3

Data Mining/Data Preprocessing Techniques and Error Metrics

Preprocessing techniques play a vital role in the construction of ensemble framework for solar irradiance forecasting. These techniques facilitate a better understanding of underlying patterns in the time-series and improves the forecast accuracy. Data is better understood and subsequent data analysis is performed more accurately and efficiently. A preprocessing also helps in matching the time frequency or the scale. There are other reasons for preprocessing, such as, reduction of features, handling missing information, the effect of seasonality etc. Properly preprocessed solar radiation data reduces complexity of the data significantly and further improves the forecast accuracy. Several data preprocessing algorithms have been used, based on machine learning and decomposition techniques, to characterize irregular datasets. These data preprocessing techniques enable handling small groups of data separately and further improves accuracy of the final forecast by identifying anomalies and neglecting outliers. Sometimes, these preprocessing techniques are used to divide the input data sets into small groups of sub-datasets which are easier to forecast and also leads to better forecast accuracy. A few frequently used preprocessing techniques are presented in Table 3.1 along with their strengths and limitations. To quantify the improvement of forecast accuracy, after using preprocessing techniques, the error metrics are used. It is also used for comparative analysis between the existing and the proposed models.

The chapter is organized as follows. Section 3.1 discusses various data preprocessing techniques and the respective governing equation of solar irradiance forecasting based on these preprocessing techniques. Error metrics used in the thesis are discussed in Section 3.2. Finally conclusions are drown in Section 3.3.

3.1 DATA MINING/DATA PREPROCESSING TECHNIQUES

3.1.1 Normalization

Data with wider or non-homogeneous range can cause imprecise data fitting. If the data is scaled down to smaller range, before fitting into the model, there is a possibility of better precision. One of the frequently used solutions to the problem is normalization. Normalization restricts data series in a range and minimizes the regression error while maintaining correlation among data set. The governing equation for normalization based solar irradiance forecasting can be given as

$$\begin{aligned} X_N(t) &= N(X(t)) \\ \hat{Y}_N(t+h) &= f(X_N(t)) \\ \hat{Y}(t+h) &= N^{-1}(\hat{Y}_N(t+h)) \end{aligned} \tag{3.1}$$

where, X(t) is the original time-series data, $X_N(t)$ is the normalized data, h is the forecast horizon, $\hat{Y}(t)$ is predicted value, N^{-1} represents denormalization back to original scale and f(.) is the predictor function. Here, we present two most frequently used normalization techniques used in the field of renewable energy forecasting.

(i) Min-Max Normalization: Normalization refers to the creation of shifted and scaled versions of statistics. The normalized value obtained by formulation of Min-Max normalization is given by,

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{3.2}$$

where, $\min(x)$ = minimum value of attribute (x) and $\max(x)$ = maximum value of attribute x. This eventually converts the data series in the range of 0 to 1.

(ii) Z-Score Normalization: Z-score normalization is used to convert each feature in the data to have zero mean and unit variance. The calculation of Z-Score normalization starts with the evaluation of the mean and standard deviation for each feature. Subsequently, the normalized values are obtained by subtracting the mean from each feature and dividing this by the standard deviation,

$$x' = \frac{x - \text{mean}(x)}{\text{std}(x)}, \tag{3.3}$$

where, $\text{mean}(x)$ = mean of attribute (x) and $\text{std}(x)$ = standard deviation of attribute (x). This technique is also useful to merge different data sets into one.

3.1.2 Wavelet transform

Fourier Transform (FT) can be used for non-stationary time series, if we are only interested in what spectral components exist in the signal and not interested where these occurs. However, if we want to know, what spectral component occurs at what time (interval), then Fourier transform is not the optimum choice to use. The frequency and time information of a signal at some certain point in the time-frequency plane cannot be known. In other words: We cannot know what spectral component exists at any given time instant. The best we can do is to investigate what spectral components exist at any given interval of time. Higher frequencies are better resolved in time, and lower frequencies are better resolved in frequency. This means that a certain high-frequency component can be located better in time (with less relative error) than a low-frequency component. On the contrary, a low-frequency component can be located better in frequency compared to high-frequency component.

Wavelet transform is capable of providing the time and frequency information simultaneously, hence giving a time-frequency representation of the signal. The wavelet transform possesses good time and frequency resolution simultaneously by using short windows at high-frequency and long windows at low-frequency. Effective applications of wavelet heavily depend on the choice of wavelet family, mother wavelet, and its order. A discrete wavelet transform (DWT) uses successive high pass and low pass filtering operations to decompose the time series into approximations and detailed series for the analysis purpose. The governing equations for DWT based solar irradiance forecasting are given by [Ren *et al.*, 2015]:

$$\begin{aligned}
&[X_{D_i}(t), X_{A_n}(t)] = DWT(X(t)) \\
&\hat{Y}_{D_i}(t+h) = f_1(X_{D_i}(t)) \\
&\hat{Y}_{A_i}(t+h) = f_2(X_{A_i}(t)) \\
&\hat{Y}(t+h) = \sum_{i=1}^{n} \hat{Y}_{D_i}(t+h) + \hat{Y}_{A_i}(t+h)
\end{aligned} \tag{3.4}$$

where, X(t) is the original time series, $X_{D(i)}(t)$ is the ith detailed component, $X_{A_i}(t)$ is the approximation component, h is the forecast horizon, $\hat{Y}(t)$ is the predicted value and $f_1(.)$ and $f_2(.)$ are the predictor functions. The DWT utilizes two set of functions; the scaling function $\phi(t)$ and the wavelet function w(t). The scaling function $\phi(t)$ and the wavelet function w(t) is associated with the low pass and high pass filters respectively with their respective filter coefficients b(n), $n \in Z$

and c(n), $n \in Z$. According to Mallat's algorithm [Mallat, 1989], DWT is realized by means of the filters b[k], c[k] that are related to each other through

$$c[k] = (-1)^{k+1} b[N-k-1],\ (for)\ k = 0, 1, ..., N-1, \tag{3.5}$$

where, N is the length of filter. These filters are constructed recursively from the scaling function $\phi(t)$ and the wavelet function w(t) as,

$$\phi(t) = \sqrt{2} \sum_k b[k] \phi(2t-k), \tag{3.6}$$

$$w(t) = \sqrt{2} \sum_k c[k] \phi(2t-k). \tag{3.7}$$

The output of these filters are sets of low and high-frequency components named as approximations $a_j[k]$ and details $d_j[k]$ respectively. The components are defined as

$$d_{j+1}(k) = \sum_n a_j[n] c(2k-n), \tag{3.8}$$

$$a_{j+1}(k) = \sum_n a_j[n] b(2k-n). \tag{3.9}$$

Here, $a_j[k]$ represents the series which is characterized by slow dynamics (less variability), while $d_j[k]$ represents the local details of time series and have fast dynamics (high variability).

3.1.3 Seasonal and trend decomposition using loess (STL)

Solar time series data has inherent seasonality patterns in it and for the efficient applications of a predictive model, the effect of seasonality needs to be identified and separated out from it. STL helps in detection and separation of seasonality patterns present in the solar time series data. The STL has the capability to handle seasonality components and outliers present in the data. It is applicable to a wide range of time series with different characteristics and sampling frequencies. STL procedure is purely based on numerical methods and does not require any mathematical modelling. This makes the method very easy to implement for a large number of time-series. In addition, it can be easily implemented in the statistical software package (R). The key equations of STL decomposition based forecasting are given in by

$$\begin{aligned} &[X_S(t), X_T(t), X_R(t)] = STL(X(t)) \\ &X_{Residual}(t) = X_T(t) + X_R(t) \\ &\hat{Y}_{Residual}(t+h) = f(X_{Residual}(t)) \\ &\hat{Y}(t+h) = \hat{Y}_{Residual}(t+h) + X_S(t), \end{aligned} \tag{3.10}$$

where, X_S, X_T and X_R are trend seasonal and remainder components respectively, h is the forecast horizon, $\hat{Y}$ is the predicted value and f(.) is the predictor function.
STL is a filtering procedure used for the additive decomposition of the time series into its three constituent components; seasonal, trend and remainder [Cleveland *et al.*, 1990]. In contrast to wavelet decompositions, STL decomposes a time series in the time domain without taking help of any deterministic function. STL filters the data through a sequence of applications of the loess smoother and moving average [Cleveland and Devlin, 1988] which applies locally weighted polynomial regression at each point of the dataset. The loess regression curve is obtained by using some of the nearest previous data as the explanatory variables.

The whole STL procedure is an iterative cycle of de-trending by using loess, that is, local regression polynomial fitting and then updating the seasonal component at each iteration. The

Table 3.1 : Data pre-processing methods, their strength and limitations

Data pre-processing methods	Strengths	Limitations
Normalization	1. Convert data into smaller and uniform range	...
	2. Maintain correlation among dataset	
Wavelet transform	1. Multi resolution analysis of the signal	1. Choice of mother wavelet
	2. Helps in understanding the frequency component present in the signal	2. Pose challenges in identifying the level of decomposition
STL decomposition	1. Decomposes a time series in the time domain without using any deterministic function	1. Not a mathematical model
	2. Can handle outliers in the data	2. The application is limited to seasonal data
	3. Can handle seasonality in the data	
K-means clustering	1. Uses simple principles for identifying clusters which can be explained in non statistical terms	1. It is not guaranteed to find the optimal set of clusters
	2. It is highly flexible	2. requires a reasonable guess to determine the initial number of clusters
	3. Efficient and performs well at dividing data into useful clusters	

iterated cycle is composed of recursive inner and outer loops. Each pass of the inner loop applies seasonal smoothing followed by trend smoothing and updates the seasonal and trend components respectively. An iteration of the outer loop consists of one iteration of the inner loop to estimate trend and seasonal components which are further used for calculation of remainder component. The detailed description about STL and loess polynomial fitting can be found in [Cleveland *et al.*, 1990; Cleveland and Devlin, 1988].

3.1.4 K-means clustering

Clustering permits a dataset to come up with compact groups of data with similar characteristics within the same cluster and isolate from those clusters which contain elements with different characteristics. K-means clustering uses some metrics to define similarity or dissimilarity among samples [Han *et al.*, 2011]. A K-means algorithm is characterized by issues such as initialization of cluster center (centroid), number of clusters (value of K), adopted similarity metric, etc. The main task of K-means clustering is to partition N observations into K homogeneous clusters. In a multidimensional feature space K-means algorithm treats each of its feature value as coordinates. The K-means begins by choosing K points in the feature space to serve as the cluster centres. These centres are treated as a basis that spurs the remaining examples to fall into place. Each observation in a cluster belongs to the cluster with the nearest mean and this mean is also called cluster center. Often, these centres points are chosen by selecting k random examples from the training dataset. Since they are selected at random, these centres could have just as easily been three adjacent points. As the K-means algorithm is highly sensitive to the starting position of the cluster centres, this means that random chance may have a substantial impact on the final set of clusters. To address this problem, K-means can be modified to use different methods for choosing the initial centres. For example, one variant chooses random values occurring anywhere in the feature space (rather than only selecting among the values observed in the data). The similarity between samples is measured by the distance from the samples to the cluster centres. Traditionally, K-means uses Euclidean distance as a measure of dissimilarity, as defined in [Jain *et al.*, 1999; MacQueen *et al.*, 1967], but Manhattan distance or Minkowski distance are also used in many cases. In the field of solar radiation prediction, clustering enables handling small groups of data in place of the whole data, which leads to better forecast accuracy. K-means utilizes unsupervised learning algorithm for data-partitioning in the field. The key equations of clustering based forecasting

techniques are given as

$$\begin{aligned} [X_{Ci}(t_i)] &= K - means(X(t)) \\ X_{Bestcluster} &= CSA(X_{Ci}(t_i)) \\ \hat{Y}(t+h) &= f(X_{Bestcluster}) \end{aligned} \tag{3.11}$$

where, X is the original time series, X_{Ci} is the ith cluster, h is the forecast horizon, CSA represents the cluster selection algorithm, $\hat{Y}$ is the predicted value and f(.) is the predictor.

Till now our emphasis was on the various preprocessing techniques. Forth in the coming section we will look upon the error metrics which are used to quantify the improvement in the forecast accuracy after the use of preprocessing techniques. These error metrics are also used for the comparison purpose of two different forecasting models to decide which model is more suitable for particular dataset.

3.2 ERROR METRICS

The number of metrics for performance evaluation are listed in the literature [Antonanzas *et al.*, 2016; Zhang *et al.*, 2015; Coimbra *et al.*, 2013]. RMSE and MAE error metrics are most commonly used for evaluation of forecast error. The standard error of forecasting is calculated by the root mean square error (RMSE) and the mean absolute error (MAE)

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(X(t) - \hat{Y}(t))^2}{n}}, \tag{3.12}$$

$$MAE = \frac{\sum_{t=1}^{n} \mid X(t) - \hat{Y}(t) \mid}{n}. \tag{3.13}$$

3.2.1 Goodness of fit (R^2)

The coefficient of determination R^2 measures how well forecast values captures the trend in predicted value that is present in observed values. It is interpreted as the proportion of the variance in the predicted values from the measured data values.

$$R^2 = 1 - \frac{Var(\hat{Y} - X)}{Var(X)} \tag{3.14}$$

where, X is the original data series and $\hat{Y}$ is the predicted data series.

3.2.2 Forecast skill (S)

Forecast accuracy depends on weather conditions and forecasts temporal/spatial resolution. Therefore, forecast accuracies are not comparable site-by-site unless normalized by a benchmark. The forecast skill (S) is a way to normalize forecast accuracy [Coimbra *et al.*, 2013; Marquez and Coimbra, 2013].

$$S = 1 - RMSEm/RMSEp \tag{3.15}$$

where, RMSEm is the RMSE value of the model under consideration and RMSEp is the RMSE value of the persistence model. As per the definition of the forecast skill, for perfect solar forecast $S = 1$, and if $S = 0$, the forecast uncertainty is as large as variability. By convention, for persistence model $S = 0$.

3.3 CONCLUSIONS

Preprocessing techniques are often used prior to the prediction of time series data. Various frequently used preprocessing techniques and their characteristics are discussed. This chapter provides insights for applicability that helps in integration of preprocessing techniques with various predictors and develops an ensemble framework. The chapter also covers frequently used error metrics in the field and highlights their applications in estimation of error and comparative analysis among competing candidate forecast model.

…