INPUT DATA SELECTION TECHNIQUE FOR ANN LOAD FORECAST MODELS: A HYBRID APPROACH

Badar Islam, Zuhairi Baharudin and Perumal Nallagownden  
Department of Electrical & Electronics Engineering.,  
Universiti Teknologi PETRONAS, 32610, Bandar Seri Iskandar, Perak, Malaysia  
Corresponding Author: badar.utp@gmail.com

ABSTRACT - Selection of best dataset for ANN training and testing process is a key to building an effective load demand forecast model. This paper focuses on a load forecasting framework with an emphasize on a hybrid data preprocessing technique to enhance forecast accuracy. In this technique the correlation coefficient and genetic algorithm are used for the selection of a reduced and optimized dataset for a multilayer perceptron neural network (MLPNN). As the Artificial Neural Network (ANN) has a tendency to map and memorize the non-linear relations between inputs and output variables, they are extensively implemented in modern predictive model development including electrical load forecast models. The non-linear behavior of the load can only be narrated effectively by ANN, if a group of most appropriate inputs is identified prior to the training process. Results show that this method is effective and forecasts one day ahead load with reduced mean absolute percentage error (MAPE).

Keywords - Artificial neural network, artificial intelligence, correlation analysis, genetic algorithms, load forecasting.

1. INTRODUCTION

Electrical load forecasting is the prediction of future load, which plays a very important role in the energy management system to provide a better environment for future planning and decisions [1, 2]. These forecasts/predictions are categorized into short-term, medium-term and long-term forecasts with respect to planning horizon’s duration. Among the others, short term load forecast is vital, as it supports the operation managers in their day to day operational decisions, such as financial planning of generating capability, arrangement of fuel, approximation of peak demand and system safety assessments [3-5].

Many statistical and computational techniques are used by the researchers for these predictions. Artificial neural networks (ANN) are abundantly used in the development of forecast models for short term load in the literature in the past few years [6]. The effectiveness of the forecast model is based on multiple factors, such as selection of ANN architecture, choice of training algorithm and input data set used for the training and testing of the network [7]. Among the other factors, the selection of most influential input variables and training data set for ANN is focused in this paper.

The objective of this research is to find a technique for the selection of the input variables and training and testing dataset that leads to the minimum forecast error and reduced computational time for ANN based load forecast model. In general, an increase in the number of inputs of a model network may result in the form of a wider training capacity and better forecasting accuracy. However, the erroneous choice of the additional input variables and bad data would increase the forecast error [8].

ANN has the ability to map and store the non-linear relations between inputs and outputs, but if a group of non-relevant variables are selected as an input variable set, the training time of ANN is increased and the errors become bigger [9]. The nonlinear relation of the load can be effectively explicated only when a group of appropriate input variables is found. The impact of the correlation coefficient and data dispersion is used for input variable selection in the first stage of ANN based short-term load forecast model. The genetic algorithm is used to select the best individuals among the population in the second stage on the basis of minimum mean absolute error (MAE) used as the objective function.

Genetic algorithm (GA) is a guided random search method [10], which was initially introduced by John Holland in 1970s at the University of Michigan. The major inspiration is adopted from the natural system of evolution for the designing of an artificial system that retains its robustness and adoption properties. Since the discovery of GA, these techniques are consistently improved by other researchers and are now widely implemented in various fields (business, science and engineering) to address and solve multiple optimization problems. GA implements the biological processes to perform a random search in a defined N-dimensional search space [11].

Several attempts have been made for the identification of appropriate input data set for the training and testing of ANN based models. Most of these efforts emphasize on the correlation analysis of the data, however some of the researchers also focused on the combination of mathematical formulation to resort the derived variables by squaring, averaging, adding or differencing the data sequences to determine the appropriate input variables [12]. These mathematical and statistical techniques for the manipulation of data have a common attribute that they attempt to identify almost completely linear relationships among a set of input variables and their dependent variables [13]. Because of their linear and consistent approach, these methods are unable to track the unusual and brisk variations occurring in the real time input data.

A variety of input variables have been tried out for ANN based STLF, such as a historical load of multiple intervals, meteorological variables and economic variables [14]. Normally, hit and trial methods of deploying multiple combinations of above mentioned variables are reported in the literature in conjunction with the multiple neural network topologies.
2. DATA HANDLING AND QUANTITATIVE ANALYSIS

Five year past data of load demand and weather related variables of an Australian utility at half hour sampling frequency are used in the research. A complete quantitative examination is conducted to study multiple trends of the data. The analysis reports statistical facts of the data including, least and highest values, mean, mode, median, variance and standard deviation. Table 1 encapsulates the numerous statistical features of the database of input variables based on this analysis. Matlab 2013b, GMDH and Neuroshell 2 are used as simulation software.

**TABLE 1. QUANTITATIVE ANALYSIS OF HISTORICAL LOAD AND METEOROLOGICAL VARIABLES OF DATA SET**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>System Load</th>
<th>Dry Bulb</th>
<th>Dew Point</th>
<th>Wet Bulb</th>
<th>Humidity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min. Value</td>
<td>5498.36</td>
<td>7</td>
<td>-2.5</td>
<td>-8.4</td>
<td>3.7</td>
</tr>
<tr>
<td>Max. Value</td>
<td>14274.1</td>
<td>100</td>
<td>24.2</td>
<td>43.8</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>8894.00</td>
<td>68.9007</td>
<td>14.8772</td>
<td>11.9235</td>
<td>18.2599</td>
</tr>
<tr>
<td>Median</td>
<td>8992.58</td>
<td>70</td>
<td>15.1</td>
<td>12.45</td>
<td>18.5</td>
</tr>
<tr>
<td>Mode</td>
<td>9450.98</td>
<td>66</td>
<td>18.5</td>
<td>12.3</td>
<td>21</td>
</tr>
<tr>
<td>Range</td>
<td>8775.79</td>
<td>93</td>
<td>23.8</td>
<td>32.6</td>
<td>40.1</td>
</tr>
<tr>
<td>Variance</td>
<td>198541</td>
<td>284.127</td>
<td>18.4244</td>
<td>29.8862</td>
<td>23.9327</td>
</tr>
</tbody>
</table>

In this analysis the system load and meteorological variables are divided into nineteen different slots to analyze the frequency of occurrence of each sample in a certain range. The analysis returns a percentage of individual and cumulative frequency of occurrence of samples from a certain system load range. This dispersion and repetition analysis of the system load is beneficial for finding the data samples which are repeated frequently and have the maximum influence in the prediction process. This analysis of the system load is depicted in Table 2.

Fig. 1 reflects the frequency of occurrence of a certain load data range along with its cumulative frequency in graphical order. In Fig. 2, the frequency of occurrence of meteorological variables is shown with data ranges and cumulative frequency. The ranges with maximum and minimum number of samples of these variables are shown in Table 3.

Correlation analysis is most frequently reported method to verify the significance of data used for prediction purpose in the literature. It provides a comprehensive measure of the relationship between the variables in a data set. In general, the bigger correlation coefficient of the input variable with expected output values indicates a strong relationship between them. The general form of the correlation coefficient r can be formulated as:

\[ r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}} \]  

(1)

Where, n shows the total number of x and y pairs, the value of r lies between -1 and +1. The positive and negative signs are used to indicate positive linear coefficient and negative linear coefficient respectively. It provides a comprehensive measure of the relationship between the variables in a data set. The
results of this analysis between all the projected input variables and load demand are shown in table 4.

**Table 3. Ranges For Frequency Of Occurrence Of Weather Variables.**

<table>
<thead>
<tr>
<th>Input Variable</th>
<th>Maximum no. of Samples</th>
<th>Minimum no. of Samples</th>
<th>Moderate Sample Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dry Bulb Temperature</td>
<td>21-24°C</td>
<td>42-45°C</td>
<td>27-30°C</td>
</tr>
<tr>
<td>Wet Bulb Temperature</td>
<td>15-18°C</td>
<td>3-6°C</td>
<td>9-12°C</td>
</tr>
<tr>
<td>Humidity</td>
<td>65-70%</td>
<td>10-15%</td>
<td>50-55%</td>
</tr>
</tbody>
</table>

**Table 4. Correlation Analysis Of Load And Weather Related Variables.**

<table>
<thead>
<tr>
<th>Load Related Variables</th>
<th>Weather Related Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same day and time in the previous week (W-1) : 0.8821</td>
<td>Dry bulb temperature : 0.111</td>
</tr>
<tr>
<td>Same time in the previous day (D-1) : 0.8901</td>
<td>Dew point : -0.116</td>
</tr>
<tr>
<td>Same day previous hour (H-1) : 0.9220</td>
<td>Wet bulb temperature : -0.024</td>
</tr>
<tr>
<td>Same day, two hours earlier (H-2) : 0.9018</td>
<td>Humidity : -0.271</td>
</tr>
</tbody>
</table>

The results of correlation analysis show that the load related variables have higher values than weather variables. The correlation coefficient values of the input variables used in this research are shown in Fig. 3.

**Fig. 3. Importance of input variables according to correlation values.**

### 3. METHODOLOGY

All the samples of input variables are divided in multiple ranges as shown in table 2. Half of the samples from these data ranges are selected on the basis of frequency of occurrence and effect of correlation with respect to the forecast error. Higher correlation and frequency of occurrence of the data samples is the primary information which is used for data selection based on probability and discard information provided in table 5 in this first stage of the experiment. This technique ensures that the data samples from each range are present in the final data set. The presence of data samples related to all ranges is necessarily required for effective mapping between inputs and target outputs during the ANN training procedure. Another other important consideration is to discard the data samples having the repeated values because the similar data play an insignificant role in the training process of ANN.

The remaining data samples are subjected to ANN to formulate a set of factors based on various input combinations that contain the correlation and data distribution impact of inputs and output that is constituted on the basis of forecast error. The forecast load values, which have a higher correlation coefficient and maximum number of occurrences with expectation output values, are chosen from this factor set as input variables. By means of this method, a preferable input variables set can be determined. The high correlation between the input variables and the forecasting points along with repeated occurrences of a certain date range indicate that the forecasting results are more exact.

**Table 5. Probability and discard information for initial data samples.**

<table>
<thead>
<tr>
<th>Ranges</th>
<th>Discard Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range 1 (maximum frequency of occurrence)</td>
<td>90 %</td>
</tr>
<tr>
<td>Range 2 (frequency of occurrence less than R-1)</td>
<td>80 %</td>
</tr>
<tr>
<td>Range N (Least frequency of occurrence)</td>
<td>10 %</td>
</tr>
</tbody>
</table>

In this section, the reduced dataset obtained from statistical handling during, as explained above is subjected to genetic manipulation of chromosomes for further optimization and reduction to constitute the most appropriate input variable dataset. Mean absolute error (MAE) is the fitness function for this selection mechanism. The shortlisted samples are subjected to the genetic algorithm for the final selection of the data set. The most crucial part of the genetic-based algorithm is to define the fitness function. In this case the fitness function is based on the mean absolute error (MAE).

The various steps of this algorithm are explained below.

2. [Fitness] Evaluate the fitness $f(x)$ of each chromosome $x$ in the population.
3. [New population] Create a new population by repeating following steps.
4. [Selection] Select two parent chromosomes from a population according to their fitness.
5. [Crossover] Cross over the parents to form new offspring.
7. [Replace] Use new generated population for a further run of the algorithm.
8. [Test] If the end condition is satisfied, stop, and return the best solution in current population.
9. [Loop] Go to fitness function evaluation step again.

In the first step of this genetic-based algorithm the chromosome length and index of input variables are defined. Real coded GA is implemented with a fixed length chromosome equal to 5 and the value of each genome in a particular chromosome is the index of the input variable. In the second stage genetic algorithm is used for the selections of the final data set. Mean absolute error (MAE) is the fitness function for this selection mechanism. Phases of research are highlighted in Fig. 4.

A real coded GA using an elitism selection method with a single point crossover and multipoint mutation is implemented. The crossover and mutation probabilities are selected as 0.8% and 0.1% respectively. The fitness function of the genetic algorithm part of this hybrid model is calculated on the basis of minimum error which is computed as shown in Eq.2. The fitness function used by GA is represented by Eq.3.

$$\text{err} = \sum_{i=1}^{p} \left| \frac{A_i - P_i}{P_i} \right|^4$$  \hspace{1cm} (2)
fitness = \frac{1}{1+err} \tag{3}

Where, \( A_t \) and \( P_t \) are the actual and forecasting values at time point \( A_t \) and \( P_t \) are the mean of actual and forecasting values respectively.

\[ 
\text{Training} \quad \text{Divide all the input variable data samples into multiple ranges according to their frequency of occurrence (Table 2)}
\]

\[ 
\text{Find correlation impact of all data ranges. Select the data samples from each range according to the probability and discard information provided (Table 5)}
\]

\[ 
\text{Apply genetic algorithm on the remaining samples for the selection of most influential data set, choose 20\% best samples where fitness function is MAE}
\]

\[ 
\text{Training & Testing of ANN using BP Algorithm with actual and optimized data set}
\]

\[ 
\text{Compare the results of both data sets}
\]

\[ 
\text{Performance with optimized data set is better?}
\]

\[ 
\text{YES}
\]

\[ 
\text{NO}
\]

\[ 
\text{END}
\]

Fig. 4. Flowchart for the proposed input variable selection technique

4. RESULTS AND DISCUSSION

Data dispersion and frequency of occurrence of a certain data range is found and a proportion of the data samples are selected on the basis of inverse probability and discard scheme discussed in the previous section. The correlation coefficient method is applied to further shortlist the data samples. Finally, the selected data samples are brought under genetic manipulation of chromosomes after applying these statistical methods. Mean absolute error (MAE) is used as a fitness function and the best samples are further acquired to constitute a final data set for the training and testing of artificial neural network using the back propagation training mechanism. The 8-10-1 NN architecture based on a BP training algorithm is used to compare the results with optimized data set and the original data set.

The scatter plot of learning and predicted residuals is shown in Fig. 5 (a). The autocorrelation impact of all the historical load data samples is depicted in Fig. 5 (b) and the number of occurrences of the error residuals is shown in Fig. 5(c). Half hourly recorded samples for 24 hours of a day over a five year period are presented below in graphical form in Fig. 5. The graph shows a plot of both actual and forecast loads in MW against half hourly data samples. The mean absolute percentage error (MAPE) of 4.35 % is calculated in this technique. The results obtained from testing the neural network on the new dataset which has a significantly lesser number of samples as compared to the original dataset produced better results in terms of forecast accuracy and network training time. The MAPE is recorded as 3.19 % using the optimized and reduced data set. A significant reduction of 1.06% shows the efficacy of the proposed method. The forecast results based on this optimally selected data set are depicted in Fig. 5.

5. CONCLUSION

ANN performs well in the domain of short term load forecasting because of their significant ability to map the nonlinear relationship between applied inputs and target outputs. However, if a non-distinctive dataset is selected for the training of neural network the training time and error can be drastically increased that decays the network performance. Selection of most appropriate set of input variables not only gives rise to enhanced forecast accuracy, but saves the computational efforts and time. In this paper, a new hybrid technique based on statistical methods and genetic algorithm is used to select the optimal data set of input variables. To verify the efficacy of this data selection method, the ANN based on a BP training algorithm is employed using this optimized data set and the original data set containing all the recorded inputs. Training and testing results of these two data sets show that this data selection method is efficient as a considerable reduction in error is observed.

REFERENCES


