A multi-task learning framework for multi-location short-term load prediction

Zhaorui Meng *, Jinhua Sun

School of computer and information engineering, Xiamen University of Technology, Xiamen, Fujian, 361024, China

*Corresponding author e-mail: mengzhaorui@xmut.edu.cn

Abstract. Short-term power load prediction is crucial to management of power system. The traditional load prediction methods are based on learning data from single location. However, load consumption is related among different locations. In this paper, we propose a novel approach to train load predictors for multi-locations in a collaborative way based on multi-task learning. Specifically, the load predictor in each location is split into two parts: a general one and a location-specific one. The general predictor is to capture information shared by various locations. And the location-specific predictor is to obtain location-specific information. In addition, a location similarity graph is built and incorporated into the model as regularization. Extensive experiment result shows that our approach outperformed single task learning method and another two multi-task learning methods in at least 80% of locations.

1. Introduction

Power load forecasting is crucial to management of power system. Short-term load prediction is important to day-to-day operation, scheduling and load-shedding plans of power utility. Advanced predicting technology is highly required to improve the accuracy of prediction of electric demand.

As development of modern technology and popularity of smart grid, theory and technology of load prediction have gained dramatic development. In recent years, there are variety approaches have been adopted in the literature to model load prediction. Classic time series analysis has been applied to this field, such as Auto-Regressive Moving Average (ARMA) [1], Holt-Winters smoothing [2] and Seasonal-adjusted Auto-Regressive Integrated Moving Average (SARIMA) [3]. In the past decades, as the fast development of artificial intelligence and machine learning technology, more and more machine learning models have been adopted to solve the problem, including support vector regression (SVR) [4], neural networks (ANN) [5] and kernel-based support vector machine (SVM) [6].

Most of research articles on load forecasting is based on single time series representing data from single task only. As latest development in transfer learning [7], researchers realized that learning performance can be improved by incorporating data from other related tasks. Transfer learning approaches have been applied successfully in many research fields, such as natural language processing [8], computer vision [9] and voice recognition [10]. In power load prediction, several researches adopting transfer learning have been published recently. In the transfer learning method proposed in [11], a source task selection algorithm is developed and the transfer learning model based on Gaussian process is constructed. Therefore, Prediction errors of the target tasks can be greatly reduced by utilizing...
the knowledge transferred from the source tasks. In [12], the authors explored kernel-based multi-task learning techniques to forecast the demand of electricity measured on multiple lines of a distribution network. In [13], a multi-task Gaussian process method for nonstationary time series prediction is introduced and applied to the power load forecasting problem.

In approaches listed above, data from source domain are transferred successfully into target domain. As a result, accuracy of prediction has been improved. However, prior knowledge among different locations are not considered during training process. If prior location similarity knowledge can be incorporated into learning model, the accuracy of prediction can be improved.

Motivated by above observation, a novel multi-location power load prediction approach is proposed. In our approach, the power load prediction is splitted into two parts: a general load predictor and a location-specific load predictor. The general load predictor shared by all locations is trained by data from all locations to get better generalization ability. While the location-specific load predictor is trained by data from single location only. In order to utilize prior load similarity knowledge among different locations, a load similarity graph is built and incorporated into the prediction model. An optimization method based FISTA is introduced to solve the convex optimization problem of our model. The performance of our approach is evaluated against several baseline methods with data collected from 15 different locations. The experiments show our approach outperformed traditional machine learning method and another two multi-task learning methods in at least 80% of locations.

The rest of paper is organized as follows. In section 2 we present our multi-location power load prediction approach in detail. And finally, the experiments are given in section 3.

2. Proposed Framework

In this section, our proposed framework for multi-location short-term load forecasting is presented in detail. In addition, the optimization algorithm based on FISTA is introduced.

2.1. Preliminaries

Let \( M \subset \mathbb{R}^2 \) be a set of geo-referenced locations, where each location \( m \in M \) is associated with a set of features. One of the fields is the electric load to be predicted, while the other fields are predicting variables, for instance: temperature, wind and load history etc.

Denote \( \{X^i \in \mathbb{R}^{N_i \times D}, y^i \in \mathbb{R}^{N_i \times 1}\} \) as the samples in location \( i \), where \( N_i \) is the number of samples and \( D \) is the feature size. \( x^i_j \) is the transpose of the \( j \)th row of \( X^i \), represents the feature vector of the \( j \)th sample in location \( i \) and \( y^i_j \) is its electric load value.

Denotes the base model shared by all the locations as \( w \in \mathbb{R}^{D \times 1} \), and the location-specific predictor of location \( i \) as \( w_i \in \mathbb{R}^{D \times 1} \). Matrix \( W \in \mathbb{R}^{D \times M} \) is used to represent all the location-specific electric load predictors, where the \( i_{th} \) column of \( W \), i.e., \( W_i = w_i \), represents for the electric load forecasting model of location \( i \). Denote the loss function of predicting under parameters \( w \) and \( W \) as \( f(x^i_j, y^i_j, w + W_i) \).

Vector \( S \in \mathbb{R}^{M \times M} \) is used to represent the prior location similarity knowledge. \( S_{i,j} \in [0,1] \) is the similarity between location \( i \) and \( j \), measured by time series similarity desribed in section 2.4. The diagonal elements of \( S \) are set to 1.

2.2. Models

The following model is used in this paper.

\[
\arg \min_{w,W} L(w, W) = \sum_{i=1}^{M} \sum_{j=1}^{N_i} f(x^i_j, y^i_j, w + W_i) + \alpha \sum_{i=1}^{M} \sum_{j \neq i} S_{i,j} \|W_i - W_j\|_2^2 + \lambda_1 \|w\|_1 + \lambda_2 \|W\|_1
\]

(1)
Where $\alpha$, $\lambda_1$ and $\lambda_2$ are non-negative regularization coefficients for prior location similarity knowledge and model parameters respectively. L1 regulation penalty are added to ensure that each base model depends only on a small subset of the predictor variables.

From the model in Eq. (1), the final load predictor of each location is the combination of the general load predictor model and location-specific load predictor model. The general load predictor model is shared by all the locations and trained using the sample data in all locations. The location-specific load predictor model is trained using the sample data with the location. By splitting the load predictor of each domain into two parts, the global power load knowledge shared by different locations can be better utilized.

From Eq. (1), location similarity information is incorporated into objective function as regularization. Minimizing $\sum_{i=1}^{M} \sum_{j \neq i} S_{ij} \| W_{ij} - W_{i,j} \|_2^2$ means that two locations share high similarity, load predictors of two locations should be more similar.

### 2.3. Optimization Method

Because of L1 regulation penalty, optimization problem in Eq. (1) is non-smooth. Therefore, in this paper, an accelerated algorithm based on FISTA [14] is introduced to solve the optimization problem in Eq. (1).

In each iteration of FISTA, there are two kinds of points updated sequentially. At first, define the search points as the linear combination of last two solutions:

$$v^{k+1} = w^k + \alpha_k (w^k - w^{k-1}),$$

$$v^{k+1} = w^k + \alpha_k (W^k - W^{k-1}),$$

Then the approximate points are updated as below:

$$W^{k+1} = T_{\lambda_2/l_k} \left( v^{k+1} - \frac{1}{l_k} \frac{\partial g(w,W)}{\partial w} |_{W = v^{k+1}, W = V^{k+1}} \right)$$

$$W_{ij}^{k+1} = T_{\lambda_2/l_k} \left( V_{ij}^{k+1} - \frac{1}{l_k} \frac{\partial g(w,W)}{\partial W_{ij}} |_{W = v^{k+1}, W = V^{k+1}} \right)$$

where $T$ is the soft thresholding operator and is defined as $T_k(x) = [x - k]^+ - [-x - k]^+$. $\frac{1}{l_k}$ is the step size at the $k_{th}$ iteration and its value is selected to satisfy following inequation:

$$g(w^{k+1}, W^{k+1}) \leq g(v^{k+1}, v^{k+1}) + \frac{L_k}{2} \| W^{k+1} - V^{k+1} \|_2^2 + \frac{\| W^{k+1} - V^{k+1} \|_2^2}{2}.$$
Algorithm 1. The optimization algorithm of solving proposed model when loss function f is smooth

1: Input: \((x_j^i, y_j^i)\), \(i=1, \ldots, M, j=1, \ldots, N_i, S, \alpha, \lambda_1, \lambda_2, \eta > 1, L_0\).
2: Output: \(w, W\)
3: Initialize: \(w^0 = 0, W^0 = \frac{1}{L_0}\).
4: While the convergence condition is not satisfied do
5: \(k=k+1, \alpha_k = \frac{k}{k+3}\)
6: \(v^{k+1} = w^k + \alpha_k (w^k - w^{k-1})\),
7: \(V^{k+1} = W^k + \alpha_k (W^k - W^{k-1})\),
8: Compute \(\frac{\partial g(w, W)}{\partial w} \bigg|_{w=v^{k+1}, W=V^{k+1}}\) according to Eq. (5)
9: \(w^{k+1} = T_{k/2/L_k} \left( v^{k+1} - \frac{1}{L_k} \frac{\partial g(w, W)}{\partial w} \bigg|_{w=v^{k+1}, W=V^{k+1}} \right)\)
10: for \(i=1\) to \(M\) do
11: \(W_{i,i}^{k+1} = T_{k/2/L_k} \left( v_{i,i}^{k+1} - \frac{1}{L_k} \frac{\partial g(w, W)}{\partial W_{i,i}} \bigg|_{w=v^{k+1}, W=V^{k+1}} \right)\)
12: end for
13: while Eq. (9) doesn’t hold do
14: \(L = \eta L\).
15: \(w^{k+1} = T_{k/2/L_k} \left( v^{k+1} - \frac{1}{L_k} \frac{\partial g(w, W)}{\partial w} \bigg|_{w=v^{k+1}, W=V^{k+1}} \right)\)
16: for \(i=1\) to \(M\) do
17: \(W_{i,i}^{k+1} = T_{k/2/L_k} \left( v_{i,i}^{k+1} - \frac{1}{L_k} \frac{\partial g(w, W)}{\partial W_{i,i}} \bigg|_{w=v^{k+1}, W=V^{k+1}} \right)\)
18: end for
19: end while
20: end while
21: \(w = w^{k+1}\)
22: \(W = W^{k+1}\)

2.4. Task Relationship Matrix
According to work in [15], Pearson Correlation Coefficient can be used to measure distances between two times series. Since power load is a typical time series, we use Pearson Correlation Coefficient to estimate the spatial correlation between different locations in this study. Specifically, Pearson Correlation Coefficient between two time series \(X\) and \(Y\) is given by

\[
\frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum(x_i - \bar{X})^2 \sum(y_i - \bar{Y})^2}}
\]  

(10)

The prior location similarity knowledge in Eq. (1) is computed based on the Person Correlation Coefficient for two locations:

\[
S_{i,j} = \begin{cases} 
1 & \text{if } i = j \\ 
\frac{1}{\text{sim}(i,j)} & \text{otherwise}
\end{cases}
\]  

(11)

Where sim \((i, j)\) denotes Person Correlation Coefficient between two locations.

3. Experiments
3.1. Data and pre-processing
To verify forecasting performance of the proposed multi-location short term load forecasting approach, a real-world dataset from GEFCOM2012 [16] is used for validation purpose. In the dataset, there are hourly loads (in MW) from 20 geographical zones with hourly temperatures from 11 weather stations in the United States.
The complete dataset includes data from January 1, 2004 to July 7.2008. Since the objective of this study is short-term load forecasting, we are going to use data from June 1st, 2007 to June 29th, 2008. Specifically, training data is from June 1, 2007 to May 30th, 2008. And testing data is from June 1st, 2008 to June 29th, 2008.

Before learning process, there are some preprocessing procedures needed. Based on analysis in [16], Zone 3 and 7 contain identical data. Data of Zone 2 are 92.68% of the values of Zone 3 and Zone 7. Zone 4 and Zone 9 contain erratic demand patterns which is totally different with other zones. Also, Zone 10 has a big jump in demand in year 2008. Therefore only 15 zones out of 20 demands data are included in our study.

3.2. Performance Metrics:
Forecasting performance can be evaluated for each location. The mean absolute percentage error (MAPE) is calculated to examine the forecasting accuracy, which is defined as follow:

$$\text{MAPE} = \frac{\sum_{i=1}^{n} \left| \frac{P_i - A_i}{A_i} \right|}{n} \times 100\%$$  \hspace{1cm} (12)

Where $P_i$ and $A_i$ are the $i_{th}$ predicted and actual values respectively, and $n$ is the total number of predictions. The MAPE is a widely used metric, which measures the percentage error between the actual and predicted values. The smaller the MAPE values, the closer are the predicted values to the actual values.

3.3. Selected baseline methods and performance evaluation
To evaluate the benefits of the proposed method in this paper, two traditional STL machine learning methods and another two MTL methods have been selected as counterparts for comparison purpose. These four counterparts and our proposed approach are abbreviated as follows:

1) GBDT: We applied GBDT to the data set at each zone independently. The GBDT results serve as a baseline for single-task learning.
2) Random Forest Regression: As GBDT, Random Forest Regression results serve as another baseline for single-task learning.
3) TRACE: The TRACE is an algorithm developed in [17] by solving trace norm regularized problems.
4) SRMTL: This is an MTL algorithm proposed in [18], which allows to model the relation between tasks in terms of a kernel function that uses a task–coupling parameter.

We use implementation given in MALSRL software package [19] for both TRACE and SRMTL.

All our experiments were carried out in MATLAB 2012a using a computer with Intel Core i7-7500U CPU, 2.90 GHz, and 8 GB RAM.

Results of our proposed approach compared to GBDT model are shown on Figure 1. We only display the result of Zone 1. The rest of zones are similar. Table 1 compares the MAPE for different methods when applied to the experimental data set (which has 1 year of training data and 1 month testing data). The results show that our method outperforms other baselines for most of the zones. The best method with the lowest rank is highlighted in bold. In fact, our method outperforms other baselines in at least 12(80%) out of 15 zones. Furthermore, by comparing STL methods (GBDT and Random Forest Regression) against the other two MTL methods, we observe that GBDT outperforms Trace and SRMTL in 7 (46.67%) out of 15 zones. And Random Forest Regression GBDT outperforms Trace and SRMTL in 6 (40%) out of 15 zones. It implies that our approach can take good advantage of task relatedness for multi-location power load prediction. In our approach, the power load prediction models of different locations share a general load predictor model. Therefore, our method can better leverage the common power load knowledge shared by all locations and the location-dependent load knowledge can be captured at the same time. In addition, the similarities between different locations are also incorporated into our approach to better model the different relatedness of different pairs of locations.
Table 1. The accuracies of different methods on all 15 zones (1 year of training data and 1 month testing data)

| Zone | Random Forest | GBDT | Trace | SRMTL | Our Approach |
|------|---------------|------|-------|-------|--------------|
| 1    | 10.43         | 8.96 | 8.61  | 8.01  | 19.18        |
| 2    | 14.6          | 13.34| 12.32 | 11.86 | 25.12        |
| 3    | 8.53          | 7.59 | 7.25  | 6.78  | 14.32        |
| 4    | 10.03         | 9.81 | 11.01 | 8.85  | 22.95        |
| 5    | 9.47          | 8.05 | 8.95  | 7.95  | 22.56        |
| 6    | 11.25         | 9.71 | 11.32 | 7.35  | 19.30        |
| 7    | 7.88          | 9.88 | 7.95  | 8.73  | 22.56        |
| 8    | 10.65         | 8.45 | 8.95  | 6.63  | 22.95        |
| 9    | 9.74          | 11.66| 9.87  | 8.89  | 22.56        |
| 10   | 13.35         | 7.29 | 10.85 | 7.63  | 22.56        |
| 11   | 10.03         | 9.82 | 10.95 | 9.88  | 22.56        |
| 12   | 5.9          | 10.3 | 9.88  | 9.8     | 22.56       |
| 13   | 10.03         | 4.96 | 9.8    | 22.56   | 22.56       |
| 14   | 7.66          | 8.63 | 9.87  | 9.88  | 22.56        |
| 15   | 18.82         | 7.32 | 7.23  | 6.82   | 22.56        |

Figure 1. Forecasted and actual load for June 2008, Zone 1

Next, we investigate the relative performance of our method against other methods as the training set size decreases from 1 year to 3 months. Table 2 shows the results when the training set size decreases. Comparing the result between Table 1 and Table 2, we can see that MAPE value increase dramatically as training data size decreases. The result also shows that our method outperforms the baseline methods in 13(85%) out of 15 zones, which is even higher when there are fewer training data. It confirms that our method can effectively train local models when there are limited training data.

Table 2. The accuracies of different methods on all 15 zones (3 months of training data and 1 month testing data)

| Zone | Random Forest | GBDT | Trace | SRMTL | Our Approach |
|------|---------------|------|-------|-------|--------------|
| 1    | 19.46         | 21.81| 19.33 | 19.18 | 22.88        |
| 2    | 22.95         | 22.56| 22.61 | 25.12 | 22.56        |
| 3    | 17.21         | 16.06| 15.30 | 25.12 | 22.56        |
| 4    | 16.86         | 15.93| 16.40 | 25.12 | 22.56        |
| 5    | 16.3          | 16.56| 17.50 | 25.12 | 22.56        |
| 6    | 24.98         | 24.62| 29.98 | 25.12 | 22.56        |
| 7    | 26.66         | 25.69| 29.98 | 25.12 | 22.56        |
| 8    | 22.11         | 20.95| 29.98 | 25.12 | 22.56        |
| 9    | 33.2          | 33.63| 33.21 | 29.15 | 22.56        |
| 10   | 27.87         | 26.96| 33.21 | 29.15 | 22.56        |
| 11   | 26.31         | 26.46| 33.21 | 29.15 | 22.56        |
| 12   | 20.89         | 20.14| 33.21 | 29.15 | 22.56        |
| 13   | 22.49         | 22.82| 29.15 | 29.15 | 22.56        |
| 14   | 29.3          | 22.16| 29.15 | 29.15 | 22.56        |
| 15   | 18.82         | 22.65| 29.15 | 29.15 | 22.56        |
| 16   | 20.39         | 22.65| 29.15 | 29.15 | 22.56        |
| 17   | 29.3          | 22.16| 29.15 | 29.15 | 22.56        |
| 18   | 22.88         | 18.64| 29.15 | 29.15 | 22.56        |
| 19   | 22.49         | 22.82| 29.15 | 29.15 | 22.56        |
| 20   | 29.3          | 22.16| 29.15 | 29.15 | 22.56        |

4. Conclusion

This paper presents a novel spatial-temporal multi-task learning framework for multi-location power load prediction. Our approach assumes that load predictor of each location is the combination of the
general load predictor model and location-specific load predictor model. In addition, we propose to use Person Correlation Coefficient to model the similarities between different locations. The location similarities are incorporated into our approach as regularization over the location-specific models. Besides, we introduce an algorithm to solve the optimization problem in our approach efficiently. Experimental result on a real-world power load data set show that the proposed approach outperformed other baseline algorithms.

For future work, we plan to develop a distributed version of our approach for improving its scalability and making it able to handle training data in a parallel way. Secondly, we will extend the approach to an online multi-task learning setting for better scalability.

References
[1] Berk, K., A. Hoffmann, and A. Mueller. "Probabilistic forecasting of industrial electricity load with regime switching behavior." International Journal of Forecasting 34.2(2018): 147-162.
[2] Navarro, Maricar Misola, and B. B. Navarro. "Optimal Short-Term Forecasting using GA-Based Holt-Winters Method." IEEM 2019 (IEEE) International Conference on Industrial Engineering and Engineering Management IEEE, 2019.
[3] Jesús Silva, et al. "Forecasting Electric Load Demand through Advanced Statistical Techniques." Journal of Physics: Conference Series 1432.1(2020): 012031 (7pp).
[4] Gao, Xin, et al. "Short-Term Electricity Load Forecasting Model Based on EMD-GRU with Feature Selection." Energies 12.6(2019):1140-.
[5] Pourdaryaei, Alireeze, et al. "Hybrid ANN and artificial cooperative search algorithm to forecast short-term electricity price in de-regulated electricity market." IEEE Access PP.99(2019):1-1.
[6] Shine, P., et al. "Annual electricity consumption prediction and future expansion analysis on dairy farms using a support vector machine." Applied Energy 250.PT.1(2019):1110-1119.
[7] Pan, Sinno Jialin, and Q. Yang. "A Survey on Transfer Learning." IEEE Transactions on Knowledge & Data Engineering 22.10(2010):1345-1359.
[8] A, Junchi Zhang, et al. "Multi-task and multi-view training for end-to-end relation extraction." Neurocomputing 364(2019):245-253.
[9] Ionescu, Radu Tudor, and M. Popescu. "Knowledge Transfer between Computer Vision and Text Mining." Advances in Computer Vision & Pattern Recognition (2016).
[10] Khanna, Rohan, D. Oh, and Y. Kim. "Through-Wall Remote Human Voice Recognition Using Doppler Radar With Transfer Learning." IEEE Sensors Journal (2019):1-1.
[11] Zhang, Yulai, and G. Luo. "Short term power load prediction with knowledge transfer." Information Systems 53.C(2015): 161-169.
[12] Fiot, Jean Baptiste, and F. Dinuzzo. "Electricity demand forecasting by multi-task learning." PowerTech, 2017 IEEE Manchester IEEE, 2017.
[13] Zhang, Y., G. Luo, and F. Pu. "Power load forecasting based on multi-task Gaussian process." IFAC Proceedings Volumes 47.3(2014): 3651-3656.
[14] Wu, Fangzhao, and Y. Huang. "Collaborative Multi-domain Sentiment Classification." IEEE International Conference on Data Mining IEEE, 2016:459-468.
[15] Berthold, Michael R., and F. Höppner. "On Clustering Time Series Using Euclidean Distance and Pearson Correlation." (2016). 
[16] Taieb, Souhaib Ben, and R. J. Hyndman. "A gradient boosting approach to the Kaggle load forecasting competition." International Journal of Forecasting 30.2(2014):382-394.
[17] Ji, Shuiwang, and J. Ye. "An accelerated gradient method for trace norm minimization." International Conference on Machine Learning 2009:457-464.
[18] Evgeniou, Theodoros, and M. Pontil. "Regularized multi--task learning." Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM, 2004: 109-117.
[19] Zhou, Jiayu, Jianhui Chen, and Jieping Ye. "MALSAR: Multi-task learning via structural regularization." Arizona State University (2011). http://www.MALSAR.org.