IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

Data Mining Approach to Wind Data Preprocessing

Oleena Thomas

Assistant Professor, Department of CSE, Kottayam Institute of Technology and Science, Chengalam, Kottayam, Kerala, India

Abstract: Wind energy being a major source of energy has become an interesting area of research. Wind farm power curve monitoring and wind power prediction are the constituent elements of the integrated wind energy research. As absolute modeling of wind source is nearly impossible and as wind turbines are nonlinear, data mining methods are preferred over analytic method to obtain high level information from low level data collected from data acquisition systems. The inputs required are wind power and wind speed magnitudes. Based on that raw wind data are classified into valid or invalid data using unsupervised algorithm. The categorization of data is done into six categories mainly valid, missing, constant, exceeding, irrational and unnatural. Outlier detection is done to filter out the data. Variance and bias are taken into consideration while using different approximation power curve models for data detection. Local Outlier Factor is incorporated, along with the similarity measures used. Weighted distance is calculated to avoid non detection of relevant data points.

Keywords: Data Mining, Wind Data Preprocessing, Wind energy, Local Outlier Factor, Weighted distance.

1 INTRODUCTION

Data preprocessing is one of the major techniques used in the process of data mining. Data mining is highly sensitive to the data being dealt with [3]. So the data collected must be preprocessed to maintain an integrity between considered data. Unsupervised learning is more preferred in data mining because size of the data database to be considered will not stick to the small size. There are cases when large databases are to be considered where supervised learning is not quite practical.

Wind energy integration research generally depends on complex sensors located at remote areas or sites. The generation of high-level synthetic information from databases containing large amounts of low-level data are exposed to factors like possible sensor failures and imperfect input data. Data quality must be maintained, which relies on the data input. To address this problem, an empirical methodology is proposed that can efficiently preprocess and filter the raw wind data using only aggregated magnitudes which are being considered i.e. aggregation of power output and the corresponding wind speed values at the wind farm.

2 LITERATURE REVIEW

Wind farm power curve monitoring and wind power prediction [5]-[6] are the basic constituents of wind energy integration research. As absolute modeling of the wind source is tedious, and because of the high non-linearity of the wind turbine, researchers prefer data-mining methods over analytic methods to gener-ate high-level synthetic information - which could be termed as knowledge- from the low-level data collected by real-time data acquisition systems.

The acquisition and transmission of wind data de-pend on the reliability of sensors. The sensors are located at remote sites exposed to an open, uncontrolled, and even harsh environment, there is a rel-atively high probability of the occurrence of incorrect data.

On the other hand, unnatural operating states of a wind farm cause unnatural data. Wind curtailment is an example. Due to the congestion or load balancing purposes or wind turbine shutdown be-cause of mechanical faults or maintenance wind curtailment possibilities are likely to occur.

This could re-sult in unnatural data, which have normal wind speed and abnormal wind power output below the theoretical values corresponding to the wind speed. Both incorrect and unnatural data affect the performance of the data-based research, as data-mining methods are highly sensitive to data quality. The detection of incorrect and

DOI10.17148/IJARCCE.2017.6866

unnatural data thus is of greater importance. Such data should be preprocessed [4] before the integration studies. Different approaches have been proposed to address preprocessing problems for various types of data including load data, remote terminal unit (RTU) data, geophysical data, fingerprint image data, and photo-voltaic data.

The raw wind data preprocessing is done in different steps to ensure more effectiveness. Raw wind data properties are analysed, and all the data are divided into six categories according to magnitudes of their at-tributes from a statistical perspective. The weighted distance, a novel concept of the degree of similarity between the individual objects in the wind database and the local outlier factor (LOF) algorithm, is incor-porated to compute the outlier factor [9]-[11] of every individual object, and this outlier factor is then used to assess which category an object belongs to. The valid data is then subjected to graphical plotting.

3 PROPOSED APPROACH

Wind data preprocessing method may include va-lidity check, data scaling, missing data processing and lag removal. The validity check involves a data range check that detects data values exceeding the physical limits. Data scaling normalizes data with the ratings. Missing data processing involves either neglecting or approximating the missing values. Lag removal uses the cross correlation function to identify the lag be-tween input and output, which is useful when dealing with time-series analysis.

However, many approaches do not consider unnat-ural data. The conclusion of unnatural data is done by classifying the raw data according to the magnitude of the wind speed and the wind power output. A neural network classifier was trained with the wind speed and wind power as input and the classification result as output. Then, the neural network was used to classify more data. As a type of supervised learning algorithm, this method can achieve relatively high accuracy as long as the classification result is accurate, i.e., the correct class is precisely determined for every single data point.

In real-world applications, artificial judgment is lim-ited and inconvenient when the size of the database is large, and the wind farm operation state records are often unavailable. Thus, these data classification procedures are unfeasible or unreliable, which causes difficulties in applying supervised learning algorithms.

Therefore, the alternative solution is to use unsupervised algorithms. To use unsupervised algorithms, an unsupervised learning approach is adopted based on the local outlier factor (LOF)-identifying algorithm introduced by Breunig. The LOF of every data point is computed using a novel concept of the degree of similarity among the individual data points, and hence invalid data are detected as abnormal outlier factors.

Analysis was conducted on the data set that was collected from the wind farm meteorological mast in the period of 01 January 2012 to 03 October 2015. The only parameters needed were wind speed and wind power.

3.1 Wind Data Properties

The constant data, missing data, and exceeding data can easily be detected by the different methods proposed already [1]. Among the remaining data [the irrational, unnatural, and valid (IUV) data], the un-natural data and the irrational data should be given more attention. To study the distribution of the different data categories, the raw wind data in a complete calendar year from the same wind farm have been analysed. The raw wind data distribution histogram (shown in Fig.1) indicates that the number of invalid data points (including the irrational and the unnatural data) is much smaller than the number of valid data points, even in winter, when more wind curtailments result in more unnatural data [see Fig. 1(a)].

Within the areas of the unnatural and the irrational data, the density is considerably lower than the density in the area of the valid data. Both the unnatural and the irrational data can be considered outliers or noise compared to the valid data. Therefore, outlier detection, which tries to identify exceptional cases that deviate substantially from the majority patterns, can be used to exclude the unnatural and the irrational data. Furthermore, from the simplicity point of view, as a type of unsupervised learning, outlier detection [2] can learn relationships and structure from the attributes of the data themselves, so that the classification steps in earlier methods are no longer necessary.

In the raw wind data preprocessing technique, raw wind data is initially considered and is then considered to filter out the constant data which is then followed by filtration of missing data. Then physical range for the obtained data is checked. All unfit data are filtered out. Then data scaling is performed. The filtering of unnatural and irrational data is done which ultimately results in filtered data.Fig.2 shows the structure of the preprocessing method.

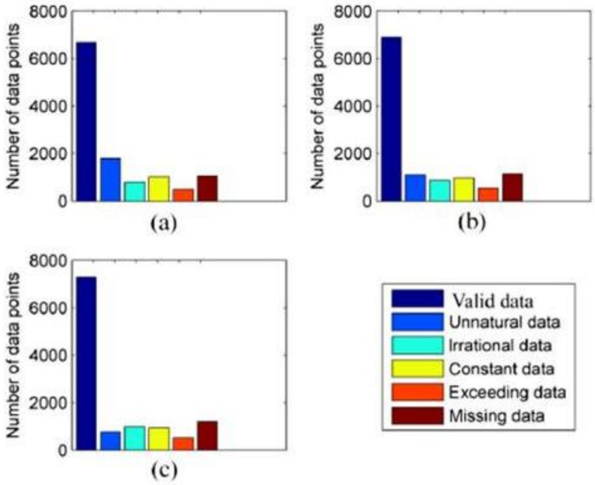


Figure 1: Distribution of raw wind data: (a) period from 10/1/2010 to 1/31/2011; (b) period from 2/1/2011 to 5/31/2011; and (c) period from 6/1/2010 to 9/30/2011.

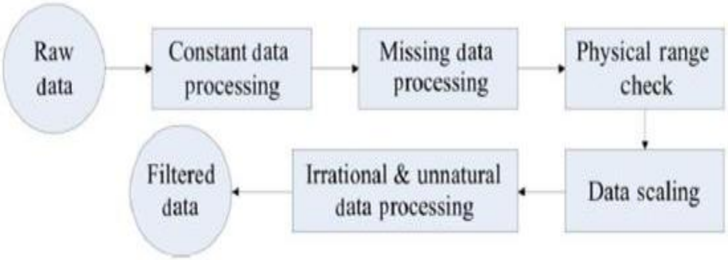


Figure 2: Structure of preprocessing system

3.2 Preprocessing of Wind Data

The constant data processing block, the missing data processing block, and the physical range check block can easily be implemented via several if then judgment sentences. Regarding imputation of the in-valid data, the major imputation approaches are to fill or predict the missing values based on the nearby observed values. However, because the invalid data are often consistent for a relatively long time, there are insufficient data to make a smooth imputation, which may only introduce more incorrect data to the database. Moreover, there is a large amount of data available, so we can obtain sufficiently interesting pat-terns from the remaining data that the effect of the pattern losses with the removal of the invalid data is limited. Therefore, no approximation is performed after removing the invalid data.

Data scaling is performed by applying the following equation: $_x$ = x=x_r where x is a variable, x_r is the rating, and $_x$ is the normalized value of the variable. Similarly, we use the bar notation to denote the normalized value of a variable. For the irrational and the unnatural data processing block, the LOF identifying algorithm is applied.

3.3 Local Outlier Factor

Outlying is not a binary property. Instead each ob-ject is assigned with an outlier [8] factor depending on the degree of outlierness. The LOF values are calculated for each object based on different notions. The notions considered are as follows:

1) k-distance of object: For any positive integer k, the k-distance of object p, denoted as k-distance (p), is defined as the distance d(p,o) between p and an object o €D such that:

(i) For at least k objects o € D:{p}it holds that d(p; q) <= d(p; o) , and

(ii)For at most k-1 objects o €

D{p} it holds thatd(p; q) < d(p; o) .

2) k-distance neighborhood of object: Given the k-distance of p, the k-distance neighborhood of p contains every object whose distance from p is not greater than the k-distance, i.e.$N_{k \text{ distance}(p)}$(p) = fq 2 D/fpg|d(p; q) k distance(p)g. These objects q are called the k-nearest neighbors of p.

3) Reachability distance of object: Let k be a natural number. The reachability distance of object p with respect to object o is de ned as reach dist$_k$(p; o) = maxfk distance(o); d(p; o)g

Fig. 3 illustrates the idea of reachability distance with k = 4. Intuitively, if object p is far away from o (e.g. p2 in the figure), then the reachability distance between the two is simply their actual distance. How-ever, if they are sufficiently close (e.g., p1 in the figure), the actual distance is replaced by the k-distance of o. The reason is that in so doing, the statistical fluctuations of d(p,o) for all the p's close to o can be significantly reduced. The strength of this smoothing effect can be controlled by the parameter k. The higher the value of k, the more similar the reachability distances for objects within the same neighborhood.

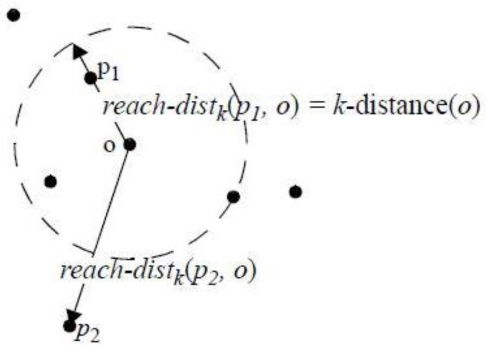

p_1
$reach\text{-}dist_k(p_1, o) = k\text{-distance}(o)$
o
$reach\text{-}dist_k(p_2, o)$
p_2

Figure 3: Reachability distance

4) Local reachability density of object: The local reachability den-sity of y is de ned as Lrd$_{MinP\ ts}$(x) = ^{jN}MinP ts$^{(x)j=}$ yNMinP ts(x)^{reachdist}MinP ts(x;y)$^{\cdot}$

The local reachability density of an object p is the inverse of the average reachability distance based on the MinPts nearest neighbors of p. The local density can be 1 if all the reachability distances in the summation are 0. This may occur for an object p if there are at least MinPts objects, different from p, but sharing the same spatial coordinates, i.e. if there are at least MinPts duplicates of p in the dataset.

Local Outlier Factor:The local outier factor of p is defined as LOFMinP ts(p) = $^{(}$o2N$_{MinP\ ts(p)}$ ^{lrd}MinP ts$^{(o)=lrd}$MinP ts$^{(p))=}$j^{N}MinP ts(p)j

The outlier factor of object p captures the degree to which we call p an outlier. It is the average of the ratio of the local reachability density of p and those of ps MinPts-nearest neighbors. It is easy to visualize that the lower p's local reachability density is, and the higher the local reachability densities of p's MinPts-nearest neighbors are, the higher is the LOF value of p.For most objects in a cluster, the outlier factors are approximately equal to 1. For the outliers, the outlier factors are larger than 1. We can generally de ne an LOF-threshold value, which is determined by trial and error to obtain the best performance. The threshold used is 1.1 here. Objects with outlier factors greater than the LOF-threshold value are outliers.

3.4 Similarity Measurement

LOF algorithm should be incorporated with distance formula to obtain the distance between the datapoints in the datasets so as to obtain the degree of outlierness. Many distance formula has been put forward to calculate the distance between the desired points. Each formula has its own specification.

One such case is the Euclidean distance formula. With the use of Euclidean distance formula the dis-tance between the datapoints are calculated and the test results evidently show that with the use of Eu-clidean distance along with the LOF algorithm, the unnatural datasets remain undetected. This is a se-rious drawback which must be attended. Euclidean distance formula fails with unnatural data because, the formula do not take into consideration the greater impact of wind power dimension.

As the Euclidean formula shows d(x; y) =

q $(\overline{V}(x)\overline{V}(y))^2 + (\overline{P}(x)\overline{P}(y))^2$ Where $\overline{V}$(x)denotes

wind speed dimension and P (x) denotes the wind power dimension both the wind speed and wind power dimensions are taken equally. But in real, the wind power dimension has more impact over the wind speed dimension.

DOI10.17148/IJARCCE.2017.6866

Based on the inference from Euclidean distance measure, a new concept in distance calculation is put forward which is called weighted distance [1]. It mea-sures the similarity between objects, where the outlier attribute is assigned a larger weight. To determine a proper form of the weight, prior knowledge about the wind power curve should be considered. Wind power curve [6] has got three region theoret-ically as shown in Fig. 4. The points corresponding to the valid data are distributed near the power curve, whereas the points corresponding to the unnatural and the irrational data are far away.

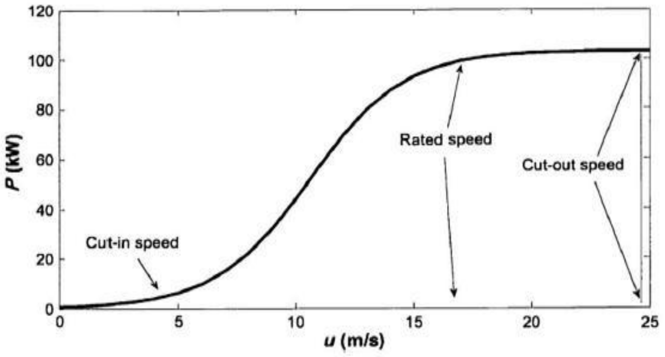


Figure 4: Wind turbine power curve

Therefore, the weight can be formulated based on the difference between the measured and the true value of wind power, as in Fig. 5 where V_{ci}; V_r; andV_{co} represent the cut-in, rated, and cut-out speed of the wind turbine. P $_T$ denotes the normalized true value of wind power.

1) When $V_{ci} \leq V < V_r$,

$$\omega = \begin{cases} 1, & \left|\bar{P}_T - \bar{P}\right| \leq 0.1 \\ \left|\bar{P}_T - \bar{P}\right| /0.1, & \left|\bar{P}_T - \bar{P}\right| > 0.1. \end{cases}$$

2) When $V < V_{ci}$ or $V_r \leq V < V_{co}$,

$$\omega = \begin{cases} 1, & \left|\bar{P}_T - \bar{P}\right| \leq 0.05 \\ \left|\bar{P}_T - \bar{P}\right| /0.05, & \left|\bar{P}_T - \bar{P}\right| > 0.05. \end{cases}$$

3) When $V \geq V_{co}$,

$$\omega = 1$$

Figure 5: Weight dimensions

An object that is close to the power curve is defined as being located in the [P $_T$ - 0.1, P $_T$ + 0.1] interval whenV_{ci} V < $V_r$and in the [P $_T$ - 0.05, P $_T$ + 0.05] interval when V < V_{ci} or V V_r, based on domain experiences and past studies. The weights assigned to these objects are 1, identical to the Euclidean distance. Additionally, the weights of the data whose wind speed values are larger than the cut-out speed, are also equal to 1, to extract the natural properties from the wake effects data. The weight of the other data is larger than 1, in proportion to the difference between the measured and the accurate value of the wind power. The farther away an object is located, the greater the weight, and the more likely the object is to be detected as an outlier. With the introduction of the weighted norm, the equation for calculating the weight between datapoints can be calculated from the modified Euclidean distance as

q
d(x; y) = (V (x)V (y))2 + !T (P (x)P (y))2. T 0 is a tuning parameter, to be determined separately.

4 EXPERIMENTAL RESULTS

The e effectiveness of the weighted distance is cal-culated while classifying the raw wind data. The weighted distance shows that the degree of oulierness can be clearly specified. This is far better when com-pared with the Euclidean distance. The equation ap- plied is d(x; y) = (V (x)V (y))^2 + !^{T} (P (x)P (y))^2. T 0 is a tuning parameter, to be determined separately.

DOI10.17148/IJARCCE.2017.6866

Table 1: Valid Wind Data Classification

	Speed(mps)	Power(kW)	Euclidean	Weighted
Missing	4.25	NaN	Invalid	Invalid
	4.892	NaN	Invalid	Invalid
	5.369	NaN	Invalid	Invalid
	6.324	NaN	Invalid	Invalid
Constant	13.376	139684.344	Invalid	Invalid
	13.376	139684.344	Invalid	Invalid
	13.376	139684.344	Invalid	Invalid
Irrational	0.000	120150.322	Invalid	Invalid
	0.000	120000.421	Invalid	Invalid
	0.000	134211.982	Invalid	Invalid
Exceeding	5.890	9999999.9	Invalid	Invalid
	10.345	-999999.99	Invalid	Invalid
	12.212	9999999.22	Invalid	Invalid
Unnatural	13.778	15000.78	Invalid	Valid
	11.890	13622.431	Invalid	Valid
	13.842	1076.200	Invalid	Invalid

5 OBSERVED DATA

The observed data are plotted along the graph and then the efficiency of the farm is obtained. Fig. 6 shows the scatter plot of data.

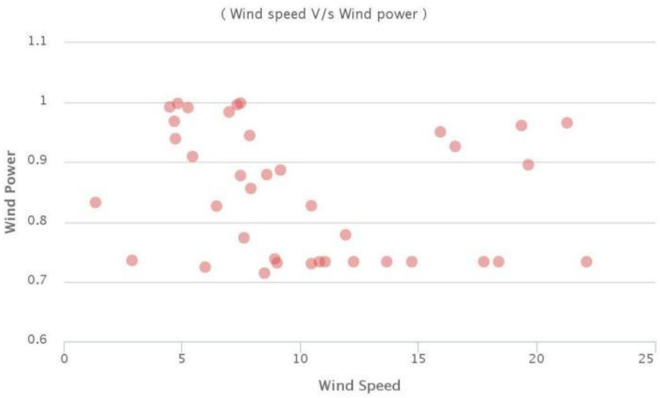


Figure 6: Scatter plot of valid data

From the graph, it can be inferred that the real data set deviate a lot from the ideal curve because of the is-sues of practicality. The raw wind data is processed for constant data initially which is followed by processing of missing, exceeding scaling and irrational and un-natural data. Finally, the filtered output is obtained which is plotted in the graph.

5.1 Performance Analysis

The raw wind data was initially preprocessed. This was then followed by the classification accuracy tests. In the outlier detection test, the Euclidean method was employed first. In this method, both the wind speed and wind power parameters were given the same importance. So the test resulted in misclassification of the unnatural data as valid data. This happened because the Euclidean method considered wind speed and power with equal impact. In case of unnatural data, the wind speed values remain original but the wind power values will not be in respondent to the speed values.

The weighted Euclidean distance concept was the next method. In this method, the wind power values were given more weightage than that of wind speed values. Thus the unnatural data were classified cor-rectly as invalid data.

6 SUMMARY AND CONCLUSIONS

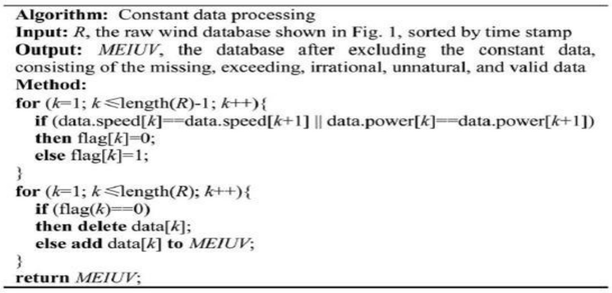
Raw wind data properties were analysed as per the data acquired from the wind mill which act as the source of the data. Invalid data can be categorized into five types. A wind data preprocessing method-ology has been adopted to classify and prioritize the data. Because identifying the unnatural and the ir-rational data is challenging, they are treated as outliers and uses the LOF algorithm to detect and remove these outliers. To incorporate prior knowledge regarding the wind data, a new type of similarity measurement is designed and applied in the algorithm. Numerical experiments have verified the e effectiveness of the algorithm and the similarity measurement. The performance evaluation of the algorithm is tested against the collected set of data.

Being a type of unsupervised learning algorithm is one of the greatest advantages. Therefore, it can detect and classify the raw data using solely the at-tributes of the data themselves. It is easier and more convenient to perform in practice, especially when the operation records are not available. However, as there is no universal data-mining algorithm that can handle all problems, this methodology has its limitations. First, the total number of the data points should not be too small. An empirical minimum value is approximately 1000. Second, if most of the data are invalid, the accuracy cannot be guaranteed. This situation indicates that either the data acquisition and trans-mission system is broken down or manual actions are frequent. In short, the wind farm is faulty, and the data acquired from it should not be used for research.

As a matter of the future enhancement works needs to be done in the area regarding the weighted distance. The idea of weighted distance could be incorporated in outlier detecting applications other than wind data related applications. This concept could be used in the wake of cluster identification applications as well. Thus future research should focus on the outlier or cluster detection using the weighted distance used in this application.

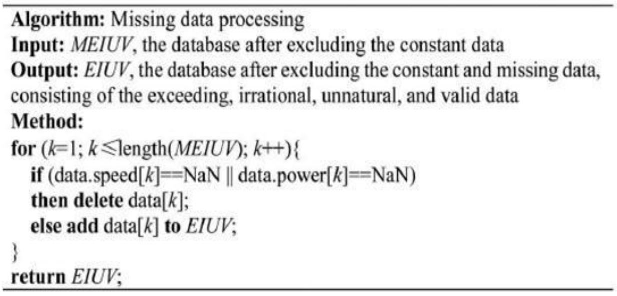
APPENDIX

The following algorithms describes the preprocessing of constant, missing and exceeding types of wind data.



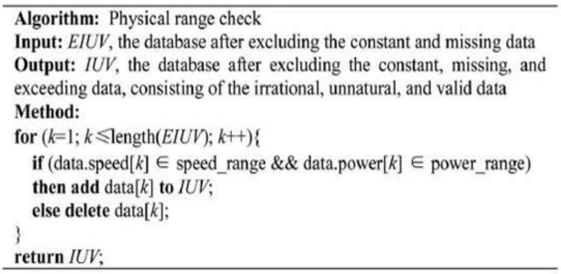
```
Algorithm: Constant data processing
Input: R, the raw wind database shown in Fig. 1, sorted by time stamp
Output: MEIUV, the database after excluding the constant data,
consisting of the missing, exceeding, irrational, unnatural, and valid data
Method:
for (k=1; k⩽length(R)-1; k++){
   if (data.speed[k]==data.speed[k+1] || data.power[k]==data.power[k+1])
   then flag[k]=0;
   else flag[k]=1;
}
for (k=1; k⩽length(R); k++){
   if (flag(k)==0)
   then delete data[k];
   else add data[k] to MEIUV;
}
return MEIUV;
```

Figure 7: Constant data processing algorithm



```
Algorithm: Missing data processing
Input: MEIUV, the database after excluding the constant data
Output: EIUV, the database after excluding the constant and missing data,
consisting of the exceeding, irrational, unnatural, and valid data
Method:
for (k=1; k⩽length(MEIUV); k++){
   if (data.speed[k]==NaN || data.power[k]==NaN)
   then delete data[k];
   else add data[k] to EIUV;
}
return EIUV;
```

Figure 8: Missing data processing algorithm



```
Algorithm: Physical range check
Input: EIUV, the database after excluding the constant and missing data
Output: IUV, the database after excluding the constant, missing, and
exceeding data, consisting of the irrational, unnatural, and valid data
Method:
for (k=1; k⩽length(EIUV); k++){
   if (data.speed[k] ∈ speed_range && data.power[k] ∈ power_range)
   then add data[k] to IUV;
   else delete data[k];
}
return IUV;
```

Figure 9: Exceeding data processing algorithm

DOI10.17148/IJARCCE.2017.6866

REFERENCES

[1] Le Zheng, Wei Hu, and Yong Min, Raw Wind Data Preprocessing: A Data-Mining Approach," IEEE Transactions On Sustainable Energy, vol. 6, no. 1, Jan. 2015.
[2] N. Zhang and W. F. Lu, An Efficient Data Pre-processing Method for Mining Customer Survey Data," in Proc. IEEE Renew. Energy, vol. 34, no. 6, Jun. 2009.
[3] Pramod Kumar, V. K. Chandna, and Mini S. Thomas, Intelligent Algorithm for Preprocess-ing Multiple Data at RTU," in IEEE Trans. Power Systems, vol. 18, no. 4, Nov. 2003.
[4] Paulo M. Goncalves Jr. and Roberto S. M. Barros, Automating Data Preprocessing with DMPML and KDDML," in IEEE Trans. Knowl-edge Engineering , vol. 4, no. 3, Jul. 2011.
[5] Ziqiao Liu, Wenzhong Gao, Yih-Huei Wan and Eduard Muljadi, Wind power plant prediction by using neural networks," in IEEE Trans. Energy Conversion, Arlington, VA, USA, 2006.
[6] M. Lydia, A. I. Selvakumar, S. S. Kuma, and G. E. P. Kumar, Advanced algorithms for wind turbine power curve modeling," IEEE Trans. Sustain. Energy, vol. 4, no. 3, Jul. 2013.
[7] Barry P. Hayes, Irinel-Sorin Ilie, Antonios Por-podas, Sasa Z. Djokic,and Gianfranco Chicco, Equivalent power curve model of a wind farm based on field measurement data," IEEE Trans-actions On Power Systems, vol. 37, no. 1, 2011.
[8] M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, LOF: Identifying density-based local outliers," Proc. Int. Conf. Manage. Data, 2000.
[9] Emmanuel Muller, Ira Assent, Uwe Steinhausen and Thomas Seidl, OutRank: ranking outliers in high dimensional data," in IEEE Trans. Knowledge Engineering,vol. 35, no. 5, Sep. 2007.
[10] Yunxin Tao and Dechang Pi, Unifying Density-Based Clustering and Outlier Detection," in IEEE Trans. Knowledge Discovery, vol. 5, no. 1, Jan. 2005.
[11] Amineh Amini, Teh Ying Wah, Mahmoud Reza Saybani and Saeed Reza Aghabozorgi Sahaf Yazdi, A Study of Density-Grid based Clus-tering Algorithms on Data Streams," in IEEE Trans. Energy Conversion, vol. 24, no. 1, Mar. 2009.

BIOGRAPHY

Oleena Thomas, working currently as Assistant Professor CSE in Kottayam Institute of Technology and Science, Kottayam has pursued her M. Tech in CSE from Amal Jyothi College of Engineering Kottayam and B. Tech from College of Engineering Kottarakkara. She has published papers named “Literature Analysis on Reputation Models for Feedback in E- commerce” and “Automated Social Media Mining System in Health Care” in IJARCCE.

DOI10.17148/IJARCCE.2017.6866