EAI Endorsed Transactions
on Energy Web

Research Article **EAI.EU**

Improved Prediction of Wind Speed using Machine Learning

Senthil Kumar P

School of Information Technology & Engineering, VIT University, Vellore. India

Abstract

The prediction of wind speed plays a significant role in wind energy systems. An accurate prediction of wind speed is more important for wind energy systems, but it is difficult due to its uncertain nature. This paper presents three artificial neural networks namely, Back Propagation Network (BPN), Radial Basis Function (RBF) and Nonlinear AutoRegressive model process with eXogenous inputs(NARX) with Mutual Information (MI) feature selection for wind speed prediction. The MI feature selection identifies the significant features and reduces the complexity of wind speed prediction model without loss of information content. The performance of prediction model is evaluated in terms of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The results show that the performance of all three neural network models are highly satisfied. Moreover, NARX model with mutual information feature selection is more accurate in dealing with wind speed prediction.

Keywords: Wind Speed, prediction, neural network, feature selection, renewable energy

Received on 25 September 2018, accepted on 19 January 2019, published on 21 March 2019

doi: 10.4108/eai.13-7-2018.157033

Corresponding author. Email:senbe@rediffmail.com

1. Introduction

Renewable energy sources are available in abundance naturally and can be utilized for power generation to satisfy industrial and commercial needs. The renewable sector is witnessing a phenomenal growth, and accurate prediction is essential. In recent years, the renewable nature of wind power makes it more attractive. Nowadays, many countries face lots of problems with their energy sources due to many environmental factors and the uncertainties of origin. The power generated from wind energy is better than other energy sources. The wind turbine layout uses advanced technologies to decrease the cost of wind power generation and permit large-scale integration into the energy grid. The primary input needed for producing wind power generation is the natural wind. When the wind moves through wind turbines, the turbines convert the wind energy into mechanical energy. Then, the wind power is produced from that mechanical energy [1]. The advantages of wind energy are combined with a few difficulties like high uncertainty, limited predictability and wind power energy is not entirely deliverable [2].

The generation of wind power can be forecasted using numerous methods on different time scales. Traditional approaches based on physics and mathematical modeling cannot handle the wind as a matter of course, considering the unpredictable variation inherent. There are two methods used for the purpose. The first method is to forecast wind power output directly based on the history of wind power data, and other process uses the power curve model to convert the natural wind speed forecast into wind power output. The wind turbine converts power from the wind that is conveyed with a power curve. The wind turbine performance can be measured by using wind speed where wind speed measured by using an anemometer. This research more focuses on the wind speed forecast by using meteorological datasets.

The physical and statistical methods are two standard methods commonly used for wind speed forecasting. The combination of both methods is also utilized in some models for integrating the advantages of them. The physical method considers parameters related to physical description of wind movement in and around the wind farm. It relies on weather forecasting data like atmospheric variable and also the characteristics of wind farm environment like farm layout, obstacles, and roughness. These data are used for predicting wind power by considering the wind speed and transforming it to wind turbine at the wind farm. The physical method does not need any training input from past data. This has the offset against the problem involved in obtaining physical data. The statistical methods consider the training of past history of wind speed data and generates an output without considering physical phenomena. A statistical approach includes artificial neural networks, fuzzy logic, regression tree and support vector machine etc. The statistical approach produces a good result in wind speed forecasting. The hybrid approach is a combination of different physical and statistical approaches like a combination of Numerical Weather Prediction (NWP) and neural networks [3], [4]. Forecasting of a wind speed is an essential measure to find the uncertainty and can also be used for some purposes such as power commitment decision, power increase or decrease decision, maintenance arrangement and energy storage optimization. The forecasting system predicts wind speed for wind power generation. Based on time horizon the wind speed forecasting is classified into three types. The first-time horizon is a very short term which is very useful for trading in intraday markets. It represents few minutes to one hour only. The second time horizon is a short term which is suitable for maintenance scheduling. It represents one hour to 12 hours. The third time horizon is a medium or long term which is useful for the maintenance of non-renewable power generation. It represents several hours to a few days.

Advanced technologies have empowered taking in the huge volume of data on a continuous or periodic basis in various disciplines. These data represent the potential to discern valuable information and knowledge. Manual processing of a large amount of data to determine useful information is difficult. The dataset in the modern era has high dimensionality and huge volume, which required an automated tool for processing. This confinement needs automated tools to mine useful information and knowledge from the huge volume of data [5]. Nowadays, the enormous amount of data is collected and stored at an extraordinary rate. Consequently, the existing data analysis tools and statistical methods are inadequate for analysing this large size of data. Data mining techniques are used to provide interesting interpretation by extracting useful information from large data sets. In recent years, Knowledge Discovery in Database (KDD) plays an important role in processing large quantities of data. KDD is the process of analysing the database to explore or discover useful patterns. In this process, the data mining acts as a central process which involves deriving of algorithms to explore hidden information, develop a suitable model and discover interesting patterns [6]. On the other hand, machine learning more emphases on prediction based on known properties learned from the training data. It is capable of generalizing or optimizing data from large quantities of data. It is more faster and produces accurate results in order to identify patterns or solution. Relating machine learning with artificial intelligence will create it more effective in processing large quantities of data. Liu et al.[7] have designed the combination of Wavelet and Improved Time Series Method (ITSM) for prediction of wind speed and power. The wind speed data in time series form is decomposed into multiple subseries by using wavelet method. Then, ITSM is used to construct the prediction model for each of that subseries to find the forecasting in every subseries. Finally, the outputs of each of the subseries are collected for obtaining the forecast wind speed. As a result, the hybrid technique effectively enhances an accuracy of the wind speed. Artificial Neural Network (ANN) is the vastly utilized statistical approach for prediction of wind speed. ANN is trained from the data sets of past observation for learning the dependencies of the output based upon input values. ANN model has the capability of self-learning and self-adaption. Fadare [9] has proposed a model to predict the wind speed using ANN method. In the study, they have utilized the gradient descent backpropagation algorithm in a neural network to train ANN. For this network, latitude, longitude and altitude meteorological parameter were given as input and the forecasted monthly mean wind speed was produced as output. The minimum MAPE of an output found from this network is 8.9%.

2. Importance of Feature Selection Process

A feature selection method helps reduction of computational complexity of learning algorithm, improve prediction performance, better data understanding and reduce data storage space. Feature selection has gained immense popularity in machine learning applications. The feature selection finds the minimum number of feature subsets that retains high accuracy to represent the original features. When the choices of a number of subsets of features are very small, the chances of information content may be low. On the other hand, the presence of noise as an irrelevant data is highly probable when many features are selected. Hence, feature selection should be on the right selection of subsets, avoiding too large or too small number of features. There are many advantages of using the feature selection methods like data visualization and understanding, reduction of memory storage, reduction of training time and reduction of the dimensionality which may improve prediction and classification performance. The feature selection process discards redundant and irrelevant attributes while retaining the model accuracy. Irrelevant features are not useful for improving the performance of the model and should be removed. Instead, an unnecessary feature represents the same feature or relevant of another feature when both have the same impact on the model accuracy the feature selection process can remove one of them. Consider the input data $Y = \{ y_1 , y_2 , \dots , y_m \}$ of size M and number of features $y_i = \{ f_1 , f_2 , \dots , f_L \}$ and the target prediction feature T. Feature selection process chooses the best features from y_i that effectively characterizes T. The feature selection is optimal if and only if it produces highest-class separability with minimum representation. The method considers the generalization problem to construct a minimum paradigm for defined size dataset. This process deals with bias-variance problems in data pre-processing based on the specified dataset. The processing algorithm guarantees to reduce complexity by using limited features from the available datasets which should be identified to obtain the best generalization.

The feature selection methods are broadly classified into three groups namely filter, wrapper and embedded methods. The filter method uses heuristics based on the prominent characteristics of the data handled instead of an algorithm for evaluation of the quality of feature subsets, with the application of a statistical measure for assignment of a rank to each feature. The ranking measure is used for scoring the features, and a threshold is used to eliminate features below the threshold. The filter method uses proper threshold value for the selection of subsets with most relevant features. It is an independent measure for evaluating features subsets without utilizing any machine learning algorithms. There are many independent criteria employed for assessment including interclass distance measures, uncertainty measures, dependency and consistency measure and the probability of error measures [10]. This method is faster than the wrapper method and practical for use on high dimensional datasets. Wrapper-based feature selection method searches for best subset of features using predetermined accuracy from an induction algorithm for evaluating generated subsets of features. The method may produce improved performance but is expensive to run. This method utilizes the learning algorithms to evaluate the subsets, while dataset is very large. In the wrapper method, optimal subset generation is the process of the heuristic search algorithm. Many search algorithms can be utilized for finding feature subsets. A sequential search algorithm is an iterative like algorithm which starts with an empty feature or complete feature and add or remove one feature at a time until the target is obtained. Heuristic search algorithms evaluate different subsets for optimizing the best solution. Wrapper methods make use of the classifier for providing a score of the feature subset depend on their predictive power. The embedded method combines the advantages of filter and wrapper approaches. In this method, the feature set is evaluated through the use of independent criteria and a learning algorithm. The independent criterion is used for the choice of an optimal subset with prearranged cardinality. The learning algorithm uses various cardinalities to select the finest subset from the optimal subsets. These methods are appropriate for feature selection mainly used on high dimensional datasets [11].

3. Mutual Information Feature Selection

Mutual information (MI) is a non-linear correlation based method which is used to calculate the information shared by two, three or more number of features.[12] If the features are independent they do not have any common information and one feature does not give any information about other feature. The amount of information shared by each feature with another feature is called entropy [19].Let X and Y are two features, then I(X; Y) represents the mutual information between them. The Mutual information I(X;Y) is zero when both X and Y are independent. The Mutual information is symmetric and non-negative. [20],[21],[22],[26]

For discrete random variables (X,Y), the MI is

$$\mathrm{I(X;Y)} = \sum_{\mathrm{x\epsilon X}} \sum_{\mathrm{y\epsilon Y}} \mathrm{P(x,y)} \log\left(\frac{\mathrm{P(x,y)}}{\mathrm{P(x)P(y)}}\right) \tag{1}$$

P(x,y) represents the joint probability of ‘x’ and ‘y’. p(x) and p(y) are marginal probability density function of x and y respectively.

For continuous random variables (X,Y), MI is

$$I(X;Y) = \int_x \int_y P(x,y) log\left(\frac{P(x,y)}{P(x)P(y)}\right) dxdy \tag{2}$$

$$P(x) = \int_y P(x,y)dy \tag{3}$$

$$P(y) = \int_x P(x,y)dx \tag{4}$$

Alternate way to express mutual information in the form entropy is

$$I(X;Y) = H(X) + H(Y) - H(X;Y) \tag{5}$$

H(X) and H(Y) represent the marginal entropies. H(X;Y) represents the joint entropy of X and Y

$$H(X) = -\int_x P(x)logP(x)dx \tag{6}$$

$$H(Y) = -\int_y P(y)logP(y)dy \tag{7}$$

$$H(X;Y) = \int_x \int_y P(x,y)logP(x,y)dxdy \tag{8}$$

The MI between two random variables can be calculated by estimating the Probability Density Function (PDF). The estimation can be done in three ways by using parametric, non-parametric and the combination of parametric and non-parametric density estimations. The non-parametric density estimations are statistical methods such as KNN, Kernel density, Wavelet density, etc. which allows functional form of regression function to be flexible whereas parametric density estimations are density functions like Bayesian, Maximum Likelihood, Maximum posteriori, etc. which assumes data from known group of distributions such as normal, Gaussian, etc. and then optimizes the parameters of the function by fitting the model to the dataset. The mixed density estimation takes the benefits of both parametric and non-parametric estimations and increases the quality of estimation. In the proposed model, the KNN estimator is utilized for probability density function estimation. First, the KNN method estimates the probability density function, then it calculates mutual information. While the value of K is small the bias also small but the variance is high, whereas the size of K is large the bias also large, but the variance is small. The procedure of Mutual Information Feature Selection (MIFS) algorithm is as follows

1. Let X be an input set, T be an output set and F be a final selected features set
2. Initialize X with the set of inputs and F with an empty set

3. Find mutual information for each input feature ‘x’ with an output

   I(x; T) where x € X

4. Select max[I(x; T)] then set X and F as follows

   X🡐X-{x}

   F🡐{x}

5. Find mutual information between each feature in X and F

   I(x; f) where x€X and f€F

   5.1 Select x as the next input if it maximizes

$$max\left[\left(I(x;T)\right)-\left(\beta/|F|\right)\sum_{f€F}\left(I(x;f)\right)\right]$$

   5.2 Set X🡐X-{x}

   Set F🡐FỦ{x}

6. Selected features are in set F

4. Wind Speed Forecasting using Neural Network

The basic concept behind the ANN is to develop a tool that should perform computation for demonstrating the brain function. This tool must carry out various computations at a rate faster than the computed rate of the conventional framework. The ANN can be utilized for various purposes like clustering, classification, prediction, etc. During the learning procedure, known patterns of a particular problem are presented to the network to improve its performance and its generalization ability. The generalization capability is an ability to respond to the pattern correctly, which was not used during the training process. An optimization method based on gradient descent is applied to reduce the error or maximize the accuracy of the neural network. There are two major categories of learning called supervised and unsupervised learning. In supervised learning system, the class label of the pattern presented to the network is known, and it is used in the training process. If the class label is unknown or unused, the learning process is unsupervised.

ANN has the capability of learning and generalizing the given input by assigning or adjusting its weights and biases for making useful decision.

4.1 NARX Model

The nonlinear autoregressive network with exogenous inputs provides very accurate chaotic time series prediction that perfectly fits to wind speed forecasting. This network with delayed inputs, delayed recurrent (feedback) outputs, the nonlinearity and dynamic character allow computation and determination of tasks that are almost impossible to solve for conventional methods or linear (time invariant) systems. NARX is a recurrent dynamic neural network, with feed-backward connecting with many layers of the neural network [13].

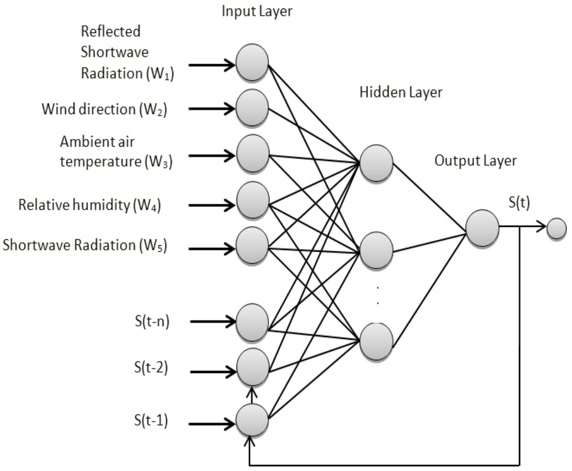

Figure 1. NARX model structure

In the non-linear systems, the NARX model has utilized for the time series oriented data [14].The learning method used in NARX model is more efficient than other neural networks. In NARX the gradient descent is improved which makes NARX becomes more active. Figure 1 shows the NARX network structure. There are two portions of inputs are given to the NARX network.

The NARX uses faster network connection with good generalization ability. It can be trained by using series-parallel mode or only parallel mode. The series-parallel mode feeds an original output directly as part of the input to the network but the only parallel mode feeds back the output as a part of the input [15, 16, 17, and 18]. Let us consider, w (t) as the external input, s(t) as the output of the network at particular time t and n_i as the delay time of external input. This NARX network can be defined as shown in equation (5).

$$s(t)=f(s(t-1)s(t-2)\ldots..s(t-n_0),\ ,w(t-2)\ldots.w(t-n_i)) \qquad (9)$$

The output of NARX network s(t) depends on current and previous input w(t-i) and also on the previous output of the network s(t-i).

4.2 Back-Propagation Network (BPN)

In the area of the ANN, back-propagation is the most important network utilized in many of the real-world applications. It uses multi-layer feed forward network with input, hidden and output layers. The BPN training process consists of three stages namely, feed forward of the input training, backpropagation of error and updation of weights [8]. During the feed forward stage, the signal x_i is given as an input to each input node X_i(i=1 to n) and send it to the hidden layer. For each hidden node z_j (j-1 to p) the net input is calculated by summing the weighted input signal.

v_{ij} is the connection weight between an input node i and hidden node and v_{oj} is bias on hidden node

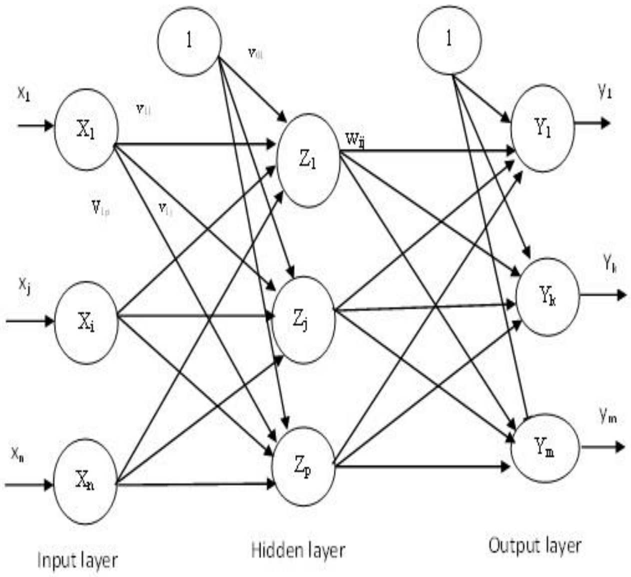

Figure 2: Back-Propagation Network

$$z_{netj} = v_{oj} + \sum_{i}^{n} x_i v_{ij} \tag{10}$$

The output of the hidden node is calculated by applying an activation function.

$$z_j = f(z_{netj}) \tag{11}$$

The commonly used activation function is binary sigmoid or bipolar sigmoid function. The sigmoid function is defined as

$$f(x) = \frac{1}{1 + e^{(-x)}} \tag{12}$$

For each output node $y_k (k = 1\ to\ m)$ the sum of its weighted input signal is given as net input. w_{jk} represents the connection weight from hidden to output node and w_{ok} is bias on output node. The net input for the output node is calculated as follow

$$y_{netk} = w_{ok} + \sum_{i}^{p} z_j w_{jk} \tag{13}$$

The output of the output layer node is calculated by applying its activation function

$$y_k = f(y_{netk}) \tag{14}$$

In the second stage is backpropagation of error is calculated from output layer to hidden layer. Error correction δ_k is computed as follows

$$\delta_k = (t_k - y_k) f'(y_{net}) \tag{15}$$

$$f'^{(y_{net})} = f(y_{net})[1 - f(y_{net})] \tag{16}$$

Where $t_k$represents the target vector. Based on the error correction, update the changes in the weight and bias

$$\Delta w_{jk} = \alpha \delta_k z_j \tag{17}$$

$$\Delta w_{ok} = \alpha \delta_k \tag{18}$$

Calculate an error term δ_j between hidden and an input layer

$$\delta_{netj} = \sum_{k=1}^{m} \delta_k w_{jk} \tag{19}$$

$$\delta_j = \delta_{netj} f'^{(z_{netj})} \tag{20}$$

Based on the error correction update the changes in the weight and bias

$$\Delta v_{ij} = \alpha \delta_j x_i \tag{21}$$

$$\Delta v_{oj} = \alpha \delta_j \tag{22}$$

In the third stage, each output node updates the weight and bias

$$w_{jk}(new) = w_{jk}(old) + \Delta w_{jk} \tag{23}$$

$$w_{ok}(new) = w_{ok}(old) + \Delta w_{ok} \tag{24}$$

Each hidden node updates the weight and bias

$$v_{ij}(new) = v_{ij}(old) + \Delta v_{ij} \tag{25}$$

$$v_{oj}(new) = v_{oj}(old) + \Delta v_{oj} \tag{26}$$

Initially random values are assigned as the interconnection weights values and bias values. Later, these values are changed based on the BPN training process.

4.3 Radial Basis Function (RBF) Network

The radial basis function is a functional approximation neural network which uses the most common nonlinearities such as Gaussian kernel functions and sigmoid. The input layer has a set of input units receiving an input signal and forwards it to the hidden layer. The activity function is usually Gaussian function that is regulated by hidden layer. The output layer executes linear transformation from the hidden unit to the output space.[23],[24],[25]

The RBF neural network uses three parameters namely the centre of the basis function, the variance and weights between hidden layer and an output layer. The Gauss function in RBF network is represented as

$$R(x_p - c_i) = exp\left(-\frac{1}{2\sigma^2}\|x_p - c_i\|^2\right) \tag{27}$$

Where $\|x_p - c_i\|$ represents Euclidean distance, c represents centre of Gaussion function and σ represents variance of Gauss function. The output of the RBF neural network is shown in equation (28).

$$y_j = \sum_{i=1}^{h} w_{ij} exp\left(-\frac{1}{2\sigma^2}||x_p - c_i||^2\right)\ j = 1,2 \ldots.,n \tag{28}$$

Where $x_p = (x_1^p, x_2^p, x_3^p, \ldots \ldots \ldots \ldots . x_m^p)^T$ represents the p^{th} input sample with p ranges from 1 to P.

P, c_i and w_{ij} represents the total number of samples, centre of the hidden layer nodes and connection weight from hidden layer to output layer respectively. i =1, 2…h, h represents number of hidden layer nodes and y_j is the actual output of the j^{th} output node corresponding with the input sample.

5. Performance Evaluation of Forecast Accuracy

An accurate forecasting is the primary task in uncertain wind speed. The forecasting model is trained with training data and evaluated by using the testing data. The accuracy of the prediction can be measured by the difference between forecasted value and the an observed value. Mean absolute error (MAE), and root mean square error (RMSE) are the most frequently used parameters to evaluate the wind speed forecasting.

$$\mathrm{RMSE}(k) = \frac{\sum_{t=1}^{N}(y(t)-\hat{y}(t))^2}{N} \qquad (29)$$

$$\mathrm{MAE}(k) = \frac{\sum_{t=1}^{N}|y(t)-\hat{y}(t)|}{N} \qquad (30)$$

Where $y(t)$ represent an actual value $\hat{y}(t)$ represents the forecasted value and N represents the number of forecasting used for the evaluation.

6. Results and discussion

Time series meteorological data was collected from the University of Waterloo weather station from 2013 to 2015 with 15 minutes intervals. The meteorological dataset features consists of Shortwave radiation (X1), Shallow soil temperature(X2), Relative humidity(X3), Reflected shortwave radiation(X4),Deep soil temperature(X5), Ambient air temperature(X6), Barometric pressure(X7), Precipitation(GeoNor) (X8), Wind direction(X9), Sonic range sensor for snow depth(X10), Precipitation (Tipping Bucket) (X11),Deep Soil moisture(X12). The data collected in the first two years was considered as training data and the last one year data was utilized as testing data.

Mutual information feature selection finds the significant features to reduce the complexity of the wind speed forecasting model. This feature selection method is meant for determining the minimum number of feature subsets from an original meteorological dataset that maintains a high accuracy to represent the original features of the dataset. This algorithm estimates the weight for each feature. The rank for each feature is calculated based on its weight, and top five features are considered for further evaluation. This algorithm identifies reflected shortwave radiation, wind direction, ambient air temperature, and relative humidity and shortwave radiation. The investigation datasets consisting of selected features produced by the MI feature selection are used as input for RBF, BPN and NARX algorithm to forecast the wind speed. All the three neural network models are evaluated in terms of RMSE and MAE. Figure 3 shows actual and forecast wind speed values using back propagation model with MI features selection. Figure 4 shows actual and forecast wind speed values using radial basis function model with MI feature selection. Figure 5 shows actual and forecast wind speed values using NARX model with MI features selection. Table 1. shows the comparison of forecasting models performance.

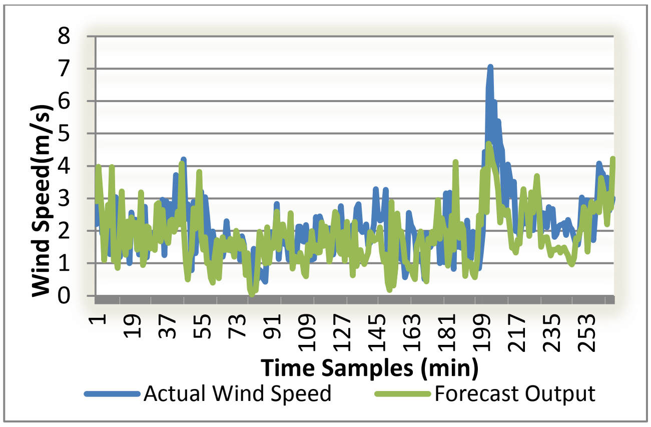

Figure 3. Actual and forecast wind speed output using BPN with feature selection

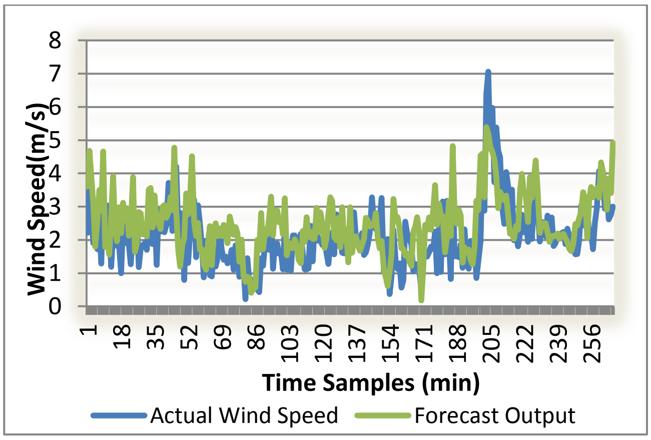

Figure 4. Actual and forecast wind speed output using RBF with feature selection

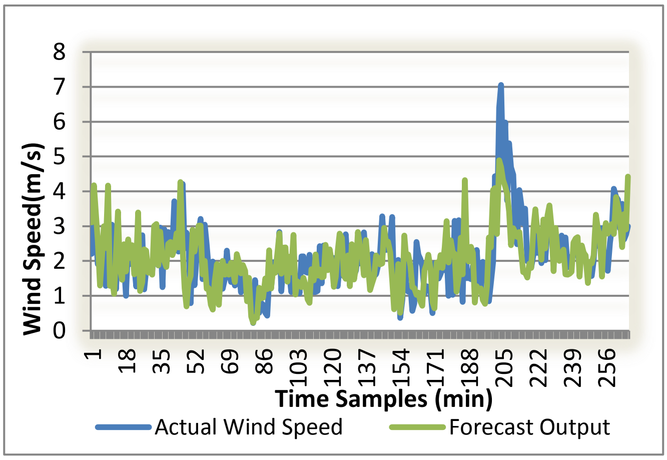

Figure 5. Actual and forecast wind speed output using NARX with feature selection.

Table 1. Comparison of forecasting model accuracy

Wind Speed forecasting model	RMSE	MAE
BPN	1.4854	1.1191
BPN with MIFS	0.8431	0.6437
RBF	1.2963	1.0352
RBF with MIFS	0.7040	0.5732
NARX	1.1859	0.8531
NARX with MIFS	0.5814	0.4381

7. Conclusion

Wind energy is the rapidly growing source of renewable energy. An accurate prediction of wind speed becomes a complicated task due to its uncertain nature. In this paper, MI feature selection method was utilized to identify significant features for improving the wind speed prediction using neural network. The NARX, BPN and RBF neural network models were developed for wind speed prediction and the performance of the models with MI feature selection were compared in terms of RMSE and MAE values. The results show that NARX with MI feature selection model outperforms other models.

References

[1] Zhu, X., & Genton, M. G. (2012). Short-term wind speed forecasting for power system operations. *International Statistical Review*, 80(1), 2-23.

[2] Kusiak, A., Zheng, H., & Song, Z. (2009). Wind farm power prediction: a data-mining approach. Wind Energy: *An International Journal for Progress and Applications in Wind Power Conversion Technology*, 12(3), 275-293.

[3] Soman, S. S., Zareipour, H., Malik, O., & Mandal, P. (2010, September). A review of wind power and wind speed forecasting methods with different time horizons. *In North American power symposium (NAPS)*, 2010 (pp. 1-8). IEEE.

[4] Chang, W. Y. (2014). A literature review of wind forecasting methods. *Journal of Power and Energy Engineering*, 2(4).

[5] Ye, N. (2003). *The Handbook of Data Mining* , Mahwah, NJ/London: Lawrence Erlbaum Associates.

[6] Venugopal, K., Srinivasa, K. G., & Patnaik, L. M. *Soft Computing for Data Mining Applications. Studies in Computational Intelligence*, Springer-Verlag, Berlin Heidelberg, 190, 1-15.

[7] Liu, H., Tian, H. Q., Chen, C., & Li, Y. F. (2010). A hybrid statistical method to predict wind speed and wind power. *Renewable energy*, 35(8), 1857-1861.

[8] Sivanandam, S.N. and Deepa, S.N., 2007. Principles of Soft Computing . John Wiley & Sons

[9] Fadare, D. A. (2010). The application of artificial neural networks to mapping of wind speed profile for energy application in Nigeria. *Applied Energy*, 87(3), 934-942.

[10] Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine learning*, 53(1-2), 23-69.

[11] Kira, K., & Rendell, L. A. (1992, July). The feature selection problem: Traditional methods and a new algorithm. *In Aaai* (Vol. 2, pp. 129-134).

[12] Rana, M., Koprinska, I., & Agelidis, V. G. (2012, November). Feature selection for electricity load prediction. In *International Conference on Neural Information Processing* (pp. 526-534). Springer, Berlin, Heidelberg.

[13] Lin, T., Horne, B. G., Tino, P., & Giles, C. L. (1996). Learning long-term dependencies in NARX recurrent neural networks. *IEEE Transactions on Neural Networks*, 7(6), 1329-1338.

[14] Diaconescu, E. (2008). The use of NARX neural networks to predict chaotic time series. *Wseas Transactions on computer research*, 3(3), 182-191.

[15] Gao, Y., & Er, M. J. (2005). NARMAX time series model prediction: feedforward and recurrent fuzzy neural network approaches. *Fuzzy sets and systems*, 150(2), 331-350.

[16] Xie, H., Tang, H., & Liao, Y. H. (2009, July). Time series prediction based on NARX neural networks: An advanced approach. *In Machine Learning and Cybernetics, 2009 International Conference on* (Vol. 3, pp. 1275-1279). IEEE.

[17] Kumar, S., & Lopez, D. (2015). Feature Selection used for Wind Speed Forecasting with Data Driven Approaches. *Journal of Engineering Science and Technology Review*, 8(5), 124-127.

[18] Li, G., & Shi, J. (2010). On comparing three artificial neural networks for wind speed forecasting. *Applied Energy*, 87(7), 2313-2320.

[19] Doquire, G. and Verleysen, M., (2011),Feature selection with mutual information for uncertain data. *In International Conference on Data Warehousing and Knowledge Discovery* (pp. 330-341). Springer, Berlin, Heidelberg.

[20] Ghiasvand, O. and Ghiasvand, A., (2011),Wind speed short term forecast by neuro fuzzy modeling with aid of Mutual Information at Manjil Wind Power Plant. *Proceedings of the 1st international econference on computer and knowledge engineering (ICCKE)*, pp.1-5.

[21] Babel, M.S., Badgujar, G.B. and Shinde, V.R.,(2015), Using the mutual information technique to select explanatory variables in artificial neural networks for rainfall forecasting. *Meteorological Applications*, 22(3), pp.610-616.

[22] Hosseini, S.H., Moshiri, B., Rahimi-Kian, A. and Araabi, B.N., (2012), April. Traffic speed prediction using mutual information. *IEEE Canadian Conference on Electrical & Computer Engineering (CCECE*), 2012 25th (pp. 1-4).

[23] Wu, X., Hong, B., Peng, X., Wen, F. and Huang, J., (2011), Radial basis function neural network based short-term wind power forecasting with Grubbs test. In Electric Utility Deregulation and Restructuring and Power Technologies (DRPT), 2011 4th International Conference on IEEE, (pp. 1879-1882).

[24] Gao-cheng, C. and Dao-huo, H.U.A.N.G., (2015). Ultra-Short-Term wind speed prediction using RBF Neural Network. *In International Symposium on Computers & Informatics (ISCI 2015)*, (pp. 2441-2448).

[25] Wu, X., Hong, B., Peng, X., Wen, F., & Huang, J. (2011). Radial basis function neural network based short-term wind power forecasting with Grubbs test. In *Electric Utility Deregulation and Restructuring and Power Technologies (DRPT), 4th International Conference on* (pp. 1879-1882). IEEE.

[26] Li, C., Wang, W., Xiong, J., & Chen, P. (2014). Sensitivity analysis for urban drainage modeling using mutual information. *Entropy*, 16(11), 5738-5752.