Short-term load forecasting based on big data technology

School of Electrical Engineering
Beijing Jiaotong University

Dr. Jinghan He
Short term load forecasting refers to forecasting load curve of coming several hours, one day or several days ahead. It is not only ensure power system to operate safely and economically, and also lay the foundation of formulating electricity dispatching schedule and transactions.

Short-term load forecasting accuracy has been listed as an important assessment indicator in Grid Corporation. With the development of power market, the requirement of forecasting accuracy, timeliness, reliability and intelligence become higher.
Current Situation

Object

- At present, short-term load forecasting mainly targets the total load, or a deeper level substation bus load, then obtained the total load by accumulating all the bus load.

Method

- The core of power load forecasting is algorithm. Most of the current methods can be divided into classical methods, traditional forecasting methods and intelligent forecasting methods.
The research of short-term load forecasting has a very long history. Many scholars have put forward different algorithms. The table below shows some of their advantages and weakness.

<table>
<thead>
<tr>
<th>Forecasting algorithms</th>
<th>Advantages &amp; Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical Method</td>
<td><strong>Advantages</strong>: Simple principle, Fast calculation</td>
</tr>
<tr>
<td></td>
<td><strong>Weakness</strong>: Weather and other factors are not considered</td>
</tr>
<tr>
<td>Exponential Smoothing, Time series</td>
<td></td>
</tr>
<tr>
<td>etc</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Advantages</strong>: Easy Application</td>
</tr>
<tr>
<td></td>
<td><strong>Weakness</strong>: Difficult to find Regression parameters</td>
</tr>
<tr>
<td>Traditional Method</td>
<td></td>
</tr>
<tr>
<td>Regression analysis . etc</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Advantages</strong>: Able to handle non-linear and stochastic problem</td>
</tr>
<tr>
<td></td>
<td><strong>Weakness</strong>: Model training need much of time; The existence of local minimum points, etc.</td>
</tr>
</tbody>
</table>
The existing algorithms ordinarily treat all users load as a whole namely system load. But this kind of method must ignore the individual load properties.

Load influential factors considered incomplete
- Most of industrial users are affected heavily by their production schedule. But now the majority of forecasting algorithms do not take these kind of factors into account.

Application of models is not accurate enough
- Because different users have different load properties. The system load can not grasps the law of load change.

Analytic Hierarchy is not detailed enough
- System load is constituted by many users load. Different users have pretty different load characteristics, because they have their own properties.
Outline

1. Background
2. Solutions
3. Case study
4. Conclusion
**Research object**

- **Normal approach:** System load
- **New approach:** User load

**User electricity behavior**
- Different users
- Different electricity law
- Different forecasting method

**Forecasting model**

- **Forecasting model**
- Massive user load data
- Data Mining
- Relationship between influential factors and load
- Proper model for each kind of load type
Solution Framework

Step 1:
Cluster

Step 2:
Association analysis

Step 3:
Deployment
Classify daily load curves of each user into several types using improved hierarchical cluster technique.

**Step 1:**

**Improved Hierarchical Cluster Technique**

Selection of proper distance algorithm is the key for clustering technique:

\[
 d_{12} = \sqrt{\sum_{k=1}^{n} \left( \frac{x_{1k} - x_{2k}}{x_{\text{max}}} \right)^2}
\]
For different users, critical influential factors are not the same. For example, temperature could be a major influential factor for residential loads. But temperature may have little impact on industrial loads.

Grey association analysis is applied to determine the critical influential factors for each individual load.
Based on cluster analysis and association analysis results, a decision tree is developed to establish relationship between clustering results and critical factors.

As shown in the figure:

**Input:**
- Critical influential factors determined by association analysis;
- The load patterns determined by hierarchical clustering analysis are used as leave node labels.

**Output:**
Decision tree = classification rules.
STEP 4

• For each type of load patterns, corresponding data set is chosen (cluster results) to train SVM model in order to determine SVM parameter shown in the figure below;

• Therefore, a set of SVM models are developed for each load. Once the forecasting day’s load pattern is determined according to the decision tree, the corresponding SVM forecasting model with appropriate parameters is used.

EXAMPLE

Target day: 2013/1/2
Target User: #123

<table>
<thead>
<tr>
<th>Input factor</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Temp</td>
<td>21 °C</td>
</tr>
<tr>
<td>Ave Humid</td>
<td>54.3%</td>
</tr>
<tr>
<td>Ave Temp</td>
<td>12.75 °C</td>
</tr>
<tr>
<td>Day Type</td>
<td>Saturday</td>
</tr>
</tbody>
</table>

Forecasting value 2013/1/2
Step 5:
Forecasting Total System Load

STEP 5

- The total system load is forecasted based on aggregation of individual load’s forecasting results as shown in Figure below;
- Repeat the four steps for each user, then aggregate all of them.

The forecasted total load can be calculated by adding all of the forecasted individual load together with line loss:

\[ L_{total} = L_{loss} + \sum_{i=1}^{n} L_{user(i)} \]
This area has 16 220(kV) substations. We use one distribution transformer load data of the yellow area to demonstrate our method.

**Historical load data:**
2012/1/1~2012/12/31

**Forecasting:**
2013/1/25~2013/1/31
Hierarchical cluster results

All daily curves from 1/1/2012 to 20/11/2012

Daily load curves in 2012

Cluster results of 2012

- The historical daily load curves in 2012 are classified into 6 clusters;
- Each color in the right page represents a cluster;
Hierarchical cluster results

Cluster results

All daily load curves in 2012

Cluster 1

Cluster 2

Cluster 3

Cluster 4

Cluster 5

Cluster 6
Typical day type in 6 clusters

<table>
<thead>
<tr>
<th>Day type</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>4</td>
<td>0</td>
<td>14</td>
<td>27</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>Tuesday</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>33</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Wednesday</td>
<td>2</td>
<td>2</td>
<td>34</td>
<td>18</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Thursday</td>
<td>2</td>
<td>1</td>
<td>31</td>
<td>20</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Friday</td>
<td>0</td>
<td>1</td>
<td>39</td>
<td>21</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Saturday</td>
<td>0</td>
<td>21</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Sunday</td>
<td>2</td>
<td>22</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>

The sixth cluster: Special holiday

The first cluster: The day after special holiday:
(2/1/2012), (29/1/2012), (5/4/2012), (2/5/2012),

◆ In the table, 3rd cluster and 4th cluster are mainly workday. 2nd cluster are mainly weekends, 5th cluster are mainly Monday.
◆ Further analysis, 6th cluster are mainly special holiday, and the 1st cluster are mainly the day after special holiday;
◆ Conclusion: cluster results can reflect that day type have strong influence on load curve shape.
Grey association analysis

Use daily load data of Jiute distribution substation and local weather information as inputs. Output the grey correlation of influential factors like below:

Gray Correlation of 6 influential factors:

<table>
<thead>
<tr>
<th>Max temp</th>
<th>Ave humidity</th>
<th>Ave temp</th>
<th>day type</th>
<th>wind velocity</th>
<th>Average precipitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7512</td>
<td>0.6588</td>
<td>0.5346</td>
<td>0.4472</td>
<td>0.2214</td>
<td>0.2137</td>
</tr>
</tbody>
</table>

• Sequence theses 6 factors, Excluding wind velocity and average precipitation which have weak correlation;
• Conclusion: The results can reflect the correlation between different factors and load. **These four critical influential factors are inputs of CART decision tree.**
Establishment of classification rules

- Set 6 cluster label as decision tree leaf node.
- According to the grey correlation, set four critical factors as attribute set to split.
- CART decision tree algorithm is used to build the classification rules.

<table>
<thead>
<tr>
<th>Input factor</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Temp</td>
<td>21℃</td>
</tr>
<tr>
<td>Ave Humid</td>
<td>54.3%</td>
</tr>
<tr>
<td>Ave Temp</td>
<td>12.75℃</td>
</tr>
<tr>
<td>Day Type</td>
<td>Saturday</td>
</tr>
</tbody>
</table>

Classification results of the forecasting days

<table>
<thead>
<tr>
<th>Date</th>
<th>Classification results</th>
<th>Date</th>
<th>Classification results</th>
</tr>
</thead>
<tbody>
<tr>
<td>25/1/2013</td>
<td>Cluster 3</td>
<td>29/1/2013</td>
<td>Cluster 5</td>
</tr>
<tr>
<td>26/1/2013</td>
<td>Cluster 4</td>
<td>30/1/2013</td>
<td>Cluster 1</td>
</tr>
<tr>
<td>27/1/2013</td>
<td>Cluster 2</td>
<td>31/1/2013</td>
<td>Cluster 6</td>
</tr>
<tr>
<td>28/1/2013</td>
<td>Cluster 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Forecasting model is SVM model:

1. Use 6 cluster load data set to train 6 corresponding SVM models;
2. According to the classification results of forecasting days, find the corresponding SVM model to finish forecasting;
3. Add up all distribution substation forecasting results to get the total load;

Parameter combination for each cluster

<table>
<thead>
<tr>
<th>Cluster</th>
<th>$\varepsilon$</th>
<th>$\delta^2$</th>
<th>$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>0.001</td>
<td>20</td>
<td>110</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0.0015</td>
<td>37</td>
<td>601</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>0.0071</td>
<td>87</td>
<td>115</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>0.0032</td>
<td>20</td>
<td>341</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>0.0016</td>
<td>28</td>
<td>200</td>
</tr>
<tr>
<td>Cluster 6</td>
<td>0.0092</td>
<td>32</td>
<td>123</td>
</tr>
</tbody>
</table>

where $\varepsilon$, $\delta^2$, $c$ is main factors of SVM model.

The parameter combination in left is find by Genetic Algorithms.
The former four steps are just one distribution transformer’s forecasting process. To get system load forecasting result. We need to repeat this for all distribution transformers and then add them up by big data technology.

Our short-term forecasting framework was implemented on Hadoop platform using four PC servers. Each PC server has two E5-2630V2 CPU and 500G memory. The table below shows the computation time.

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Size</th>
<th>Completion Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster(offline)</td>
<td>1.2 million consumers (1 yrs)</td>
<td>24 mins</td>
</tr>
<tr>
<td>Forecasting(online)</td>
<td>1.2 million consumers (daily)</td>
<td>110 secs</td>
</tr>
</tbody>
</table>
Add all distribution transformers’ forecasting results

Max relative error: 5.20%
Min relative error: 0.39%
Average relative error: 1.61%

Max relative error: 1.35%
Min relative error: 0.07%
Average relative error: 0.57%

The traditional approach which use SVM model and only focus on system load: Max relative error: 3.36% Min relative error: 0.51% Ave relative error: 1.68%.

It means the proposed method is better than traditional approach.
Outline

1. Background
2. Solutions
3. Case study
4. Conclusion
The proposed new framework of performing short-term load forecasting using smart meter data based on big data technologies.

**Conclusion:**

- Step 1 to 4 are carried out offline.
- In distribution EMS real-time application, once the date, forecasting weather data are available, the decision tree can find out which cluster the individual load belongs to.
- Then the appropriate forecasting model is selected to forecast individual load.
- Finally, the system load is forecasted.
- Case studies using a practical distribution system smart meter data indicate that the proposed forecasting method can not only guarantee prediction accuracy within the predefined range but also help distribution system operators gain better understanding of which individual load causes forecasting error.
Thank You!