Research on Power User Intelligent Load Forecasting Method Based on Data State Drive

Da Lv[1], Jianjie Xue[2], Yan Zhang[3], Mengcong Tang[4], Yannan Wu[5], Qian Chen[6]

[1]Safety supervision department, State Grid Beijing Electric Power Company, Beijing, 100075, China

[2]Department of control center, State Grid Beijing Electric Power Company, Beijing, 100075, China

[3]Science and technology department, State Grid Beijing Electric Power Research Institute, Beijing, 100075, China

[4]Department of control center, State Grid Beijing Electric Power Company, Beijing, 100075, China

[5]Department of control center, State Grid Beijing Electric Power Company, Beijing, 100075, China

[6]Department of Power grid center, State Grid Beijing Electric Power Research Institute, Beijing, 100075, China

[*]Corresponding author's e-mail: 779656332@qq.com

Abstract. this paper proposed a variable input structure SVM prediction model based on state analysis. By identifying the key load characteristics in the load data and storing the load characteristics instead of the original load data, user load data can be realized more accurate and efficient. Based on the variable input structure SVM model of state analysis, the same state load is searched according to the results of state prediction in each period of the forecast day, and the same state historical load is used as the input factor of the model to predict. It effectively overcomes the interference of user power drift effect on load forecasting., and the forecasting accuracy is effectively improved.

1. Overview

With the large-scale installation of the intelligent meter of the power information collection system, the massive user load data generated has caused the "big data" problem in many fields such as data transmission, storage, processing and application. The data generated by the user side intelligent meter is usually transmitted to the data aggregator at the distribution transformer terminal through narrow band power line carrier communication, and then uploaded to the data center of the grid company by the data aggregator through 3G/4G communication means. Because of the limited bandwidth of narrow band power line carrier communication, when serval intelligent meters are transmitting data at the same time, it will cause channel collisions sometimes, so as to reduce the reliability of data communication [1]. Obviously, big data will also cause trouble to the storage and processing of data, it will increase data storage cost and processing time. In addition, the application of load data, such as load clustering [2-5], user classification [2, 5-9], load forecasting [10], and demand-side response [11-

13], will significantly increase the computational time and complexity in large data volumes. Therefore, it is in urgent need of a precise and efficient load data processing method.

On the other hand, the massive user load data collected by the smart meter makes it possible to analyse the electricity using behaviour of each user, making the previously invisible residents visible, and deriving a series of new applications in power systems with the user load data as the basis, so as to provide unprecedented insight into user behaviour in power systems. Load forecasting, as the most typical application of load data, helps us forecast the overall power system demand in advance. For example, the system load forecast helps us predict the change of demand of the whole system. The bus load forecast helps us predict the demand change of network nodes. Such a coarse-grained load forecast in the past can no longer meet the personalized application needs of the "big data era". User load, as the "molecular" constituting the demand of the power system, the short term forecast of the user load is to predict the short term load curve of the user a few days ago, which can help us grasp the power demand changes of each household and each industry on a micro scale, thus laying the data foundation for a series of new user oriented applications such as the demand-side response, even "recognizing the whole through observation of the part" to restore the demand of the entire power system accumulatively, and improving the "insight" of the power system to reach the unprecedented "molecular level".

The traditional prediction model input factor construction strategy usually adopts the fixed input structure, that is, the time lag between the input factor and the output (usually a multiple of the load time period, such as 24h, 48h, ..., 168h and so on) is fixed, and will not change as the time. This is very effective for the rather regular forecasting object of system load, because the system load rise/peak/fall/valley occurrence time varies little in different days, and the fixed time lag makes the historical load in the same or similar status with the forecasting load as the input factor of the predication model, which can effectively improve the prediction accuracy. However, as to the predication object with high uncertainty, due to the uncertainty of the occurrence time and end time of different events on the different days, if the prediction model of the fixed input structure is still adopted, it will cause that the historical load in different status with the forecasting load becomes the input factor of the prediction model, which is an interference for the load forecasting on the contrary.

In order to solve the problems of huge user load data volume and failed traditional predication model, this paper proposes an SVM prediction model construction strategy (VISVM) with the variable input structure to perform user short term load forecasting. This paper puts forward the distribution modelling and status analysis methods of user load, and constructs the user load status matrix. In the aspect of user load forecasting, based on the user load status matrix as well, the status of the user in various periods of the forecasting day can be predicated, the predication model constructing strategy of the variable input structure is proposed. The strategy searches the same status load of the historically similar periods according to the status forecasting result of various periods in the forecasting day, and the historical load of the same status is used as the model input factor for forecasting. Compared with the traditional forecasting method of fixed input structure, VISVM method screens the same state load as the input, avoiding putting loads of different statuses into the model, thus interfering with the forecasting. Furthermore, it enhances the correlation between model input and forecast, so as to effectively improve the forecasting accuracy.

2. Basic concept and method

The VISVM method of the load prediction method proposed in this paper is based on the user load state matrix, as shown in the method framework of Figure 1.

For user load forecasting, the VISVM method searches for the historical load in the same state in the day to be predicted based on the user historical load state matrix and sends it as an input data to the SVM model, which can be divided into six steps of A-B and G-J. The four steps of G-J respectively complete state forecasting, input factor heuristic search, SVM training, and SVM forecasting.
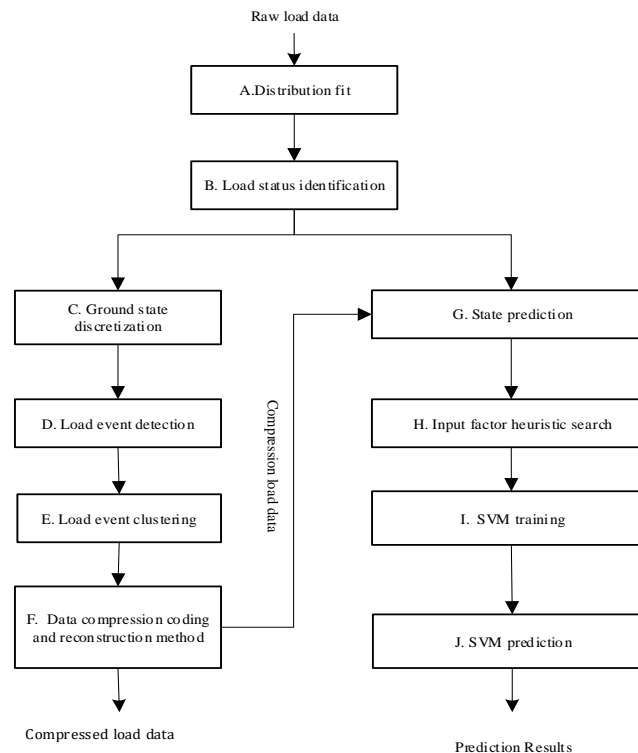
Figure 1. Method framework

3. User load characteristics and distribution modelling

3.1. Residential load characteristics

currently, smart meters widely used in the industry usually record the user's power consumption every 30 minutes at a sampling rate of 30 minutes. The residential load has two characteristics. First, the difference in electricity consumption recorded for two consecutive periods is usually small. Second, the statistical results of a large amount of data show that the residential load obeys the generalized extreme value distribution (GEV).

The "continuous small difference" characteristic of the residential load refers to the residential electricity consumption recorded for two consecutive periods, and the difference between them is usually negligible compared with the peak value of the electricity consumption on the current day. This feature was first discovered and proposed by the literature [1] in the residential load data at 1 s sampling interval. Here, we confirm this characteristic in the residential load data of the 30 min sampling interval and propose that the lower the residential load level, the more significant this characteristic.

The analysed residential load data was derived from the Sustainable Energy Authority of Ireland (SEAI) [14], which published the entire data set (user anonymity) of the Irish Smart Meter Pilot Project from 2009 to 2010 online. The data set contains more than 5,000 households and industrial & commercial user loads. Users participating in the pilot are obtained through rigorous selection and recruitment to ensure that these users are typical, and their power consumption curve can reflect the characteristics of the electricity consumption curve of users across the country [15]. The sampling interval for the Irish Smart Meter pilot data is 30 min. In order to evaluate the proportion of load sampling points with the "continuous small difference" characteristic, the continuous difference of each resident user in the Irish data set is statistically analysed here.

As shown in Fig. 2, for the cumulative probability distribution of the continuous difference rate of a typical user #1008 in the pilot project, the horizontal axis shows the continuous difference rate, and the vertical axis shows the cumulative probability, indicating that the continuous difference rate is smaller than the percentage of load points for the continuous difference rate set on horizontal axis. The continuous difference rate $r_{n,t}$ represents the ratio of the load continuous difference to the load peak value at the t time on the n day, and the calculation formula is as follows

$$r_{n,t} = \frac{P_{n,t} - P_{n,t-1}}{P_{n,max}} \tag{1}$$

Wherein, $P_{n,t}$ is the load of the time period t on the day n, $P_{n,t-1}$ is the load of the time period $t-1$ on the day n, $P_{n,max}$ is the peak load value of on the day n.
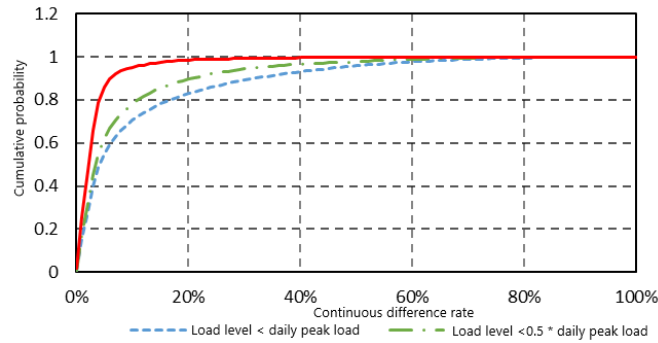

Figure 2. Cumulative probability distribution of continuous difference rate of residential users #1008

As shown by the Fig 2 blue curve, the 70% continuous difference rate for residential user #1008 is less than 10%, which means that the 70% continuous difference is less than 10% of the daily load. If the daily load error of 10% is ignored, the continuous small difference characteristic also means that 70% of the load values in the daily load curve of the residents are the same. This feature is extremely advantageous for residential load data processing because it shows that most of the residential load data are similar, using the same load data value to sub-divide the largest tiny residential load data.

If only the sampling points with load levels less than 50% of the daily load peak are counted, the green curve in Figure 2 shows that the probability of a continuous difference rate of less than 10% rises to 78%. The red curve in Figure 2 shows that if the statistical load level drops to 10% of the daily load peak, the probability increases to 95%. This shows that the lower the resident load level, the more significant the continuous small difference characteristic. This phenomenon indicates that for the daily load curve of residents, the lower the load level, the smaller the load value between adjacent points, the more stable the load, and the higher the load level, the between adjacent points. The larger the load value changes, the more the load becomes unstable.

3.2. User load distribution modelling

The GMM model is a mixed distribution model that can be used to uniformly model data from different distributions. The probability density function of a mixed distribution (PDF) can be expressed as a weighted sum of a series of known limited PDFs (usually normal distributions, but also other distributions). For a GMM model containing K finite distributions $f_k()$, the probability density of the load x under the model is as follows:

$$f(x; \Psi) = \sum_{k=1}^{K} \lambda_k f_k(x; \theta_k) \tag{2}$$

Wherein, λ_k is the weight of the finite distribution $f_k()$, Ψ is all parameters of the GMM model, θ_k is the parameter of the k finite distribution $f_k()$, and $f_k(x; \theta_k)$ is the probability density of x belongs to the k finite distribution $f_k()$, $\lambda_k f_k(x; \theta_k)$ is the weighted probability density, and $f(x; \Psi)$ is the probability density of x belonging to the GMM model. Since the probability density integral of the mixed distribution is 1, the sum of the weights is also equal to 1, as shown in equation (3):

$$\sum_{k=1}^{K} \lambda_k = 1 \tag{3}$$

In the case that the load x, the finite distribution $f_k()$, and the distributions number K are known, λ_k can be solved by the Expectation Maximum (EM). The class k with the largest posterior probability is selected as the classification of P, and the 1-posterior probability P belongs to the risk of the class k. The decision of K can be increased from $K=2$, and the minimum K value whose risk sum is less than the threshold β is selected as the optimal K value

4. Variable input structure support vector machine prediction model based on state analysis

The broadcasting model of a fixed input structure means that the input factors of the model are fixed. Due to the drift, inertia and re-entry characteristics of residential electricity, it may cause the difference for residents to use the power state (ground state/excited state) on the day of the forecast and the historical day. Since the residential load of different states is very large, if the traditional fixed input structure prediction model is adopted, the historical load opposite to the predicted daily power consumption state will be input into the prediction model, which will cause interference to the load prediction.

In this section, a variable input structure SVM prediction model based on state analysis is proposed, and is different from the fixed input structure prediction model, the input factors of the SVM prediction model constructed by this method are variable, and the relevant factors are obtained through heuristic search, which makes the input factors of the model to be most relevant to predictive objects. This strategy can cope with the prediction scenarios in which the prediction object has a high degree of uncertainty, ensuring that the input of the prediction model does not interfere with the prediction.

4.1. User load status prediction

The resident load status is more easily affected by the type of day, therefore, when the resident load status is statistic, a sub-day type is required. Because the resident load state has a significant "near-large and small" effect, the correlation coefficient with the historical load state is generally high, and the correlation coefficient with the historical load state is low. When counting the resident load status, it should also consider the recent Historical load status. Based on these two principles, this paper proposes a statistical method for the near-day load state using the day-of-day type, and gives a probability estimate of the load state to be predicted.

First, the number of statistical dates m in the recent days and the number of statistical dates d on the same type date are selected, and the number of statistical dates is $m + d$ days. If the forecasting date is F, then the number of days before the m day, that is, $F-1, F-2, ..., F-m$, is counted as the relevant recent date. In the period of 7, the same number of days of the same type are successively forward, that is, $F-7, F-2\times7, ... F-d\times7$.

The residential load state of the above $m + d$ day t period is composed of a historical state vector $S_H = [S_{F-1,t} \ S_{F-2,t} \ ... \ S_{F-m,t} \ S_{F-7,t} \ S_{F-2\times7,t} \ S_{F-d\times7,t}]$. There are many possible values for this historical state vector, and the conditional probability that $S_{F,t}$ takes 0 and 1 for each value is counted. If the conditional probability that $S_{F,t}$ takes 0 is greater than the conditional probability of taking 1, then $S_{F,t}$ is taken as 0, and vice versa.

4.2. Forecasting model of the fixed input structure

As shown in Fig. 3, the forecasting model of the fixed input structure has a fixed time-delay relationship between the input related factors and the output. For example, if the occurrence time of the output amount y is t, the difference $\Delta t_i = t - t_i$ between the occurrence times t_i and t_0 of the respective input quantities x_i is a fixed value τ, as shown in (4)
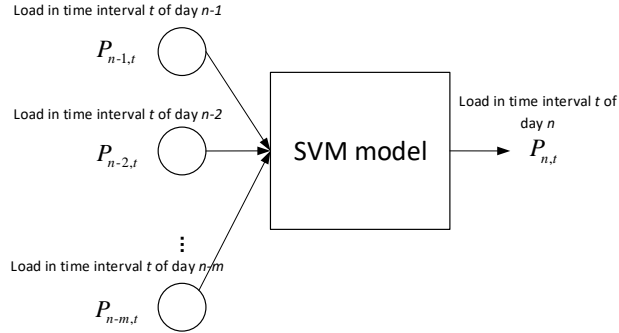
$$\Delta t_i \equiv \tau \tag{4}$$

Figure 3. SVM model of a fixed input structure

4.3. Support vector machine forecasting model for variable input structure

In the forecasting model of the variable input structure, there is no fixed time-delay relationship between the input factor and the output, that is, the difference $\Delta t_i = t - t_i$ between the occurrence time t for the output y and the occurrence time t_i for input factor x_i is variable. For the forecasting model of the variable input structure, the occurrence time t_i of the input correlation factor is not selected according to the fixed time-delay relationship, but according to their correlation relationship to determine the occurrence time t_i of the input factor, that is, for the occurrence time t_i of the input factor x_i, $y(t)$ must be most relevant to $x_i(t_i)$.

The user load forecasting model based on the variable input structure is mainly different from the user load forecasting model of the fixed input structure in the selection of historical load. According to the prediction model definition of the variable input structure, the historical load of the input is filtered according to the degree of correlation, so that it is most relevant to the daily load to be predicted, and the prediction model of the fixed input structure is selected according to the fixed time lag period, because the user load has High uncertainty, the historical load is chosen according to a fixed time lag period, and the correlation between output and input factors cannot be guaranteed. Therefore, the user load prediction model of the variable input structure is more advanced in the prediction principle than the user load prediction model of the fixed input structure. According to the above analysis, a practical user input prediction model framework with variable input structure is proposed as follows:
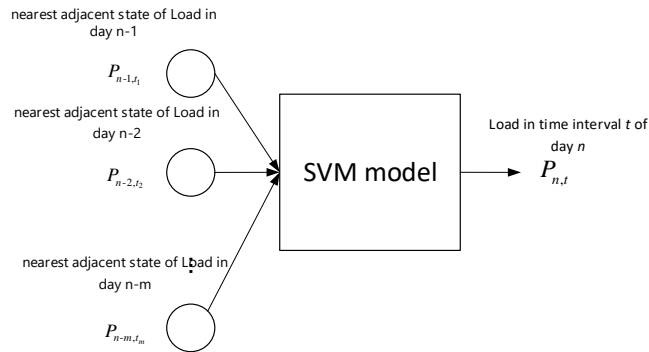

Figure 4. SVM model of variable input structure

Regarding how to find the historical load most relevant to the daily load to be predicted, the nearest neighboring state selection method is proposed here. It can be known from the "user load status statistics" that the load state of the user in each period of the day to be predicted can be obtained by statistical methods. Therefore, it is only necessary to find the historical day n and the period load t_{near} which is the same with the load state $S_{n,t}$ at the time t on the day n of the day to be predicted., so

$$t_{near} - t = \min_{S_{h,t_h} = S_{n,t}} (t_h - t) \tag{5}$$

5. Case analysis

5.1. database

The databases come from the Irish smart meter pilot data released by SEAI. The database covers a total of 4,225 users. The recording period is 536 days from July 14, 2009, to December 31, 2010. The sampling frequency is 30 minutes. 48 points of load data per day was recorded for a single user. The database is stored in double precision floating point numbers. The single-family household needs $48\times8=384$ bytes in a single day, and the total data volume is 829.32 MB.

5.2. Prediction results

Based on the processed Irish dataset, the variable input structure SVM prediction model is used to predict the short-term load of the user. The training period is from August 2, 2009, to August 1, 2010, and the forecasting period is August 2, 2010. On August 29, 2010, the forecasting method is the current popular load forecasting method SVM. The SVM prediction model of the fixed input structure and the SVM prediction model of the variable input structure are respectively constructed for the three methods, and the prediction accuracy is compared. The effect of the structured SVM prediction model of the variable input on prediction accuracy is analysed.

Figure 5 shows the average prediction error (MPPE) of the randomly selected 20 users applying the fixed input structure SVM and the variable input structure SVM model. It can be seen that the SVM model MPPE value of the variable input structure is significantly lower than the fixed input structure SVM. Table 1 shows the average prediction accuracy of 4225 users, with a fixed input structure SVM of 18.21% and a variable input structure SVM of 18.04%, which is 0.17% lower than the fixed input structure.
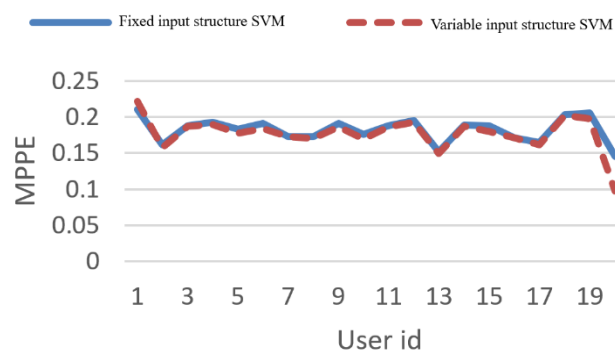

Figure 5. average prediction error of the 20 user

Figure 6 is a comparison of the prediction results of the typical user #1008 for the variable input structure SVM and the fixed input structure SVM. The predicted date is August 2nd. The solid line in the figure is the actual load, the dotted line is the fixed input structure SVM prediction result, and the broken line indicates the predicted result of the variable input structure SVM. It can be seen that the variable input structure SVM and the fixed input structure have similar prediction performances at the valley load, and the difference between the two is not large, but in the peak load period, the prediction result of the variable input structure SVM model is closer to the actual load. The peak load of the user occurred at 15:00 in the afternoon, the peak load level was 3.137 kW, the predicted result of the variable input structure SVM model was 2.328 kW, and the predicted result of the fixed input structure SVM model was only 0.223 kW, which is far apart from the actual peak load. This is because the variable input structure SVM model, by selecting the same state load as input, avoids the interference caused by the load input of different states into the model, and enhances the correlation between the model input and the predicted amount, thereby effectively improving the prediction accuracy.

Table 1. Comparison of prediction accuracy of 4225 households

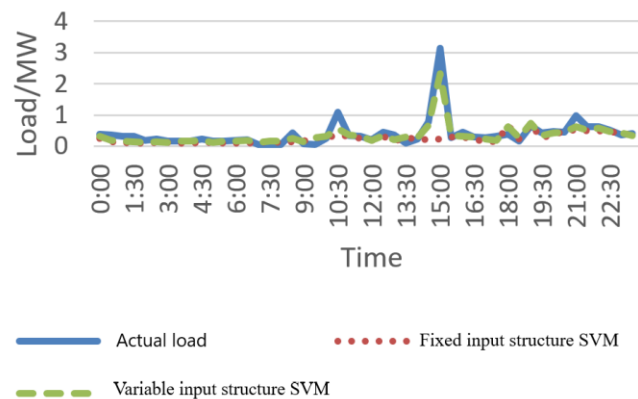| Forecasting method	Average Prediction Error（MPPE）
Fixed input structure SVM	18.21%
Variable input structure SVM	18.04%

Figure 6. the prediction results of the typical user #1008 (August 2)

6. conclusion

This paper proposes a variable input structure SVM prediction model based on state analysis. By identifying the key load characteristics in the load data and storing the load characteristics instead of the original load data, the accurate and efficient processing of the user load data is realized. The variable input structure SVM model based on the state analysed searches the same state load in similar period according to the state prediction result of each period of the forecast day and takes the same state load of the historical period as a model input factor to predict, effectively overcoming the interference caused by the user's power drift effect to load forecasting, also effectively improves the prediction accuracy.

The results of the example show that the method of this paper for the 4225 user load data in the Irish smart meter test has the advantages of simplicity, high efficiency, and remarkable effect. The average reconstruction error is only 5.57% of the daily peak load, and the average prediction accuracy for 28 days in succession is increased by 0.17%.

References
[1] Unterweger A, Engel D. (2015)Resumable Load Data Compression in Smart Grids. IEEE Transactions on Smart Grid, vol.6(2): 919-929.
[2] Pitt B D, Kitschen D S. (1999)Application of data mining techniques to load profiling, Proceedings of the 21st 1999 IEEE International Conference, vol.6:131-136.
[3] Minghao P, Ho S S, Jong Y L, et al. (2014)Subspace Projection Method Based Clustering Analysis in Load Profiling. IEEE Transactions on Power Systems, vol.29(6): 2628-2635.
[4] Luo C, Ukil A. (2015)Modeling and Validation of Electrical Load Profiling in Residential Buildings in Singapore. IEEE Transactions on Power Systems, vol.30(5): 2800-2809.
[5] Stephen B, Mutanen A J, Galloway S, et al. (2014)Enhanced Load Profiling for Residential Network Customers. IEEE Transactions on Power Delivery, vol.29(1): 88-96.

[6] Chicco G, Akilimali J S. (2010)Renyi entropy-based classification of daily electrical load patterns. IET Generation, Transmission & Distribution, vol.4(6): 736-745.

[7] Tsekouras G J, Hatziargyriou N D, Dialynas E N. (2007)Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers. IEEE Transactions on Power Systems, vol.22(3): 1120-1128.

[8] Chicco G, Ionel O M, Porumb R. (2013)Electrical Load Pattern Grouping Based on Centroid Model With Ant Colony Clustering. IEEE Transactions on Power Systems, vol.28(2): 1706-1715.

[9] Tiefeng Z, Guangquan Z, Jie L, et al. (2012)A New Index and Classification Approach for Load Pattern Analysis of Large Electricity Customers. IEEE Transactions on Power Systems, vol.27(1): 153-160.

[10] Espinoza M, Joye C, Belmans R, et al. (2005)Short-Term Load Forecasting, Profile Identification, and Customer Segmentation: A Methodology Based on Periodic Time Series. IEEE Transactions on Power Systems, vol.20(3): 1622-1630.

[11] Safdarian A, Fotuhi-Firuzabad M, Lehtonen M. A (2014)Distributed Algorithm for Managing Residential Demand Response in Smart Grids. IEEE Transactions on Industrial Informatics, vol.10(4): 2385-2393.

[12] Chen C, Jianhui W, Kishore S. A (2014)Distributed Direct Load Control Approach for Large-Scale Residential Demand Response. IEEE Transactions on Power Systems, vol.29(5): 2219-2228.

[13] Federal Energy Regulatory Commission. [2016-03-26]Assessment of Demand Response and Advanced Metering [EB/OL]. http://www.ferc.gov/legal/staff-reports/12-20-12-demand-response.pdf.

[14] Commission for Energy Regulation.[2016-03-26]Electricity Smart Metering Customer Behaviour Trials (CBT) Findings Report[EB/OL]. http://www.cer.ie/docs/000340/cer11080(a)(i).pdf.

[15] Prescott P, Walden A T. (1983)Maximum likeiihood estimation of the parameters of the three-parameter generalized extreme-value distribution from censored samples. Journal of Statistical Computation and Simulation, 16(3-4): 241-250.