Methodology formation of the training sample short-term forecasting electricity load for an energy supply company using data mining technologies

S O Khomutov¹, R N Khamitov², A S Gritsay², N A Serebryakov¹
¹Polzunov Altai State Technical University, Barnaul, 656038, Russia
²Omsk State Technical University, 11, Mira ave., Omsk, 644050, Russia

Abstract. For effective management of the power system operation mode the predictive information about hourly electrical load of all consumers is required. Forecasting errors, as a rule, lead to a decline of the technological and economic indicators of the power system operation, due to unreasonable changes of the generating equipment operating mode, as well as the selection of a non-optimal scheme of electrical networks. This article is devoted to improving the accuracy of short-term load forecasting of delivery points cluster of energy sales company with the use artificial neural networks. One of the most important conditions for achieving high prediction accuracy is the quality of the data sample required for training and testing neural network algorithms for short-term loading forecast. The proposed methodology is based on the authors' analysis uses the factors influencing the hourly power loading. The proposed methodology is based on the authors' analysis uses the factors influencing the hourly power loading as well as methods for improving the convergence of learning algorithms for artificial neural networks.

Keywords: artificial neural network, training example set, forecast electricity loading, factor analysis.

1. Introduction

The most important condition for the reliable functioning of the power system is the observance at any time of the balance of consumption and generation of electricity. Most of the power system generating equipment is inert, and the process of putting this equipment into operation takes more than 8 hours. The predictive information on the hourly power loading of all consumers is required for effective management of power systems. The purchase of electricity on the wholesale electricity and capacity market (WECM) requires short-term load forecasting (STLF). The electricity load forecasts accuracy in the conditions of the electricity market functioning significantly affects the technological and economic indicators of the power system [1]. Trading in deviations of the actual electricity loading from the forecast on the wholesale electricity market takes place on the balancing market at a disadvantageous price. The forecasting errors are increasing the cost of electricity for consumers in the retail electricity markets. It's also increasing the value of imbalance on the electricity market.

The main buyers of electricity on the wholesale electricity market are guaranteed electricity suppliers (ES), who purchase electricity using groups of supply points (SP). The SP is a conditional point, which can include several electrical substations with a voltage of 110 kV and above. As it was found out in [2], the hourly power consumption of the SP ES depends on a large number of factors, some of which are non-deterministic. In this regard, the problem of short-term forecasting of power consumption belongs to the class of poorly formalized problems. Under conditions of uncertainty,
traditional methods of mathematical statistics or simulation modeling do not allow building adequate models of objects [3]. Artificial intelligence and deep machine learning technologies are rapidly developing today. The considered information technologies are a highly effective tool for solving a wide range of tasks that are poorly formalized or non-formalized [4]. The accuracy of short-term forecasting of electricity load depends on the predictive algorithm, as well as on the quality of the sample of related statistical data [5]. The choice of highly correlated input data is critical for power prediction models based on artificial intelligence [6]. Consequently, special attention should be paid to the formation of a training sample in the short-term forecasting of the SP electricity load using machine learning tools.

2. Predicting electricity load methodology for the formation of a data sample intended for training and testing neural network algorithms

In [2] formulated the behavior of the hourly power electricity load time series of the SP short term depends on different factors:

– time factors (time of year, day of the week, an hour of the day, length of daylight hours, holidays, rallies, vacations in educational institutions, etc.);

– factors associated with the reliability of electricity, water, and heat supply to consumers related to this SP;

– meteorological factors;

– a factor that takes into account the powerful consumers operating mode of electricity-related to this SP.

Temporal factors are deterministic. Statistical data on the magnitude of temporary factors are included in the training sample in the form of ordinal numbers of the day in the year, an hour of the day, day of the week, etc., as well as signs of vacations, holidays, and pre-holidays, rallies, etc. The volume of undersupply of electricity, as a result of planned outages of power lines supplying consumers (feeder), it is possible to determine the average hourly electricity load for several typical days:

\[ V_i = \frac{1}{n} \sum_{j=1}^{n} j_i V_j \]  

where \( V_i \) – the volume of a shortage of electricity due to the disconnection of the feeding feeder per hour \( i \); \( n \) – number of regular days, over which averaging is performed; \( V_j \) – the volume of electricity load by the disconnected feeder per hour \( i \) of days \( j \).

The outside air temperature is the most significant external factor in the electricity load short-term forecasting [6]. In [7] to reduce the dimension of the factor space of the training sample, it is more expedient to take into account the factors of temperature and wind speed using the Stedman wind-cold index [8]:

\[ WCT = 0.2141 \cdot V + 1.162 \cdot T + 0.0124 \cdot V^2 + 0.0185 \cdot TV \]  

where \( WCT \) – wind-cold index; \( V \) – wind speed, m/s; \( T \) – outdoor temperature, °C.

In the research [7], it was found that changes in the SP electrical load are delayed to changes in the air temperature. This fact is due to the inertia of the thermal processes of buildings and structures. Changes in temperature in heated rooms lag behind changes in outdoor air temperature in time. Consequently, changing the mode of climatic devices arises with a delay to changes in ambient temperature.

It is proposed to take into account the factor of inertia of changes in electricity load caused by changes in ambient temperature using the dispersion of daily air temperature values:

\[ k = (-1)^r \cdot D = (-1)^r \cdot \frac{1}{24} \sum_{n=1}^{24} (t_i - \bar{t}) \]  

where \( r \) – the number of signs.

By adding the above factors to the training sample, it is possible to predict the hourly load of power systems using neural network algorithms.
where \( n = \begin{cases} 0, & \text{if } \frac{1}{24} \sum_{i=1}^{24} t_i \geq \frac{1}{24} \sum_{i=1}^{24} t_i \\ 1, & \text{if } \frac{1}{24} \sum_{i=1}^{24} t_i < \frac{1}{24} \sum_{i=1}^{24} t_i \end{cases} \) - correction factor for the direction of temperature changes; \( D \) - the selective variance of the last 24 air temperature values; \( t_i \) - air temperature per hour \( i \) per day \( X \); \( I(X-2) \); \( \bar{t} \) - sample mean.

An increase in the value of \( k \) signals about strong changes in the ambient temperature. With insignificant or short-term temperature changes, the increase in the dispersion of daily values of the ambient temperature will be small. Also, due to the correction factor \( n \), this factor allows you to determine the direction of temperature changes (warming or cooling).

In addition to air temperature and wind speed, the most influential meteorological factor that affects the electricity load of the SP during daylight hours is natural illumination. However, at present, there is practically no data on natural illumination in the public domain. Natural illumination mainly depends on the sky state, but there is a weak correlation between the time electricity load series and total cloudiness, or precipitation. The inclusion of these factors in the training set reduces the accuracy of predictive algorithms. In the course of this research, it was proposed to take into account the natural illumination factor on the basis of the following expression:

\[
P_t = \begin{cases} 0, & \text{if } P < 0.5 \\ 1, & \text{if } 0.5 \leq P \leq 2 \\ 2, & \text{if } 1 < P \leq 2 \\ 3, & \text{if } P > 2 \end{cases}
\]

where \( P_t \) - coded precipitation value; \( P \) - the amount of precipitation during daylight hours (from 7th to 19th hours), mm. This expression is based on the assumption that natural illumination depends on the following weather phenomena:

- clear or partly cloudy (precipitation less than 0.5 mm / 12 hours) - no stable decrease in natural illumination is observed;
- drizzling or short surface precipitation (precipitation in the range of [1 mm / 12 hours) - a slight decrease in natural illumination;
- overhead precipitation (precipitation in the range (2 mm / 12 hours) - significant decrease in natural illumination;

Let us estimate the relationship between the time series of electricity load and the factors of total cloudiness, precipitation, and coded precipitation, encoded using expression 4. Table 1 presents the data of calculating the Pearson correlation coefficient between the daily electricity load time series of the Yuzhnaya SP and the factors of total cloudiness, precipitation, and coded precipitation in Rubtsovsk for 2018.

The correlation coefficient is calculated based on the expression:

\[
r = \frac{\sum (y_i - \bar{y})(E_i - \bar{E})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (E_i - \bar{E})^2}},
\]

where \( y_i \) - factor value per day \( i \); daily electricity load of the Yuzhnaya SP per day \( i \), \( E_i \) - daily electricity load of the Yuzhnaya GTU per day \( i \); \( \bar{y}, \bar{E} \) - sample means

Pearson's correlation coefficient between the electricity load time series of the Yuzhnaya GTU and factors of total cloudiness, precipitation, and coded precipitation in Rubtsovsk for 2018: Coded precipitation \( P_k = 0.133 \); precipitation \( P = 0.072 \); Total cloudiness \( N = 0.106 \). As we can see, the coding of the precipitation factor using Expression 4 allows for an increase in conditionality at STLF.

Taking into account the factor of the mode of operation of consumers, when forming a training dataset, depends on the electricity load profile uniformity. For consumers with an electricity load uniform schedule, accounting for the operating mode of a given consumer at STLF can be carried out
using the sign of turning on (off) the main technological equipment that consumes electricity. For consumers with an uneven daily schedule of electricity load, it is necessary to take into account the hourly load as a separate parameter.

Since the statistical information on the magnitude of all influencing factors has a different value (from units to tens of thousands) and units of measurement, then with the STLF SP using artificial intelligence tools, the statistical series of data on the electrical load and the magnitude of the main influencing factors should be normalized to the range (0;1) [9]. Also, to improve the convergence of the learning algorithm for neural network predictive models, it is necessary to shift the boundaries normalization range boundaries from the saturation zone near the definition domain boundaries of the sigmoid activation function. Based on the above, the expression for normalizing static data will take the form:

\[ x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \cdot 0.9 + 0.05, \]  

(6)

where normalized statistic value range; \( x \) – actual value statistics range; \( x_{\text{min}}, x_{\text{max}} \) – the minimum and maximum values of a series of statistics.

Table 1 shows a fragment of the normalized training sample data.

| Parameter name                               | Parameter value, relative units |
|----------------------------------------------|---------------------------------|
| The ordinal number of the day of the year    | 0.87 0.87 0.87 0.87 0.87 0.87 |
| Day hour                                     | 0.05 0.09 0.13 0.17 0.21 0.25 |
| Electricity load                             | 0.16 0.19 0.30 0.40 0.46 0.48 |
| Wind-cold index                              | 0.60 0.59 0.59 0.58 0.59 0.59 |
| Temperature dispersion                       | 0.90 0.89 0.88 0.87 0.86 0.86 |
| Day of week                                  | 0.18 0.18 0.18 0.18 0.18 0.18 |
| Holiday and pre-holiday day parameter        | 0.35 0.35 0.35 0.35 0.35 0.35 |
| Vocation parameter                           | 0.95 0.95 0.95 0.95 0.95 0.95 |
| Precipitation                                | 0.05 0.05 0.05 0.05 0.05 0.28 |
| Day length                                   | 0.07 0.07 0.07 0.07 0.07 0.07 |
| Central heating cign parameter               | 0.95 0.95 0.95 0.95 0.95 0.95 |
| Hot water indication cign parameter          | 0.95 0.95 0.95 0.95 0.95 0.95 |
| Operating mode of a large consumer           | 0.95 0.95 0.95 0.95 0.95 0.95 |

After normalization, the entire training data set should be arbitrarily divided into data samples intended for training and testing the predictive algorithm, in a 9:1 ratio. Accordingly, the training of the neural network model takes place on the training set data. The neural network associative ability is evaluated on data from a test sample, unknown to the neural network after the end of the training process.

3. Results and its discussion

The training sample quality was assessed using a neural network, which is a three-layer perceptron trained using the adaptive momentum estimation (ADAM) method [10]. The structural structure of a three-layer perceptron is shown in Figure 1. This neural network consists of:

- an input layer with a size of 312 sensor elements of the input layer;
- the first hidden layer of 96 neurons with a sigmoid activation function and 1 activation threshold;
- the second hidden layer of 48 neurons with a sigmoid activation function and 1 activation threshold;
- an output layer of 24 neurons with a sigmoid activation function.
Figure 1. Three-layer perceptron block diagram designed for short-term forecasting of electricity load

Figure 2 shows the curves of changes in the average absolute percentage error of the electricity load short-term prediction for a different set of factors in the training sample. The average absolute percentage error in predicting the hourly SP ES electricity load was calculated based on the expression:

\[
MAPE = \frac{100}{N} \sum_{i=1}^{N} \sum_{k=1}^{24} \frac{|P_{ik}^{\text{fact}} - P_{ik}^{\text{predict}}|}{P_{ik}^{\text{fact}}},
\]

where \( MAPE \) – average absolute percentage error in the forecast of electricity load; \( P_{ik}^{\text{fact}} \) – actual electricity load at hour \( i \) day \( k \); \( P_{ik}^{\text{predict}} \) – forecast electricity load in \( i \) hour \( k \) day; \( N \) – number of data examples in training set.

The inclusion in the training sample of the daily values dispersion factor \( o \) of the outdoor temperature, calculated based on expression 3, makes it possible to reduce the forecast error. The training set inclusion of all the factors considered in the previous paragraph made it possible to reduce short-term hourly prediction electricity load error on the test data set after 50 epochs of the training cycle by 0.41%.

1)                                                              2)

Figure 2. Curves of changes in the average absolute percentage error of electricity load short-term forecasting with a different set of factors in the training sample on 1) training data; 2) on test data.
Conclusion.
An important aspect in the hourly short-term forecasting of electricity load using artificial neural networks is the quality of the training data set. The factors selection that have a sufficiently strong correlation with power consumption, as well as preliminary data processing when forming the training sample, contributes to a significant increase in the forecasting accuracy. The application of the proposed methodology for the formation of a training sample made it possible to reduce the average absolute short-term forecasting electricity load error of a group of supply points of a guaranteeing supplier by 0.41%.

References
[1] Domanov V and Bilalova A 2016 Forecasting power consumption based on source information. Bulletin of the South Ural State University. Series: Power Engineering vol 2. pp 59–65 https://doi.org/10.14529/power160208
[2] Gritsay A, Makarov V, Khamitov R, Tatevosyan A and Gritsay S 2019 The Method of short-term forecast electricity load with combined a sinusoidal function and an artificial neural network Proceedings of the 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering pp 523–526 DOI: 10.1109/EIConRus.2019.8656796
[3] Katasev A 2019 Neuro-fuzzy model of fuzzy rules formation for objects state evaluation in conditions of uncertainty Computer research and modeling vol 11, no 3 pp 477–492 DOI: 10.20537/2076-7633-2019-11-3-477-492.
[4] Tureczek A, Nielsen P and Madsen H 2018 Electricity consumption clustering using smart meter data Energies vol 11 pp 1–18 https://doi.org/10.3390/en11040859
[5] Ming-Wei L, Jing G, Wei-Chiang H and Yang Z 2018 Hybridizing chaotic and quantum mechanisms and fruit fly optimization algorithm with least squares support vector regression model in electric load forecasting Energies vol 11 no 9 P 2226 https://doi.org/10.3390/en11092226
[6] Potapov V, Khamitov R, Makarov V, Gritsay A, Tyunkov D and Chervenchuk I 2018 Short-term forecast of electricity load for LLC "omsk energy retail company" using neural network. Dynamics of Systems, Mechanisms and Machines: conference proceedings. (Omsk: Omsk State Technical University) pp 1–5 DOI: 10.1109/Dynamics.2018.8601430.
[7] Khomutov S, Stashko V and Serebryakov N. 2020. Improving the accuracy of short-term load forecasting of delivery point cluster of the second level default provider Bulletin of the Tomsk Polytechnic University. Geo Assets Engineering T 331 no 6 pp 128-140. https://doi.org/10.18799/24131830/2020/6/2682
[8] Quayle R and Steadman R 1998 The Steadman wind chill: an improvement over present scales. Amer. Meteorological Soc. vol 13 pp 1187–1193.
[9] Liu N, Tang Q, Zhang J, Fan W and Liu J. 2014 A hybrid forecasting model with parameter optimization for short-term load forecasting of micro-grids Applied Energy vol 129 pp 336–345
[10] Kingma D and Lei Ba J 2015 ADAM: a method for stochastic optimization ArXiv.org URL: https://arxiv.org/abs/1412.6980.