Dynamic Time Warping and FFT: A Data Preprocessing Method for Electrical Load Forecasting

Juan Huo
School of Electrical and Automation Engineering
Zhengzhou University, Henan Province, China

***Abstract*—For power suppliers, an important task is to accurately predict the short-term load. Thus many papers have introduced different kinds of artificial intelligent models to improve the prediction accuracy. In recent years, Random Forest Regression (RFR) and Support Vector Machine (SVM) are widely used for this purpose. However, they can not perform well when the sample data set is too noisy or with too few pattern feature. It is usually difficult to tell whether a regression algorithm can accurately predict the future load from the historical data set before trials. Here we demonstrate a method which estimates the similarity between time series by Dynamic Time Warping (DTW) combined with Fast Fourier Transform (FFT). Results show this is a simple and fast method to filter the raw large electrical load data set and improve the learning result before looping through all learning processes.**

Keywords*—Load forecast; Dynamic Time Warping (DTW); Fast Fourier Transform (FFT); random forest; Support Vector Machine (SVM)

I. Introduction

In electrical engineering, load forecasting speculates and predicts the future power load demand for a certain period of time from historical load data. The accuracy of load forecasts has important effect on power system operations. For power management system, the Day-ahead scheduling process consists of the following principal functions: (1) assemble and update Day-ahead transmission outages; (2) produce Day-ahead zonal load forecast; (3) tabulate and evaluate non-firm transactions; (4) perform automated mitigation of generator offers.

Historical load data is important to set up prediction model and the training features. Most of the research for the prediction methods focus on the forecast methods while did not mention too much about the data preparation process [1], [2]. In most study cases for day ahead load forecast, data of the adjacent days have been selected manually as the training data source. As well known, data pre-processing has significant impact on predictive accuracy, even for some data mining techniques which can balance error in class population of unbalanced datasets [3]. Thus one new method which combines the estimation of DTW and FFT is introduced to act as a reference for raw data pre-processing and feature selection for electrical load data. The electrical load data source is evaluated in both time and frequency domain by Dynamic Time Warping (DTW) and Fast Fourier Transform (FFT) before training. DTW and FFT are supposed to help feature reselection and data re-sampling for data pre-processing purpose. Both DTW and FFT have been widely used to identify the similarity and patterns between two data sets. In the following sections, the function of DTW&FFT for time series similarity and pattern recognition will be tested.

For the purpose of electrical load forecast, we have used Random Forest and Support Vector Machine (SVM) which are popular methods for load forecast in recent years [4]–[6]. In 2001, EUNITE network organized a world wide competition on the daily electrical load prediction problem. In this event, SVM (support vector machine) or SVR (support vector regression) surpassed the other algorithms and claimed the throne for daily electrical load forecast [7], [8]. Some recent papers have found Random Forest Regression (RFR) can also perform well for time series prediction task, sometimes it can even excel SVM for some data sets [4], [5], [9], [10]. But some other papers reserves this opinion and points out they can only be compared when parameters are fixed [6].

The data source used in this paper for test is NYISO (details see Section III-A), which is rich with more than 15 years' historical record for New York Area. For the purpose of electrical load prediction, our initial forecast result is not good. With the analysis from DTW and FFT, the reason is explained. According to the analysis, redundant data sets are filtered out and the new feature is added which results in improvement for the downtown zone (N.Y.C.) of the New York. Another analysis of DTW and FFT for suburb zone also explains why data of suburb zone (North zone) is not suitable for RFR and SVM regression and can be a reference for the other training.

This paper is organized as follows. Section I provides the background knowledge of this work; Section II has the main algorithms demonstrated; Section III introduces the features of the data source; Section IV compares the time efficiency between the traditional and the new algorithm; Section V is the conclusion of this paper.

II. Method and Algorithm

We analyzed the electrical load data by combing DTW and FFT. Besides DTW and FFT. Cross correlation has also been considered once, however it failed to identify the difference between each year's difference for NYISO North Zone data set, thus it is not used as our evaluation reference in this paper. The result of DTW for similarity reference is a distance value $D(U, V)$. For FFT analysis, the resulting common frequency components (frequency with maximum power spectrum amplitudes) are the reference parameters.

A. Dynamic Time Warping (DTW)

Dynamic time warping (DTW) is an algorithm for time series analysis, it has been used for measuring similarity between two temporal sequences which may vary in time of speed. The essence of DTW is to estimate the alignment between two time series. To align two time series, U and V, an n-by-m matrix X is constructed. The (ith,jth) element of the matrix X_{ij} contains the distance $d(u_i, v_j)$ between the two points u_i and v_j. The Euclidean distance is typically used, which corresponds to the alignment between the points u_i and v_j. A warping path, W which is a set of matrix elements that defines a mapping between U and V [11]. Its kth element is defined as

$$w_k = (i_k, j_k) \tag{1}$$

and the warping path W is

$$W = w_1, w_2, ..., w_k, ..., w_K \tag{2}$$

where $max(m, n) \leq K < m + n - 1$

The warping path W is minimized and typically subjected to some constraints such as boundary conditions,continuity and monotonicity. The warping cost can be estimated by different algorithms, the most used one is a recurrence equation that defines the cumulative distance as the distance in the current cell and the minimum of the cumulative distances of the neighbouring elements. Thus the distance between two points is minimized, which can be expressed as:

$$D(U, V) = \min_{W} \left[\sum_{K=1}^{K} d(W_K) \right] \tag{3}$$

where $D(U, V)$ is the estimated distance between two time series U and V. It is an important reference in time domain. DTW has been widely used in different areas to find the matched subsequences between two time series. The speed of DTW has been improved dramatically with different kind of methods and can deal with trillions of time series in a short time [11]–[13].

B. Fast Fourier Transform (FFT)

For the frequency domain, we analyze the amplitude spectrum of FFT (Fast Fourier Transform) to find the common frequency components of two time series. The FFT is a kind of discrete Fourier transform algorithm which reveals periodicities in input data as well as the relative strengths of any periodic components. The input data is decomposed into smaller frequency complex components. By this way, it is more convenient to find the similarity pattern in frequency domain.

C. Training Algorithm and Process

The artificial intelligent model we have used for training purpose is the Random Forest Regression (RFR) and Support Vector Regression (SVR). We have used R package libraries for the implementation of these algorithms [14]–[16]. After several trials of cross validation between different years, we find the default parameters of R package is generally ok for SVM. The main parameter of RFR are the number of trees $ntree$ and the number of variables to partition at each tree node $mtry$, which do not have remarkable impact on the resulting accuracy according to the investigation of previous papers [6]. The tree number we have chosen for RFR is 1000 and variable number is 10 which are good enough to get satisfactory result.

III. Data Profile

The data set is a public electrical load forecasting database, New York Independent System Operator (NYISO)[1]. This data source recorded real-time load demand in every five minutes measured in MW. For comparison, we have studied two zones: N.Y.C. and North. The zonal forecast models use weather information of every day gathered from stations across of New York. Each hour's load was averaged for day ahead load prediction task. Fig. 1 shows the hourly load profile of N.Y.C. while Fig. 5 is the load of North zone. As N.Y.C. is the central part of New York, the load demand is obviously different from the North Zone, where the population density is low (the population for N.Y.C. is about 8 million, while the population for the North zone is only 82 thousand).

A. N.Y.C. Zone

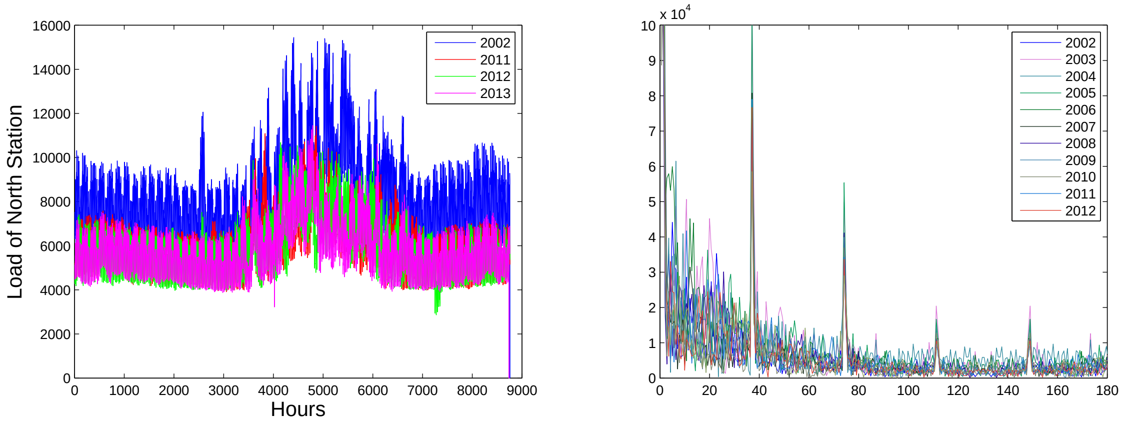
The load demand of a year varies regularly with season in N.Y.C. As in Fig. 1(a), a summer day is obviously superior to the other days each year. Fig. 1(b) is our analysis of all years' load with FFT, the amplitude of frequency spectrum of the recent years are highly overlapped over the main bands.

For load prediction task, we use year 2013 as the test target to be predicted. Before the forecast process, the load of every month from 2002 to 2012 is compared with the corresponding month's load of year 2013. The DTW distance D_{ij} ($i = 1 \ldots 12$, $j = 2002 \ldots 2012$), is calculated for every paired month. Thereafter, D_{ij} is divided by the average load of target month's adjacent days, which results in ND_{ij}. Anova analysis of ND_{ij} is shown in the two figures of Fig. 2, which are grouped by dimension month i and year j respectively. Fig. 2 shows the load of year 2002 to 2005 differs from the recent years 2006 to 2012 apparently, the average DTW distance of year 2002 to 2005 is nearly two times of the other years. Fig. 2(b) shows the different distance varies with month. Although the deviation of each month group is high, the summer period ($i = 6 \ldots 9$) has larger distance than the other months. This is consistent with our normal observation that the load variation of summer period is more uncertain than the other months.

B. RFR and SVM Prediction

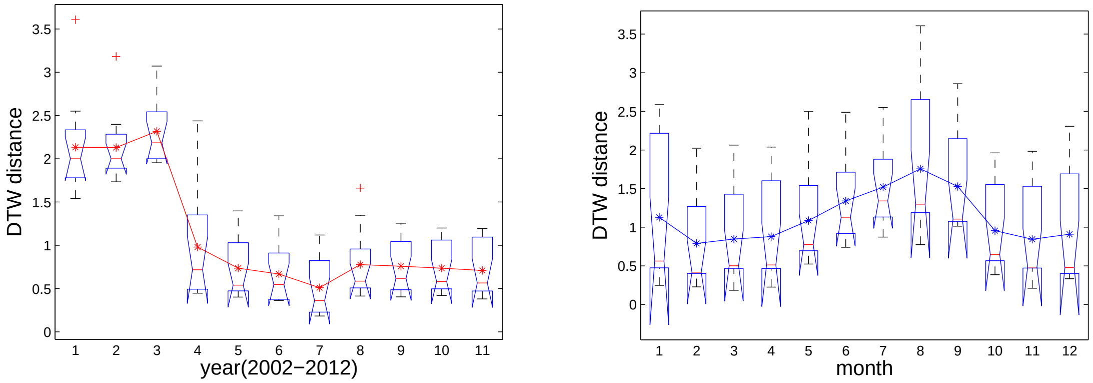
After analysis, the above data were then put into our training system of RFR and SVM to validate our hypothesis that there is noise or outlier values in the data set for regression purpose when month $i = 6 \ldots 9$ or year $j = 2002 \ldots 2005$. The features of the input vectors are initially set in Table I. The training result is evaluated as by a most common used parameter for electrical load forecast measurement, which is named as MAPE(mean absolute percentage error) [6]. Normally, MAPE calculates the average error of one day 24h. The formula of MAPE is shown in equation 4,where X_i is the predicted value

[1] http://www.nyiso.com



(a) (b)

Fig. 1: N.Y.C. load profile. (a) The hourly load profile by year. (b) FFT analysis and amplitude spectrum of different years' load.



(a) (b)

Fig. 2: Anova analysis of DTW distance $ND_{ij} = D_{ij}/AVG(L_{2013})$. The time series of each month from year 2001 to 2012 is paired with the corresponding month of test year 2013.

and R_i is the real electrical load data on the i_{th} day or hour of the prediction period (day or hour).

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \frac{|X_i - R_i|}{R_i} \tag{4}$$

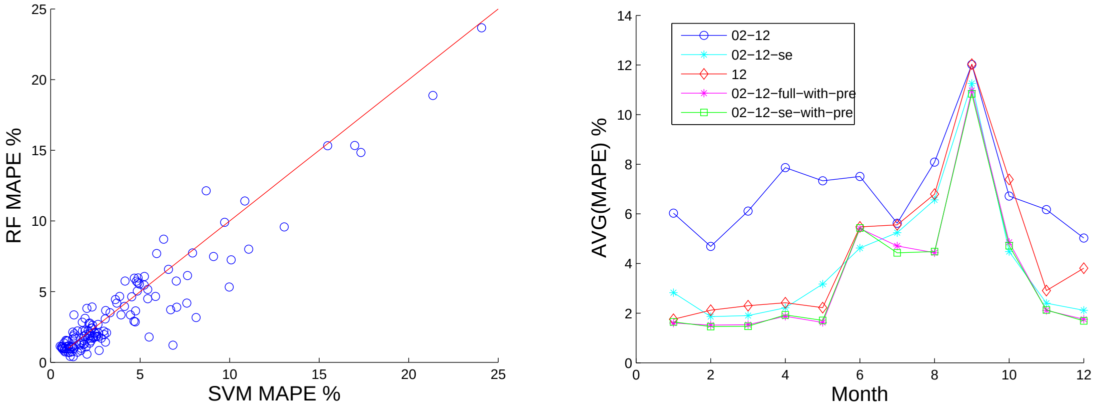
The training process is as follows: for every month of 2013, one day of the first week is randomly selected, as the first day of the test set to be predicted. The hourly load of that day to be predicted is labeled as S_{ik} $(i = 1 \ldots 12, k = 1 \ldots 24)$, in which i represents the month sequence and k represents the hour in the range of 0 to 23. All days before S_i are in the training set. The load 11 days after S_i is the test data to be predicted. The data set is grouped by hour k. Data of each hour group is trained for the corresponding hour to be predicted. For example, if hour 23:00 of March 1st 2013 is to be predicted. The load data at 23:00 is sampled from everyday before March 1st 2013. Features of table I are also collected from that day. Since each train results in 11 days' predicted value, at last we get a MAPE matrix with size 11 × 12.

We did a comparison of RFR and SVM by using 132 pairs of MAPE (all days of year 2013). T-test of these pairs proves the hypothesis that the difference between SVM and RFR prediction comes from a normal distribution with mean equal to zero and $p < 0.01$. The scatter plot of SVM vs RFR is shown in Fig. 3(a), which indicates in our load forecast task the difference between SVM and RFR is not significant.

With the initial five features in Table I, the first train has used all years' data $(2002 - 2012)$. The prediction error is very high. The average MAPE is more than 5% every month and is shown as the blue circle line labeled with the set range "02-12" in Fig. 3(a). Then we select the data of the year 2006. . . 2012 for $ND_j <$ AVG(ND) measured by DTW. The average MAPE of this selected data group "02-12-se" is

TABLE I: Input Features

ID	Features	Range	Note
1	**Month number**	$1 \ldots 12$	
2	**Weekday**	$1 \ldots 7$	
3	**Holiday**	binary	1 is holiday and weekend, 0 workday
4	**Minimum temprature**	$15.8 \ldots 102.2^{\circ} F$	
5	**Maximum temprature**	$1.4 \ldots 84.2^{\circ} F$	
6	***Load before 24h**	$0 \ldots 15503.68$ MW	not used in the initial train



(a)

(b)

Fig. 3: MAPE comparison for RFR and SVM. (a) MAPE comparison between RFR and SVM. (b) Average MAPE of RFR with different years' data.

shown as the cyan star line in Fig. 3(b). The red diamond line "12", which only has the data of year 2012, has mean MAPE($= 4.5652$). Although, the mean MAPE($= 4.0556$) of this cyan star line "02-12-se" is only slightly lower than "12", it is much better than the mean MAPE($= 6.9319$) of "02-12". This is consistent with our observation of DTW test results in Fig. 2. The time domain difference between the early years' data $j = 2002 \ldots 2005$ and the recent years' data lowered the prediction accuracy for the recent year. On the other hand, in section III-A, we have observed that although there is large deviation between the early years' data and the recent years' data on amplitude, they share the same main frequency bands. This means there is still similar pattern during their variation. Therefore in a new train, a new feature "Load before 24h" shown in Table I was added for training.

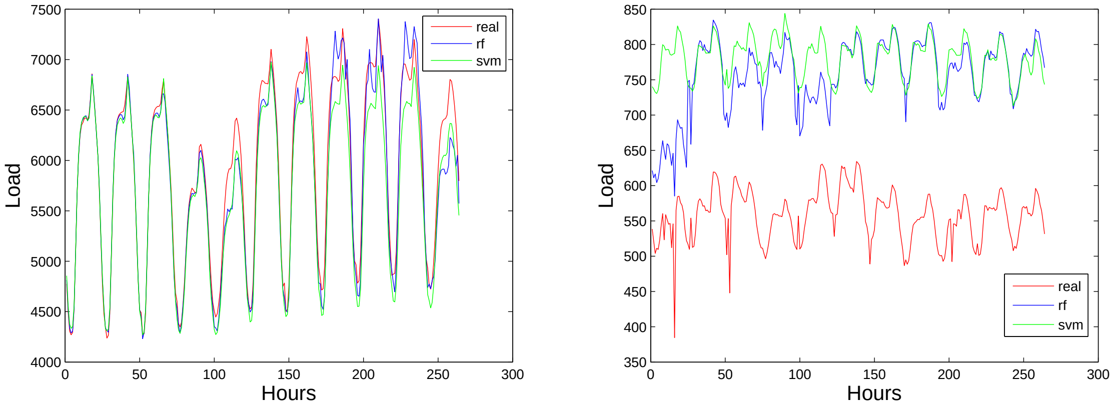
With previous day's load added, the prediction result of RFR is improved, shown as green "02-12-se-with-pre" (DTW selected group) and "02-12-full-with-pre" (all data sets) in Fig. 4(a). There is no significant difference between the results of these two groups, as RFR has classified the early years' data automatically according to the new added feature "Load before 24". The mean MAPE of "02-12-se-with-pre" (=3.4944) and "02-12-full-with-pre" (=3.5336) are both lower than any training without the new feature. The DTW selected group "02-12-se-with-pre" still performs slightly better. For all the training result, summer time is always the worst prediction period $i = 6 \ldots 9$, which is also coherent with our DTW distance hypothesis in Section III-A.

C. North Zone Data

We then use the similar method to test the North Zone's dataset. With fewer population, the load variation is irregular and has more uncertain factors shown in Fig. 5(a). We once tried cross correlation to analyze the difference between different years. However the correlation coefficient value is always above 0.95 between years, just like the N.Y.C. This does not provide much useful information. Whereas the DTW distance analysis show the DTW distance of North Zone is all more than five times of the average load, $ND_{ij} > 5$. This indicates the deviation between the north zone's data is very large. The FFT analysis in Fig. 5 also shows few coherence in frequency domain from year 2011 to 2014. We then have a trial to use the Random Forest and SVM directly to predict the day ahead hourly load of 2014. All the six features with "*Load of last 24h" in Table I were used. Fig. 6 shows the monthly average MAPE is very high and even above 50%. Thus this proves north zone's data is not suitable for prediction with regression methods.

IV. Time Cost

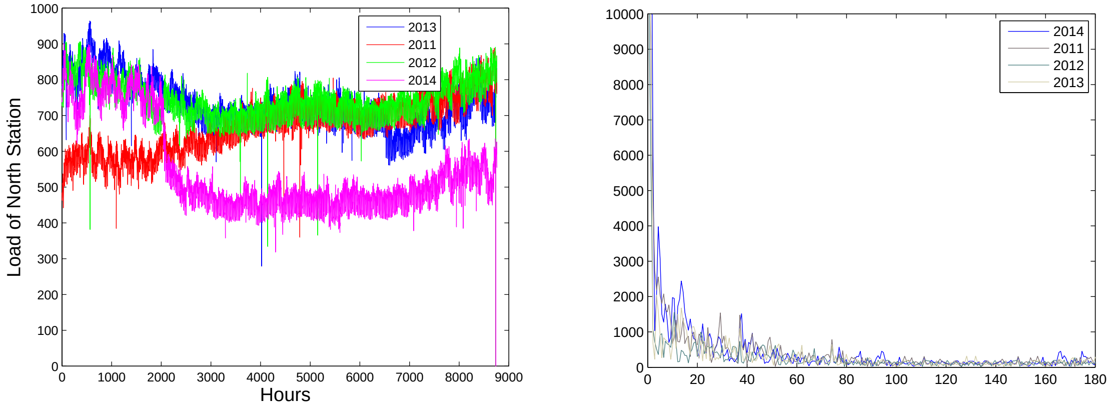
We did test to evaluate the time cost of DTW&FFT by tic-toc function in matlab. The computer has $3.4GHz$ CPU and $8GHz$ RAM. The total time cost of DTW&FFT to compare the year 2011,2012 with 2013 separately for every month is 0.2243 second. When we use data sets of 2011 and 2012 to predict the load trend of 11 days of 2013 in one trial, the average cost for SVM and RFR varies. The time cost for SVM can be as small as 0.046 second. But to have a global



(a)

(b)

Fig. 4: Load prediction of one train (December, 2013). The red line is the real load, “RF” and “SVM” represents RFR and SVM respectively. (a) The N.Y.C. zone. (b) The north zone.



(a)

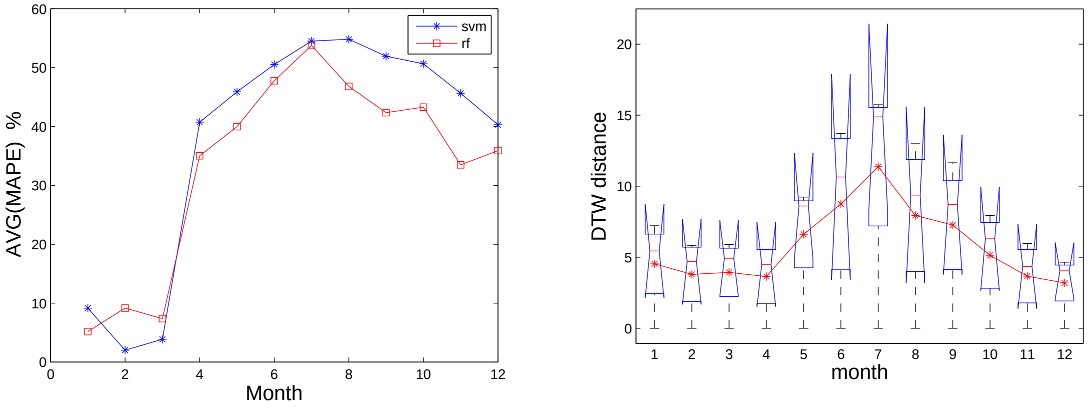
(b)

Fig. 5: North Zone. (a) The load profile. (b) FFT amplitude spectrum of north zone.

estimation of a whole year and find the special period like the summer time, we need to organize new prediction for different period multiple times and make the algorithm much more complex than DTW&FFT. This is the same for RFR, for which the situation can be worse. Because the time cost of RFR strongly depends on ntree and mtry, when $ntree = 10$ and $mtry = 2$, the time cost is less than 0.008 second, however, when ntree increases, such as $ntree = 1000, mtry = 10$, the time cost is 1.1 second on average for one trial. Generally speaking, DTW&FFT is better for global observation of the difference between time series. Some times, when only FFT is implied, it is enough to warn the low performance of regression with time cost only 0.045 second, such as the North zone data.

V. Conclusion

The above results have shown the method which combines DTW and FFT together can help to evaluate the data set for re-sampling and feature selection. Data set group selected by DTW and FFT performs better than group which has not been preprocessed. Especially for electrical load data with too few pattern features, DTW&FFT not only can identify the bad data set for prediction, but also can analyze the reason for potential prediction failure in both time and frequency domain. In addition, the computation of DTW and FFT is simpler than SVM and RFR learning process and thus is time saving compared to loop through all data sets and try different predictors. DTW&FFT algorithm has advantage to view the global features of electrical load forecast time series. Such algorithm composition can help to analyze the quality and property of the electrical load time series and should be treated as an important reference for electrical load data preprocessing.



(a)

(b)

Fig. 6: (a) RFR and SVM prediction monthly average MAPE for North Zone. (b) Anova analysis of DTW distance $ND_{ij} = D_{ij}/AVG(L_{2014})$. The time series of each month from year 2011 to 2013 is paired with the corresponding month of 2014.

References

[1] H. K. Alfares and M. Nazeeruddin, “Electric load forecasting: literature survey and classification of methods,” *International Journal of Systems Science*, vol. 33, no. 1, pp. 23–34, 2002.

[2] H. S. Hippert, C. E. Pedreira, and R. C. Souza, “Neural networks for short-term load forecasting: A review and evaluation,” *IEEE Transactions on Power Systems*, vol. 16, no. 1, pp. 44–55, 2001.

[3] S. F. Crone, S. Lessmann, and R. Stahlbock, “The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing,” *European Journal of Operational Research*, vol. 173, no. 3, pp. 781–800, 2006.

[4] Y. Chen, S. Guo, H. Chen, L. I. Wanhua, K. Guo, and Q. Huang, “Electricity customers arrears alert based on parallel classification algorithm,” 2016.

[5] X. Wu, J. He, P. Zhang, and J. Hu, “Power system short-term load forecasting based on improved random forest with grey relation projection,” *Dianli Xitong Zidonghua/automation of Electric Power Systems*, vol. 39, no. 12, pp. 50–55, 2015.

[6] A. Lahouar and J. Ben Hadj Slama, “Day-ahead load forecast using random forest and expert input selection,” *Energy Conversion and Management*, vol. 103, pp. 1040–1051, 2015.

[7] 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, November 2004.

[8] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[9] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[10] Z. Y.-w. M. B.-h. J. Huan, “Research of medium and long term precipitation forecasting model based on random forest,” *Water Resources and Power*, vol. 33, no. 6, pp. 6–10, 2015.

[11] T. C. Fu, “A review on time series data mining,” *Engineering Applications of Artificial Intelligence*, vol. 24, no. 1, pp. 164–181, 2011, fu, Tak-chung.

[12] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, “Searching and mining trillions of time series subsequences under dynamic time warping,” in *18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2012, August 12, 2012 - August 16, 2012*, ser. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, pp. 262–270.

[13] G. T, “Computing and visualizing dynamic time warping alignments in r: The dtw package.” *Journal of Statistical Software*, vol. 31, no. 7, pp. 1–24, 2009.

[14] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: https://www.R-project.org/

[15] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2015, r package version 1.6-7. [Online]. Available: https://CRAN.R-project.org/package=e1071

[16] A. Liaw and M. Wiener, “Classification and regression by randomforest,” *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: http://CRAN.R-project.org/doc/Rnews/