optimal sampling rate through several experiments. Then, we reduce the attributes that have significant impact on the power load by using learning-based methods. Finally, we provide an algorithm based on the random forest method to prevent the over-fitting problem. We evaluate the proposed model and algorithms with the real-world data provided by China Grid. The experimental results show that our model works efficiently and achieves low error rates.

Keywords: Data mining, power load, random sampling, random forest, smart grid

1 INTRODUCTION

In recent years, as the development of information systems and intelligent electronic devices in power grid companies increases [1, 2, 35], the data-oriented applications have attracted more and more attention in both industry and academic fields. On the one hand, we can take advantage of the big volume and velocity data for knowledge understanding. On the other hand, the employed intelligent and adaptive elements in smart grid require more advanced techniques for big data analysis scenarios, such as power load forecasting.

With the development of economics, the heavy and over load (HOL) accidents are increasing recently, which not only gives rise to the inconvenience of our lives, but brings a great deal of economy loss. Existing studies [6] mainly focused on applying statistical methods to analyse the data derived from production systems only. However, the power load can be affected by many factors from external sources, such as weather, human behaviour and economic pattern. Therefore, it is necessary to develop novel models and methods to achieve better performance in load forecasting with the integrated datasets from various sources.

This paper proposes a comprehensive learning-based model to forecast the HOL accidents based on the multi-source data. We obtain the real power load data from the production system of China grid. The dataset contains three-year records of load and the information of power transformers, connecting lines and customers in Shandong Province. By deeply exploring the real-world data, we found that there are three main issues that need to be solved:

1. how to practically improve a poor quality of the original data;
2. how to effectively observe the related factors with power load from the multi-source data; and
3. how to accurately and efficiently predict the HOL based on the massive data.

In order to deal with the problems mentioned above, we first presented a combined random under- and over-sampling technique for imbalanced electric data, and chose an optimal sampling rate according to experimental results. Then we provided an algorithm for associated feature analysis to extract strong related features with HOL by using association mining methods. The results can either benefit the customers from understanding their electricity consumption patterns and adjusting their electricity consumption strategies more economically, or help the company make high-quality decisions and adopt effective measures to prevent HOL accidents. In addition, it plays an important role in dimension reduction for the following learning task. Finally, we provided an algorithm based on the random forest method to classify the HOL patterns, meanwhile prevent from the over-fitting problem.

The remainder of this paper is organized as follows. Section 2 reviews the recent studies related to the topic. Through analyzing and summarizing the existing research about load classification and the characteristic of HOL, a comprehensive model for power load analysis in smart grid is provided in Section 3. Considering that the HOL cases are minority in load data, a particular sampling method is proposed in Section 4. Section 5 presents useful feature analysis methods for decision making and multi-variables reduction. The comparison of conventional classification algorithms and a specific performance evaluation metric of HOL models are presented in Section 6. Section 7 validates the performance of the model and methods proposed in this paper with extensive experiments. The paper is concluded in Section 8.

2 RELATED WORK

There are different criteria to classify the power load forecasting models applied to the large areas. In terms of forecasting interval, it can be identified as short-term (a day/week ahead), medium-term (a day/week to a year ahead) [4] and long-term (more than a year) [5]. While, with respect to forecasting outputs, it can be categorised as point forecasts, density forecasts and nominal load forecasting.

Various models and methods have been proposed for electric load forecasting [6], and most papers focus on short-term load forecasting since it is an important tool in the day-to-day operation of utility systems [7]. The enriched short-term load forecasting methods can be classified into several categories. First, there are classical statistical models on time series which include stochastic process models [14], exponential smoothing [13], ARMA [9], ARIMA [10, 12] and regression models [15]. In order to solve nonlinearity of electricity demand series artificial neural networks, fuzzy logic and some hybrid approaches have also received substantial attention in load forecasting. The work in [16, 17, 18, 19, 20] proposes artificial neural networks (ANN) models. ANN models performs well, since ANNs can learn the load series and model an unspecified nonlinear relationship between load and weather attributes. Fuzzy logic [21, 22, 23] methods are often good at drawing similarities from huge data. Particle swarm optimization is used in combination with fuzzy neural networks [24]. However, nowadays, machine learning algorithms have been used for power load forecasting and achieved relatively good performance. In general, learning-based methods are often with better self-learning and knowledge detection abilities.

From an overall investigation, we noticed that most papers in power load forecasting field focus on power load regression. There is a minimal research on power load classification and forecasting, especially on HOL forecasting. HOL forecasting is a new and promising research direction. As HOL accidents increased in the recent years it will bring big economic losses and inconvenience to everyone. It is critical issue to provide a high-accurate forecasting model.

3 A COMPREHENSIVE MODEL FOR HEAVY AND OVER LOAD ANALYSIS

In smart grids, the power load data is 'Big', which implies the data is real-time and dynamic, and the type is complex and heterogeneous. These characteristics make it rather difficult for power load forecasting. The hot learning-based method have not been widely used due to poor data quality and computational capacity. In this paper, we provide a comprehensive learning-based model for big power data forecasting. In [3], it proposes a five-phase workflow for power load classification, but the main attention is paid to the clustering model implementation phrase. In this paper, we constructed a comprehensive model with seven stages, including data, data quality assessment, data processing, feature analysis, classification model building, forecasting and result visualization as shown in Figure 1.

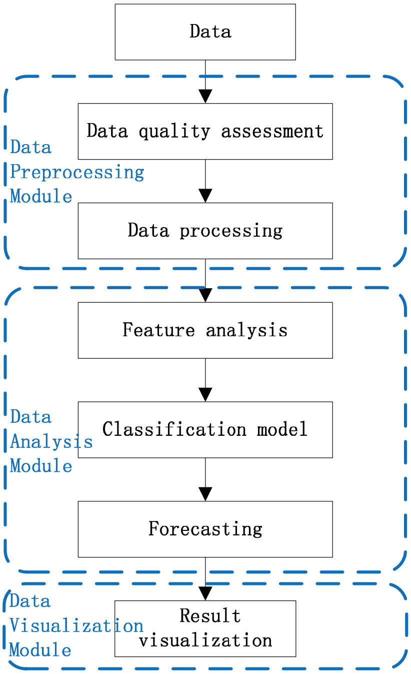


Figure 1. A comprehensive model for heavy and over load forcasting

Our seven-stage workflow can be categorised into three submodules, such as data preprocessing, data analysis and results visualization. For data preprocessing, its first step is data collecting from various information systems. For example, the equipment, consumer and load information from power load system, the weather information from internet. In a more realistic situation, the quality of the real data is poor, so data quality assessment is necessary [46, 47]. Currently, we only check the missing value rate of data set. In future, a load data quality assessment model will be provided. High quality data often performs well in learning-based model. We provide a detailed data processing statement in this paper. There are three main operations, including dispersing consecutive attributes into nominal, filling missing value and sampling for balancing HOL. In this paper, we aim at preventing

HOL accidents, so we discrete the consecutive load into normal, heavy and over load. Moreover, HOL classes are minority in power load data (the HOL defined as positive by convention, the majority class is negative). The learner for imbalanced data sets is always apt to predict the majority class better but behaves poorly to the minority class. So we provide a combined sampling technique for balancing data, related details will be shown in Section 4.

For data analysis submodule, it includes feature analysis, building classification model and forecasting. In smart grid environment, the power data is big, not only because it contains a huge number of records, but because it contains a large number of attributes. However, many attributes are irrelevant or redundant for classification. They considerably degrade the classification algorithm performance [48]:

1. greatly lower the efficiency;
2. cause the prediction deviation.

After comparing the existing feature selection and analysis methods, we put forward the association rules mining (ARM) technique [36, 37]. There are three reasons:

1. The ARM utilizes the frequent itemsets, it is efficient and applied easily for big data.
2. We want to extract the association among multi-attributes from various information systems.
3. The mining rules on IF THEN format supports decision making for regulators and visualized easily.

The details will be discussed in Section 5. With respect to building classification model and forecasting, we compare the learning-based methods with traditional time series and intelligent methods, the former is more suitable for our data. Then we further compare the well-known learning algorithms: SVM, bagging and Random Forest through experiments. Finally, we choose Random Forest method in this paper to prevent over-fitting of imbalanced HOL data and we gain high accuracy. The details will be shown in Section 6.

In recent years, the data visualization [8] submodule has attracted a lot of attention in the big data era. The knowledge discovered from the original data is required to be presented in a proper way to users, especially for decision makers. Visualization performs well making large data sets better accessible using techniques like selecting and zooming. Some visual results are displayed in the following sections.

4 DATA PREPROCESSING

The power load behaviour can be influenced by a number of factors, such as economic, social, time or environmental factors. Thus, the data are collected from various information systems. After fetching data, the quality of real data is often poor. In order to gain high accuracy, it is important to process the original data. In this paper, we process it through standardization and discretization, filling missing values and data balancing. More details are displayed in this section.

4.1 Data Collection

1. Collecting internal data from power database

   We fetch load information from electric power database mainly in 10 tables, which can be classified into Consumer Information, Transformer Information, Transformer Areas Information, Measure Points Information and Energy Meter Information.

2. Crawling external data from internet

   The power load behaviour can also be influenced by external factors (we define the information fetched outside the power database as external factors), such as season, holiday and weather information, etc. These information could be obtained by crawling or purchasing. For simplicity, we selected typical weather and holiday information for analysis. The two data sets are crawled from the related websites with high quality maintained.

   After getting the internal and external data sets, they are integrated into a big table based on relative keys.

4.2 Data Processing

1. Standardization and discretization

   In this paper, we are mainly interested in HOL pattern. So, it is necessary to discretize consecutive load data into different patterns. Firstly, we standardized the load data as the variants of equipment and lines. Basically, the load-rate (see Equation (1)) is often used to standardize the load data. The distribution transformer (DT) works normal when its *load-rate* ranges from $[0, 0.8]$. When the *load-rate* surpasses 0.8 the DT works under heavy or even over load and has to endure more heat and higher temperatures, which results in malfunctions. After discussion with the power production specialists, we discrete the power load into normal, heavy and over load, as presented in Table 1, based on the capacity of DT and the actual demand.

   $$\textit{load-rate} = \frac{a}{c} \tag{1}$$

   where a is the apparent power, and c is the capacity of distribution transform.

2. Filling missing values

   The poor quality of the original data set, is one challenge for implementing learning-based methods. For example, some attributes for customers can reach 40 % missing rate. Therefore, we provided a method based on k neighbours with the same class labels to fill missing values. For numeric attributes, the missing

Class	Daily Load Factor	Class Label
normal load	$[0, 0.8]$	0
heavy load	$(0.8, 1]$	1
over load	$(1, +\infty)$	2

Table 1. Discrete value of daily load factor

value is filled by the mean of k neighbours with the same label. For categorical attributes, the missing value is filled by the mode (the most frequent) values of k neighbours with the same label. In general, parameter k should not be small. If k is small, it is sensitive to outliers or noise. If k equals the number of samples in the same class, it transforms to fill the missing value by mean/mode of the samples in the same class. The code is shown in Algorithms 1 and 2.

```
Algorithm 1 fillingMissingValues
Input:
    Original data with missing values $D_{na}$;
Output:
    Idx of missing value: $id$, Filled Data: $D$;
$c_{oidx} \leftarrow which(D_{na}[, ncol(D_{na})] == 0)$; $c_{1idx} \leftarrow which(D_{na}[, ncol(D_{na})] == 1)$;
$c_{2idx} \leftarrow which(D_{na}[, ncol(D_{na})] == 2)$;
$D_{0na} \leftarrow D_{na}[c_{oidx}, ]$; $D_{1na} \leftarrow D_{na}[c_{1idx}, ]$; $D_{2na} \leftarrow D_{na}[c_{2idx}, ]$;
for $i$ in $1 : seq(ncol(D_{0na}) - 1)$ do
    if $any(idx \leftarrow is.na(D_{0na}[, i]))$) then
        $D_{0na}[idx, i] \leftarrow centralValue(D_{0na}[, i])$;
    end if
end for
for $i$ in $1 : seq(ncol(D_{1na}) - 1)$ do
    if $any(idx \leftarrow is.na(D_{1na}[, i]))$) then
        $D_{1na}[idx, i] \leftarrow centralValue(D_{1na}[, i])$;
    end if
end for
for $i$ in $1 : seq(ncol(D_{2na}) - 1)$ do
    if $any(idx \leftarrow is.na(D_{2na}[, i]))$) then
        $D_{2na}[idx, i] \leftarrow centralValue(D_{2na}[, i])$;
    end if
end for
$D \leftarrow$ replace $D_{na}$ by filled $D_{0na}, D_{1na}, D_{2na}$;
return $D$
```

3. Data balancing

   The real electric power databases are unbalanced, as the HOL accidents are relatively few. The forecasting system based on the original data distribution can easily be over-fitting and inaccurate. Therefore, data balancing is essential.

Algorithm 2 centralValue

```
Input:
    Vector: $x$, parameter: $k$; missing idx: $id$;
Output:
    centralValue: $cv$;
if $k/2 == 0$ then
    print parameter $k$ should be odd;
    return 0;
else
    if $is.numeric(x)$ then
        $k\_m \leftarrow \frac{k-1}{2}$;
        $cv \leftarrow median(x[id - k\_m + 1, id + k\_m], na.rm = T)$;
    else
        $x \leftarrow as.factor(x)$;
        $cv \leftarrow levels(x)[which.max(table(x[id - k\_m + 1, id + k\_m]))]$;
    end if
end if
return $cv$
```

There are different methods dealing with unbalanced data sets, such as random over-sampling, random under-sampling, threshold moving and ensemble [34]. Random over-sampling is unsuitable for big data. Threshold moving is relatively weaker for the multi-class problem. Ensemble techniques include several classic integrated algorithms, such as Bagging, Boosting and Random Forest, will be discussed in the following section.

Based on the characteristics of power load data, in this paper, we provided a hybrid method by combining random under-sampling and over-sampling. Beause the HOL samples are relatively fewer, if we only use random under-sampling to reducing regular samples, the training data decreases quickly, which may affect the performance of learning regular data. The hybrid method reduces regular samples to a certain degree, then adds HOL samples into the dataset. Each class samples are sufficient for learning and the learned classifier performs better with HOL. In order to further improve the accuracy, we chose different sample rates and tested them through the performance of classifiers. The comparative results are shown in Section 7.1.2.

5 FEATURE ANALYSIS BASED ON ASSOCIATION RULES MINING

Feature analysis plays an important role in classifying system. Nowadays, the power load pattern is influenced by many factors, such as equipment capacity, consumer behavior, environmental condition, demographic and economic condition, etc. So, the data are from various information systems. But, some of the attributes may be redundant or irrelevant and some of the attributes may be strongly associated.

In this paper, on one hand, we aim at analyzing the association among attributes from various information systems. On the other hand, we expect to select associated features and delete redundant or irrelevant attributes.

In terms of feature selection, it aims at reducing the irrelevant and redundant variables from the data set. One of the most popular unsupervised methods in this field is the principle component analysis (PCA) method. This method transforms the existing high-dimensional attributes into new low-dimensional ones which are the linear combination of existing attributes. Linear discriminant analysis (LDA) is a well-known supervised method for projecting high-dimensional data onto a low dimensional space where the data gains maximum class separability [41]. But the two methods are not immune against distortion under transformation. For example, a linear scaling of the input attributes can cause serious changes to the results. In addition, entropy-based feature selection method is used frequently. But it is computationally intensive [42], its computational complexity will be $O(N^2)$, where N is the number of samples. It is time-consuming for large power load data.

Recently, data mining methodologies are described by [43]. For example, association rule mining (ARM) method was used to select the most relevant features [36, 37]. For feature association analysis, ARM is the most famous method. It generates the most frequent feature subsets which are highly associated. In addition, the mining rules in format IF THEN are understood easily for users and support decision making. So, in this paper, we utilize Apriori (the most famous ARM algorithm) to feature selection and association analysis.

In this paper, the association rules are transformed into non binary problems with form $\{I_i = v_i, \ldots, I_j = v_j\} \Rightarrow \{I_k = v_k, \ldots, I_m = v_m\}$, where $\{I_i, \ldots, I_j\}$ and $\{I_k, \ldots, I_m\}$ are mutually exclusive subsets of attributes as A and B above, and $\{v_i, \ldots, v_j\}$, $\{v_k, \ldots, v_m\}$ are corresponding values of the attributes. Our rules mining process is presented as Algorithm 3. Firstly, we normalized the attributes into nominal. Then, we set the consequent as $\{load = 1\}$ and $\{load = 2\}$ respectively, the remaining attributes as antecedent to extract rules related with HOL (defined as $rules_1$ and $rules_2$), where $load$ is our target attribute and $\{1, 2\}$ represent heavy and over load. Rules are mining separately because of the their different rates in the actual data set. Finally, reducing the redundancy of $rules_1$ and $rules_2$, and splitting the attributes of the reduced redundancy rules as the features for HOL learning.

6 CLASSIFICATION MODEL BASED ON RANDOM FOREST

6.1 Comparison of Classification Models

Load classification is to partition various load pattern into groups. There are many different models which can be categorized into traditional time series methods and intelligent methods. The drawbacks of traditional methods such as linear regression, time series [11] model (ARIMA, exponential smoothing) are that they only take time

Algorithm 3 Feature analysis algorithm

```
Input:
    Preprocessed data $D = \{v_1, v_2, \ldots, v_n, c\}$;
    Parameters for association rules mining $s$, $c$ and $l$;
Output:
    Extracting related rules and features for heavy and over load and $rule_1 - pruned$, $rule_1 - pruned$, $feature$;
    Reduced data $D^*$
1:  $D' \leftarrow$ normalize the variables into nominal;
2:  $rule_1 \leftarrow apriori(D', parameter = list(minlen = 2, maxlen = l, supp = s, conf = c), appearance = list(rhs = c(''load = 1''), default = ''lhs''))$;
3:  while $length(rule_1)==0$ do
4:      $s = s \times 0.7$;
5:      $rule_1 \leftarrow apriori(D', parameter = list(minlen = 2, maxlen = l, supp = s, conf = c), appearance = list(rhs = c(''load = 1''), default = ''lhs''))$;
6:  end while
7:  $rule_1 - sort \leftarrow$ sort extracted $rule_1$ by $lift$;
8:  $rule_1 - pruned \leftarrow$ reduce redundant of $rule_1 - sort$;
9:  $feature_1 \leftarrow c()$
10: for $i$ in $1 : nrow(rule_1 - pruned)$ do
11:     $temp \leftarrow$ strsplit $rule_1 - pruned[i, 1]$;
12:     $feature_1 \leftarrow c(feature_1, temp)$
13: end for
14: $feature_1 \leftarrow$ single $feature_1$;
15: $rule_2 \leftarrow apriori(D', parameter = list(minlen = 2, maxlen = l, supp = s, conf = c), appearance = list(rhs = c(''load = 2''), default = ''lhs''))$;
16: $rule_2 - prunded \leftarrow$ sort, reduce redundancy $rule_2$ as $step3, \ldots, step8$;
17: $feature_2 \leftarrow$ strsplit $rule_2 - prunded$ as $step9, \ldots, step14$;
18: $feature \leftarrow$ union $feature_1$ and $feature_2$;
19: $D^* \leftarrow D'[, feature]$;
20: return $rule_1 - pruned$, $rule_2 - pruned$, $feature$, $D^*$
```

sequence features into consideration, and it is difficult to deal with non-linear nature of load pattern. Thus the intelligent methods have been practiced, such as expert system and artificial neural networks (ANN). The ANN forecasting model gives better performance with nonlinearity issue, but most of the NN models adopt the gradient descent based back-propagation learning scheme to minimizes the mean square error during training process. The error in the training data set is good, but performs badly when out-of-sample data is presented to the network, which yields limited generalization capability.

Recently, the machine learning methods were applied into this field. The widely utilized algorithms are Support Vector Machine (SVM) [26] and decision trees. SVM is to find a maximum-margin hyperplane which separates the n-dimensional data perfectly into its multi-classes, when dealing with nonlinearly separable problems, it is often computationally expensive and easily over-fitting. The tree model, such as CART and ID3, is over-fitting easily when dealing with unbalanced data set. Because it only builds a single decision tree which has limited generalization capacity. Thus, the ensemble learning method is proposed for its excellent generalization and accurate ability. Ensemble learning, such as Boosting, Bagging and Random Forest, is a combination of multiple tree predictors. In Boosting, base tree predictors pay attention to wrongly predicted points by earlier predictors, the wrongly predicted points were given extra weight to successive trees training. And taking a weighted vote for prediction in the end. In Bagging, successive trees are constructed based on a bootstrap sample of data independently instead of depending on earlier trees. And the majority vote is taken for prediction in the end.

While, Random Forest is another ensemble learning method with little difference. Firstly, bootstrap samples are taken from the original data as in Bagging. For each of the bootstrap samples, it changes how the classification trees are built, at each node, instead of using the best split among all variables as standard trees, using the best among a randomly chosen subset of predictors at that node. It turns out that the somewhat counterintuitive strategy performs better than other classifiers such as SVM and neural networks and also are robust against [33] over-fitting. In the end, a majority vote for classification or average for regression is taken for prediction as in Bagging. In addition, the generalization error for forests converges to a limit as the number of trees in the forest becomes large [33].

The objective analysis of various classification methods on real load data is shown in Section 7. In order to gain in accuracy, we also compare the generalization error of different number of trees through the experiments.

6.2 Evaluating Metric for Classification Model

In general, the performance of classifiers is evaluated in terms of testing error and training time. In this paper, we mainly pay attention to HOL classes. Because HOL accidents will cause huge damage when incorrectly forecasted to normal load. So, we provided a particular evaluation metric for the proposed model. We firstly computed the accuracy (abbreviated as A) of all correctly classified load of the classifier. Then, we observed sensitivity (abbreviated as S) of correctly classified HOL of big concern. The sensitivity is often used to evaluate the classifiers' performance on unbalanced data. In addition, we calculated the error-risk (abbreviated as R) of HOL wrongly classified as regular load, which could lead to enormous losses. In this paper, we expect error-risk towards zero as well as with high sensitivity. Table 2 illustrates a confusion matrix of our three-class problem and the computational formula of the above evaluation metrics as in Equations (2), (3), (4), where n_{ij} represents the number of original class i that is predicted into class j.

Class	Predicted Normal	Predicted Heavy Load	Predicted Over Load
normal load	n_{11}	n_{12}	n_{13}
heavy load	n_{21}	n_{22}	n_{23}
over load	n_{31}	n_{32}	n_{33}

Table 2. Confusion matrix of three-class problem

$$A = \frac{n_{11} + n_{22} + n_{33}}{n_{11} + n_{12}, \ldots, + n_{32} + n_{33}}, \tag{2}$$

$$S = \frac{n_{22} + n_{33}}{n_{21} + n_{22} + n_{23} + n_{31} + n_{32} + n_{33}}, \tag{3}$$

$$R = \frac{n_{21} + n_{31}}{n_{21} + n_{22} + n_{23} + n_{31} + n_{32} + n_{33}}. \tag{4}$$

7 EXPERIMENTS

In this section, we are going to examine the performance of our comprehensive learning model on real datasets. We will see that, the methods used in our model performs better than others.

7.1 Dataset Preprocessing

7.1.1 Data Description

In this paper, the internal power load data are fetched from the power system. It contains three years power records of Shandong Province in China, 949 952 records in total. And, it contains 24 condition attributes (see Table 3), which can be classified into 5 types: Consumer Information, Transformer Information, Transformer Areas Information, Measure Points Information and Energy Meter Information.

We choose Weather and Holiday as the external factors which were crawled from [45]. The external factors includes 7 attributes, where HIGH_TEMP and LOW_TEMP mean the highest and the lowest temperature of a day, OUKLOOK is the weather condition of a day, MEAN_HIGH and MEAN_LOW indicate mean highest and lowest temperature of a month, LAW_HOLIDAY is a binary attribute and indicates legal holiday.

In this paper, we partition the dataset into training and testing data by 7:3. All the experiments are conducted on a computer with 16 GB RAM and Intel Core i5 CPU running the Microsoft Windows 7 Professional operating system. All algorithms are implemented using R language.

Consumer Information	TRADE_CODE DATA_WHOLE_FLAG LODE_ATT_CODE ELEC_TYPE_CODE	CONTRACT_CAP SHIFT_NO HEC_INDUSTRY_CODE CONS_SORT_CODE	RUN_CAP CHK_CYCLE RRIO_CODE ORG_NO
Transformer Information	INST_DATE COOL_MODE MS_FLAG	PLATE_CAP CHG_CAP	FRST_RUN_DATE PROTECT_MODE
Transformer Areas	RUN_STATUS_CODE TYPE_CODE		
Measure Point	VOLT_CODE		
Energy Meter	T_FACTOR		
Weather Holiday Information	HIGH_TEMP MEAN_HIGH LAW_HOLIDAY	LOW_TEMP MEAN_LOW	OUKLOOK WEEK_DAY

Table 3. Data attributes

7.1.2 Balancing Rate

The original rates of normal, heavy load and over load samples are 0.957, 0.026, 0.017. We can see that, the power load data is extremely unbalanced, HOL cases as our interested classes are minority. To increase sensitivity (S) of HOL minority, we combine random under and over sampling techniques to balance real power data. In this paper, we test several different sample rates for balancing data to set a optimal sample rate. The accuracy (A), sensitivity (S) and error-risk (R) of the classifiers based on different sample rates are shown on Figure 2. We can see that, even if the accuracy (A) decreases, our concerned sensitivity (S) increases with the increasing sample rate of HOL, meanwhile, the error-risk (R) tends towards zero. But, the sample rate is not the bigger the better, the sensitivity (S) decreases when the sample rate reaches 40. So, we set optimal sample rate as 40 for the following test, where sample-rate = $(n_1 + n_2)/n_0$ and n_0, n_1, n_2 represent the number of normal, heavy and over load, respectively.

Customer Related Attribute	Meaning
CHK_CYCLE	check cycle (mouth)
PLATE_CAP	plate capacity
RUN_CAP	runtime capacity
SHIFT_NO	shifts number (refer to State Grid Corporation of China)
ELEC_TYPE_CODE	electric type code (refer to State Grid Corporation of China)
ORG_NO	power supply unit number
CONS_SORT_CODE	consumer category (01:low voltage non resident)

Table 4. Extracted consumer variables

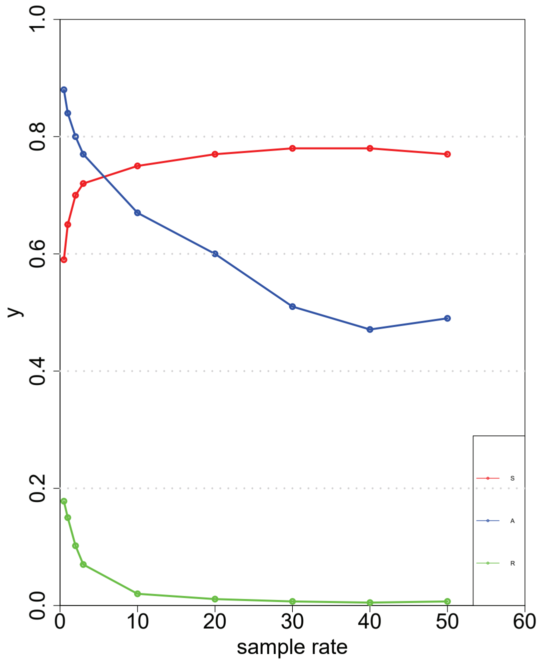
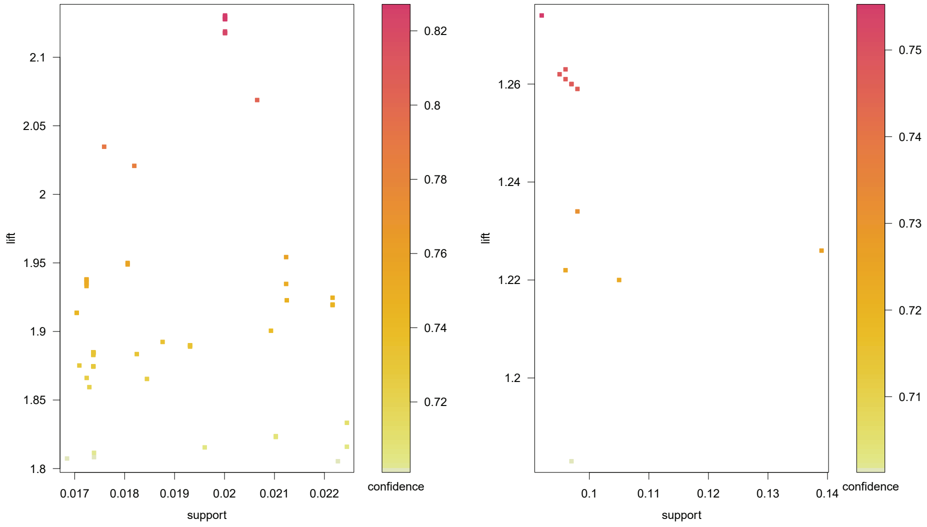


Figure 2. Comparison of different balancing rates

7.2 Feature Analysis

After balancing data, we utilize Apriori algorithm for feature selection and association analysis. Figure 3 shows the scatter plot for extracted $rule_2$ and $rule_1$, respectively. We can see that the mining rules are both with high *confidence* and *lift*, where the *confidence* and *lift* are two popular measure for association rules mining [35]. All lifts are ($\gg 1$) which means strong association with HOL. Further, we visualize the first six rules of $rule_2$ and $rule_1$ with highest lift in Figure 4.

Most attributes on the mining rules are mostly consumer related attributes, see Table 4. For example, the rule $\{\mathrm{RUN\_CAP} = 400, \mathrm{CONS\_SORT\_CODE} = 1\} \Rightarrow \{load = 1\}$ means that capacity shortage of the low voltage non resident is prone to cause heavy load. Meanwhile, holiday factor will effect load behavior too, $\{\mathrm{RUN\_CAP} = 400, \mathrm{PLATE\_CAP} = 400, \mathrm{LAW\_HOLIDAY} = 0\} \Rightarrow \{load = 1\}$ implies that on weekdays capacity shortage tends to heavy load, (where $\mathrm{LAW\_HOLIDAY} = 0$) means non-legal holiday. In addition, $\{\mathrm{INST\_DATE} = 1990/1/1\} \Rightarrow \{load = 2\}$ shows that aging facilities are with high probability of over load, so the electric power group should change device periodic. So, the min-



a) Scatter plot for rules of over load

b) Scatter plot for rules of heavy load

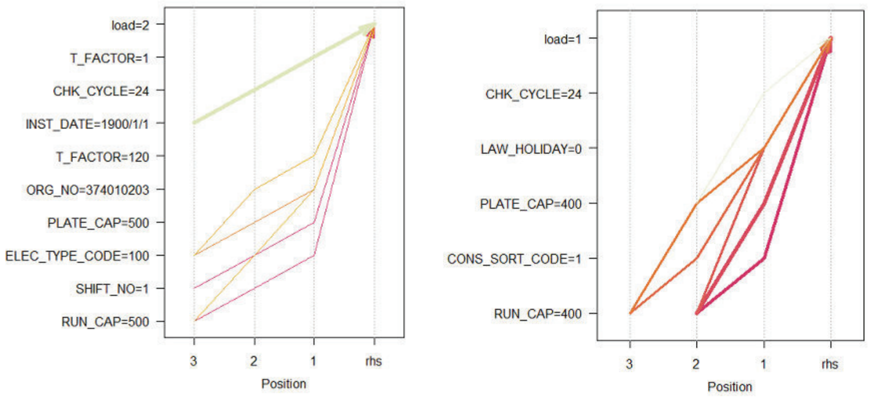
Figure 3. Scatter plot for heavy and over load

ing rules will help decision makers to take measures to prevent heavy and over load phenomenon and consumers can adjust their electricity consumption strategies more economically.

In addition, to validate the effectiveness of ARM for feature selection, we compare ARM method with the traditional feature selection method based on expertise (defined as EXPERT). The attributes based on EXPERT contains the Weather and time series attributes, where the time series attributes refers to former seven days load. The performance of the two random forest model based on ARM and EXPERT are shown in Figure 5. We can see that the sensitivity (S) of ARM-RF (Random Forest model based on ARM) performs better than EXPERT-RF (Random Forest model based on EXPERT), the error-risk (R) of the two models both approach zero. So, the ARM method for feature selection is applicable.

Data Size	RF Time (s)	SVM Time (s)
$3 * 10^3$	3	34
$3 * 10^4$	30	300
$3 * 10^5$	420	break

Table 5. Efficiency, S and R comparison of different data size



a) Rule 2 of head lift

b) Rule 1 of head lift

Figure 4. Rules of head lift

7.3 Classification Methods Analysis

After comparison with traditional time series, intelligent method and learning based method in Section 5, the latter is more suitable in big electric power data. In learning based methods, we finally utilize Random Forest algorithm for model learning, because Random Forest method prevents unbalanced power data from over-fitting and has a much higher efficiency. In this part, we test it through objective experiments with other well known machine learning methods, Bagging, SVM. Firstly we compare the sensitivity (S) and error-risk (R) of Random Forest (RF), Bagging and SVM, Random Forest performs best among these three algorithms, see Figure 6. But, we also see that the S of HOL based on Random Forest is only 0.77, this is because HOL classes are actually quite similar. In application, it is reasonable to neglect the indistinguishability between heavy and over load, because misclassification between them bring little risk. Based on our methods, we minimize the error-risk as well as with highest sensitivity, the error-risk (R) of these three classifiers is zero. In addition, we have to mention that the SVM algorithm is more time consuming than Random Forest (RF) during experiment. See Table 5, with the data size increasing, the SVM method tends to break. But, the RF model is robust with good scalability. So, our HOL forecasting system based on Random Forest is effective and accurate.

Then, we validate that the generalization error for forests converges to a limit as the number of trees in the forest becomes large, see Figure 7. So, there is no need to construct a big Random Forest for efficiency.

In summary, all these experiments validate that our comprehensive learning model and methods for HOL forecasting are precise, robust and good for application.

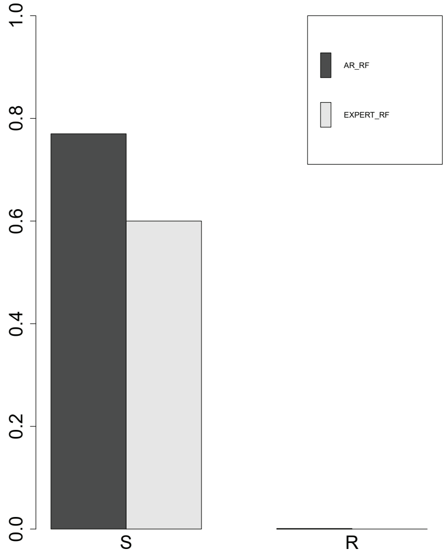


Figure 5. The S and R comparison of ARM_RF and EXPERT_RF

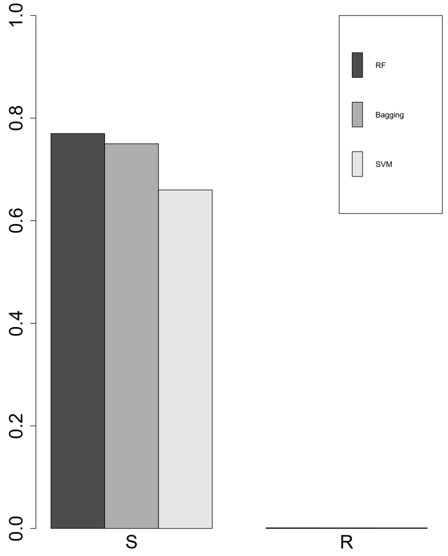


Figure 6. Performance comparison of RF, Bagging and SVM

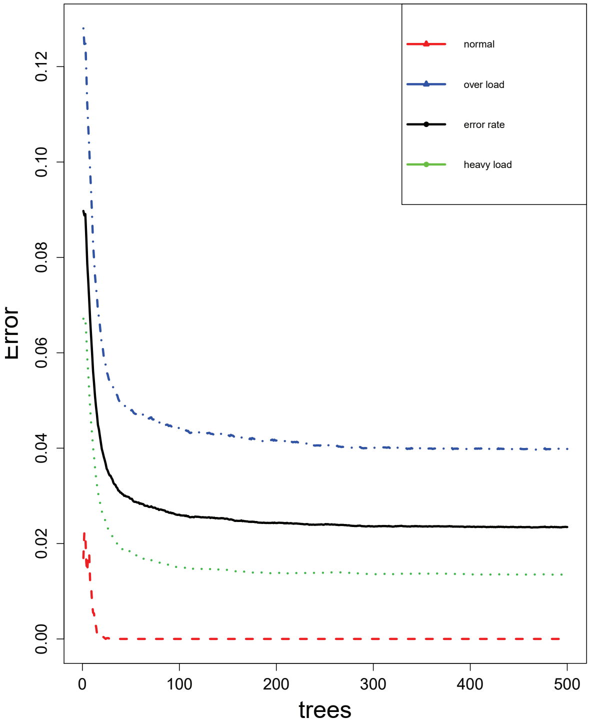


Figure 7. The error rate of Random Forest with different trees

8 CONCLUSIONS

This paper reported a comprehensive heavy and over load classification model in smart grid environment, we provide detailed information of every module. The forecasting model was generated by combined random under- and over-sampling techniques for imbalanced power data, association rules mining (ARM) for valuable feature selection and decision analysis, Random Forest model for highest precision forecasting. It is proved that such model and process architecture are efficient, accurate and implementable. The system and analysis method has already been successfully applied and evaluated in some grid corporations for decision-making and risk preventing.

The presented work will be further developed and extended. One possible direction is to develop an adapted heavy and over load forecasting model to apply in different regions and achieving the paralleled Random Forest algorithms.

Acknowledgment

This work is supported by National Science Foundation of China Grant No. 61672-088, Fundamental Research Funds for the Central Universities Nos. 2016JBM022, 2016YJS047 and 2015ZBJ007, Open Research Funds of Guangdong Key Laboratory of Popular High Performance Computers. The corresponding author is Yidong Li.

REFERENCES

[1] Hernandez, L.—Baladron, C.—Aguiar, J. M.: A Survey on Electric Power Demand Forecasting: Future Trends in Smart Grids, Microgrids and Smart Buildings. IEEE Communications Surveys & Tutorials, Vol. 16, 2014, No. 7, pp. 1460–1495, doi: 10.1109/surv.2014.032014.00094.

[2] Javed, F.—Arshad, N.—Wallin, F.: Forecasting for Demand Response in Smart Grids: An Analysis on Use of Anthropologic and Structural Data and Short Term Multiple Loads Forecasting. Applied Energy, Vol. 96, 2012, pp. 150–160, doi: 10.1016/j.apenergy.2012.02.027.

[3] Yang, S.—Shen, C.: A Review of Electric Load Classification in Smart Grid Environment. Renewable and Sustainable Energy Reviews, Vol. 24, 2013, pp. 103–110, doi: 10.1016/j.rser.2013.03.023.

[4] Nazih, A. S.—Fawwaz, B.: Medium-Term Electric Load Forecasting Using Singular Value Decomposition. Energy, Vol. 36, 2011, No. 7, pp. 4259–4271.

[5] Ekonomou, L.: Greek Long-Term Energy Consumption Prediction Using Artificial Neural Networks. Energy, Vol. 35, 2010, No. 2, pp. 512–517, doi: 10.1016/j.energy.2009.10.018.

[6] Alfares, H. K.—Nazeeruddin, M.: Electric Load Forecasting: Literature Survey and Classification of Methods. International Journal of Systems Science, Vol. 33, 2002, No. 1, pp. 23–34, doi: 10.1080/00207720110067421.

[7] Gonzalez, R. E.—Jaramillo-Moran, M. A.—Carmona, F. D.: Monthly Electric Energy Demand Forecasting Based on Trend Extraction. IEEE Transactions on Power Systems, Vol. 21, 2006, No. 4, pp. 1946–1953.

[8] Hwang, D.—Jung, J. E.—Park, S.: Social Data Visualization System for Understanding Diffusion Patterns on Twitter: A Case Study on Korean Enterprises. Computing and Informatics, Vol. 33, 2015, No. 3, pp. 591–608.

[9] Pappas, S. S.—Ekonomou, L.—Karampelas, P.: Electricity Demand Load Forecasting of the Hellenic Power System Using an ARMA Model. Electric Power Systems Research, Vol. 80, 2010, No. 3, pp. 256–264, doi: 10.1016/j.epsr.2009.09.006.

[10] Ediger, V. S.—Akar, S.: ARIMA Forecasting of Primary Energy Demand by Fuel in Turkey. Energy Policy, Vol. 35, 2007, No. 3, pp. 1701–1708, doi: 10.1016/j.enpol.2006.05.009.

[11] Vo, V.—Luo, J.—Vo, B.: Time Series Trend Analysis Based on K-Means and Support Vector Machine. Computing and Informatics, Vol. 35, 2016, No. 1, pp. 111–127.

[12] Valenzuela, O.—Rojas, I.—Rojas, F.: Hybridization of Intelligent Techniques and ARIMA Models for Time Series Prediction. Fuzzy Sets and Systems, Vol. 159, 2008, No. 7, pp. 821–845, doi: 10.1016/j.fss.2007.11.003.

[13] Taylor, J. W.: Short-Term Electricity Demand Forecasting Using Double Seasonal Exponential Smoothing. Journal of the Operational Research Society, Vol. 54, 2003, No. 8, pp. 799–805, doi: 10.1057/palgrave.jors.2601589.

[14] Weron, R.: Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach. John Wiley & Sons, 2007.

[15] Taylor, J. W.: Triple Seasonal Methods for Short-Term Electricity Demand Forecasting. European Journal of Operational Research, Vol. 204, 2010, No. 1, pp. 139–152, doi: 10.1016/j.ejor.2009.10.003.

[16] Wang, H.—Li, B. S.—Han, X. Y.—Wang, D. L.—Jin, H.: Study of Neural Networks for Electric Power Load Forecasting. In: Wang, J., Yi, Z., Zurada, J. M., Lu, B. L., Yin, H. (Eds.): Advances in Neural Networks (ISNN 2006). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 3972, 2006, pp. 1277–1283.

[17] Ferreira, V. H.—Da, S.—Alexandre, P. A.: Toward Estimating Autonomous Neural Network-Based Electric Load Forecasters. IEEE Transactions on Power Systems, Vol. 22, 2007, No. 4, pp. 1554–1562, doi: 10.1109/tpwrs.2007.908438.

[18] Sousa, J. C.—Neves, L. P.—Jorge, H. M.: Assessing the Relevance of Load Profiling Information in Electrical Load Forecasting Based on Neural Network Models. International Journal of Electrical Power & Energy Systems, Vol. 40, 2012, No. 1, pp. 85–93.

[19] Hernández, L.—Baladrón, C.—Aguiar, J.: Artificial Neural Networks for Short-Term Load Forecasting in Microgrids Environment. Energy, Vol. 75, 2014, pp. 252–264, doi: 10.1016/j.energy.2014.07.065.

[20] Hernández, L.—Baladrón, C.—Aguiar, J.: Experimental Analysis of the Input Variables Relevance to Forecast Next Day Aggregated Electric Demand Using Neural Networks. Energies, Vol. 6, 2013, No. 6, pp. 2927–2948.

[21] Hinojosa, V. H.—Hoese, A.: Short-Term Load Forecasting Using Fuzzy Inductive Reasoning and Evolutionary Algorithms. IEEE Transactions on Power Systems, Vol. 25, 2010, No. 1, pp. 565–574, doi: 10.1109/tpwrs.2009.2036821.

[22] Lou, C. W.—Dong, M. C.: Modeling Data Uncertainty on Electric Load Forecasting Based on Type-2 Fuzzy Logic Set Theory. Engineering Applications of Artificial Intelligence, Vol. 25, 2012, No. 8, pp. 1567–1576.

[23] Che, J.—Wang, J.—Wang, G.: An Adaptive Fuzzy Combination Model Based on Self-Organizing Map and Support Vector Regression for Electric Load Forecasting. Energy, Vol. 37, 2012, No. 1, pp. 657–664.

[24] Liao, G. C.: A Novel Particle Swarm Optimization Approach Combined with Fuzzy Neural Networks for Short-Term Load Forecasting. Power Engineering Society General Meeting, IEEE, 2007, pp. 1–6, doi: 10.1109/pes.2007.385688.

[25] Chen, B. J.—Chang, M. W.—Lin, C. J.: Load Forecasting Using Support Vector Machines: A Study on EUNITE Competition. IEEE Transactions on Power Systems, Vol. 19, 2004, No. 4, pp. 1821–1830.

[26] Bartok, J.—Babič, F.—Bednár, P.: Data Mining for Fog Prediction and Low Clouds Detection. Computing and Informatics, Vol. 31, 2013, No. 6, pp. 1441–1464.

[27] Lee, J. W.—Park, R. H.—Chang, S. K.: Local Tone Mapping Using the K-Means Algorithm and Automatic Gamma Setting. IEEE Transactions on Consumer Electronics, Vol. 57, 2011, No. 1, pp. 209–217.

[28] Bishnu, P. S.—Bhattacherjee, V.: Software Fault Prediction Using Quad Tree-Based K-Means Clustering Algorithm. IEEE Transactions on Knowledge and Data Engineering, Vol. 24, 2012, No. 6, pp. 1146–1150, doi: 10.1109/tkde.2011.163.

[29] Labeeuw, W.—Deconinck, G.: Residential Electrical Load Model Based on Mixture Model Clustering and Markov Models. IEEE Transactions on Industrial Informatics, Vol. 9, 2013, No. 3, pp. 1561–1569.

[30] Bidoki, S. M.—Mahmoudi, K. N.—Gerami, S.: Comparison of Several Clustering Methods in the Case of Electrical Load Curves Classification. 2011 16th Conference on Electrical Power Distribution Networks (EPDC), IEEE, 2011, pp. 1–7.

[31] Ferreira, A. M. S.—Cavalcante, C. A. M. T.—Fontes, C. H. O.—Marambio, J. E. S.: A New Method for Pattern Recognition in Load Profiles to Support Decision-Making in the Management of the Electric Sector. International Journal of Electrical Power & Energy Systems, Vol. 53, 2013, pp. 824–831, doi: 10.1016/j.ijepes.2013.06.001.

[32] Yu, D.—Yu, X.—Hu, Q.: Dynamic Time Warping Constraint Learning for Large Margin Nearest Neighbor Classification. Information Sciences, Vol. 181, 2011, No. 13, pp. 2787–2796.

[33] Liaw, A.—Wiener, M.: Classification and Regression by RandomForest. R News, Vol. 2, 2002, No. 3, pp. 18–22.

[34] Han, H.—Wang, W. Y.—Mao, B. H.: Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Huang, D. S., Zhang, X. P., Huang, G. B. (Eds.): Advances in Intelligent Computing (ICIC 2005). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 3644, 2005, pp. 878–887.

[35] Hahsler, M.—Chelluboina, S.: Visualizing Association Rules: Introduction to the R-Extension Package ArulesViz. R Project Module, 2002, pp. 223–238.

[36] Karabatak, M.—Ince, M. C.: A New Feature Selection Method Based on Association Rules for Diagnosis of Erythemato-Squamous Diseases. Expert Systems with Applications, Vol. 36, 2009, No. 10, pp. 12500–12505, doi: 10.1016/j.eswa.2009.04.073.

[37] Shahzad, W.—Asad, S.—Khan, M. A.: Feature Subset Selection Using Association Rule Mining and JRip Classifier. International Journal of Physical Sciences, Vol. 8, 2013, No. 18, pp. 885–896.

[38] Zhou, K. L.—Yang, S. L.: An Improved Fuzzy C-Means Algorithm for Power Load Characteristics Classification. Power System Protection and Control, Vol. 40, 2012, No. 22, pp. 58–63.

[39] López, J. J.—Aguado, J. A.: Hopfield CK-Means Clustering Algorithm: A Proposal for the Segmentation of Electricity Customers. Electric Power Systems Research, Vol. 81, 2011, No. 2, pp. 716–724.

[40] Nizar, A. H.—Dong, Z. Y.—Zhao, J. H.: Load Profiling and Data Mining Techniques in Electricity Deregulated Market. Power Engineering Society General Meeting, IEEE, 2006, pp. 7.

[41] Duda, R. O.—Hart, P. E.—Stork, D. G.: Pattern Classification. John Wiley & Sons, 2012.

[42] Hild, K. E.—Erdogmus, D.—Principe, J. C.: An Analysis of Entropy Estimators for Blind Source Separation. Signal Processing, Vol. 86, 2006, No. 1, pp. 182–194.

[43] Akbar, S.—Rao, K. N.—Chandulal, J. A.: Intrusion Detection System Methodologies Based on Data Analysis. International Journal of Computer Applications, Vol. 5, 2010, No. 2, pp. 10–20.

[44] Fan, S.—Hyndman, R. J.: Short-Term Load Forecasting Based on a Semi-Parametric Additive Model. IEEE Transactions on Power Systems, Vol. 27, 2012, No. 1, pp. 134–141.

[45] Weather Data. Availaible on: `http://lishi.tianqi.com`.

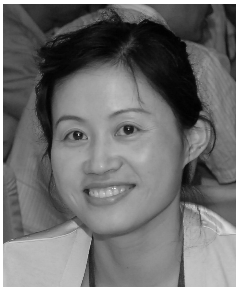
[46] Li, Y.—Shen, H.—Lang, C.—Dong, H.: Practical Anonymity Models On Protecting Private Weighted Graphs. Neurocomputing, Vol. 218, 2016, pp. 359–370.

[47] Li, Y.—Shen, H.: On Identity Disclosure Control for Hypergraph-Based Data Publishing. IEEE Transactions on Information Forensics & Security, Vol. 8, No. 8, pp. 1384–1396, August 2013.

[48] Mao, R.—Cai, T.—Li, R.-H.–Xu Yu, J.—Li, J.: Efficient Distance-Based Representative Skyline Computation in 2D Space. World Wide Web, Vol. 20, 2016, No. 4, pp. 621–638.

Huifang Li received her Bachelor's degree with honors in computer science and technology in 2013 from Henan Normal University. Now she is a graduate student at Beijing Jiaotong University majoring in computer science and technology. Her research interests mainly include data mining, deep learning and cloud computing. She has extensive research experience and practice in data mining.

Yidong Li received his B.Sc. from Beijing Jiaotong University, his M.Sc. and Ph.D. from the University of Adelaide, South Australia. He is currently Associate Professor at the School of Computer and Information Technology, Beijing Jiaotong University, China. His research interests include privacy preserving data analysis, social network analysis, web mining, and distributed computing. He has published more than 30 papers in international journals and conferences, and serves in the program committees of more than 15 international conferences.



Hairong Dong received her B.Sc. and M.Sc. in automatic control and basic mathematics from Zhengzhou University, Zhengzhou, China, in 1996 and 1999, respectively, and the Ph.D. in general and fundamental mechanics from Peking University, Beijing, China, in 2002. She is currently Professor with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. She is a Senior Member of IEEE. Her research interests include intelligent transportation systems, automatic train operation, cooperative control of multiple trains, driver assistance system, parallel control and management for high-speed railway systems, fault diagnosis and fault-tolerant control for high-speed trains. She serves as the associate editor of IEEE Transactions on Intelligent Transportation Systems, IEEE Intelligent Transportation Systems Magazine, and ACTA Automatica SINICA.