A Survey On Wind Data Pre-Processing In Electricity Generation

Mahima Susan Abraham[1] and Jiby J Puthiyidam[2]

[1] Department of Computer Science and Engineering, College of Engineering, Poonjar
[2] Department of Computer Science and Engineering, College of Engineering, Poonjar

ABSTRACT

Wind energy integration research generally relies on complex sensors located at remote sites. The procedure for generating high-level synthetic information from databases containing large amounts of low-level data must therefore account for possible sensor failures and imperfect input data. The data input is highly sensitive to data quality. To address this problem, this paper presents an empirical methodology that can efficiently preprocess and filter the raw wind data using only aggregated active power output and the corresponding wind speed value sat the wind farm. First, raw wind data properties are analyzed, and all the data are divided into six categories according to their attribute magnitudes from a statistical perspective. Next, 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 incorporated to compute the outlier factor of every individual object, and this outlier factor is then used to assess which category an object belongs to.

KEYWORDS

Data mining, data preprocessing, local outlier factor (LOF), unsupervised learning

1. Introduction

All The objective of data mining is to identify valid novel, potentially useful, and understandable correlations and patterns in existing data .Finding useful patterns in data is known by different names (including data mining) in different communities (e.g., knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing).The term “data mining” is primarily used by statisticians, database researchers, and the MIS and business communities. The term Knowledge Discovery in Databases (KDD)[7] is generally used to refer to the overall process of discovering useful knowledge from data, where data mining is a particular step in this process. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, and proper interpretation of the results of the data mining process, ensure that useful knowledge is derived from the data.

Data mining, almost by definition, is primarily concerned with the operational. The second type of data mining approach, pattern detection, seeks to identify small (but nonetheless possibly important) departures from the norm, to detect unusual patterns of behaviour. Examples include unusual spending patterns in credit card usage (for fraud detection), sporadic waveforms in EEG traces, and objects with patterns of characteristics unlike others. It is this class of strategies that led to the notion of data mining as seeking “nuggets” of information among the mass of data. In

DOI: 10.5121/ijci.2016.5243

general, business databases pose a unique problem for pattern extraction because of their complexity. Complexity arises from anomalies such as discontinuity, noise, ambiguity, and incompleteness. And while most data mining algorithms are able to separate the effects of such irrelevant attributes in determining the actual pattern, the predictive power of the mining algorithms may decrease as the number of these anomalies increase.

Data pre-processing [8] is an often neglected but important step in the data mining process. The phrase "Garbage In, Garbage Out" is particularly applicable to data mining and machine learning. Data gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: -100), impossible data combinations (e.g., Gender: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set.

Raw data is highly susceptible to noise, missing values, and inconsistency. The quality of data affects the data mining results. In order to help improve the quality of the data and, consequently, of the mining results raw data is pre-processed so as to improve the efficiency and ease of the mining process. Data pre-processing is one of the most critical steps in a data mining process which deals with the preparation and transformation of the initial dataset. Data pre-processing methods are divided into following categories:

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

Nowadays, more and more attention is paid on wind energy—a kind of clean and renewable energy. While developing the wind power, we should also keep a watchful eye on the ability of real-time data processing, in order to make the wind power develop healthily and rapidly, taking the road of sustainable development is the ultimate goal. Now wind power data is usually applied in wind power prediction, which is benefit for reducing the shock of wind power on the grid, and improving the economy of the grid operations. As we know, effective wind data pre-processing is the key to wind power forecasts. Provide clean, accurate data for, data mining, thus reduce the amount of data processing, and then deduce the valuable information. Although there are a lot of methods of data pre-processing, few of which apply in the wind power data. Wind data reprocessing existed mainly study on attribute reduction, missing values, isolated points, but offer few reference value for prediction.

Wind energy integration research generally relies on complex sensors located at remote sites. The procedure for generating high-level synthetic information from databases containing large amounts of low-level data must therefore account for possible sensor failures and imperfect input data. The data input is highly sensitive to data quality. To address this problem, this paper presents an empirical methodology that can efficiently pre-process and filter the raw wind data using only aggregated active power output and the corresponding wind speed values at the wind farm.

RELATED WORKS

There are many methods in analysing wind data. Some of those methods are discussed. Most of these methods don't consider irregular data's or values. One of these methods identifies the irregular data's or values and removes them. Some methods improve accuracy and some improves performance speed. . It's discussed below in details

1. RAW WIND DATA PREPROCESSING

[1] presented a wind data pre-processing method including four steps: 1) validity check; 2) data scaling; 3) missing data processing; and 4) 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 between input and output, which is useful when dealing with time-series analysis. In real-world applications, artificial judgment is limited 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 infeasible or unreliable, which causes difficulties in applying supervised learning algorithms. Therefore, the alternative solution is to use unsupervised algorithms. To use unsupervised algorithms, we adopted an unsupervised learning approach based on the local outlier factor (LOF)-identifying algorithm. 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.

The contribution of this paper is to develop an empirical methodology for raw wind data pre-processing. The only information required for this methodology is the aggregated wind power output of the wind farm collected from the Supervisory Control And Data Acquisition (SCADA) system, which is available at the dispatch center, and the wind speed magnitude data at the corresponding wind farm site. The availability of wind farm operation state records or wind turbine fault logs (which are not recorded or stored by most wind farm operators) will help improve the accuracy of the methodology. If these data are unavailable, this is often the case, the methodology proposed in this paper has nonetheless been proved to be adequate for the situation.

Advantage: One of the greatest advantages of the proposed methodology is that it is a type of unsupervised learning algorithm. Therefore, it can detect and classify the raw data using solely the attributes of the data themselves. It is easier and more convenient to perform in practice, especially when the operation records are not available.

Disadvantage:. 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 transmission 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. The data pre-processing method proposed in this paper can be used for many purposes, not only wind-related applications. The idea of weighted distance can also be used in other outlier or cluster-detection algorithms to develop individual detection algorithms dedicated to specific applications.

2. WIND SPEED FORECASTING STRATEGY

[2] presented a new wind speed forecasting approach based on based on the chaotic time series modelling technique and the Apriori algorithm has been developed. The new approach consists of four procedures: Clustering by using the k-means clustering approach; Employing the Apriori algorithm to discover the association rules; Forecasting the wind speed according to the chaotic time series forecasting model; and Correcting the forecasted wind speed data using the associated rules discovered previously. This procedure has been verified by 31-day-ahead daily average wind speed forecasting case studies, which employed the wind speed and other meteorological data collected from four meteorological stations located in the Hexi Corridor area of China. The results of these case studies reveal that the chaotic forecasting model can efficiently improve the accuracy of the wind speed forecasting, and the Apriori algorithm can effectively discover the association rules between the wind speed and other meteorological factors. In addition, the correction results demonstrate that the association rules discovered by the Apriori algorithm have powerful capacities in handling the forecasted wind speed values correction when the forecasted values do not match the classification discovered by the association rules.

This paper firstly analyses the historical wind speed data for a given wind farm by applying the nonlinear time series modelling techniques. The numerical simulations results indicate that chaotic characteristics obviously exist in the wind speed time series. This finding inspires us to model the wind speed as a complex non-linear dynamic system that often exhibits chaotic behaviour. If an irregular movement characterized by the time series can be regarded as a type of chaos phenomenon, a prediction with higher precision is available with the chaos theory, which is used to address the inner uncertainties of the system. As an important method for studying the characteristics of complex systems, the chaotic time series predictions have attracted significant research interests over the past few years. Some chaotic prediction methods have been developed such as the local-region method, Lyapunov Exponents method, and artificial neural network method. Among these methods, the local-region method seems more promising for wind speed forecasting. This paper employs a weighted local-region method to forecast the wind speed series.

Advantage: This can be very useful in two occasions: one is to check the predicted wind speed values for abnormal cases based on the association rules, and the other is to estimate the value ranges of the other meteorological factors, including the air pressure, air temperature and humidity.

3. WIND SPEED AND POWER FORECASTING

[3] presents an overview of existing research on wind speed and power forecasting. It first discusses state-or-the-art wind speed and power forecasting approaches. Then, forecasting accuracy is presented based on variable factors. Finally, potential techniques to improve the accuracy of forecasting models are reviewed. A full survey on all existing models is not

presented, but attempts to highlight the most promising body of knowledge concerning wind speed and power forecasting. Wind power is one of the most rapidly growing renewable energy sources, and is regarded as an appealing alternative to conventional power generated from fossil fuel. This led to a collaborative effort to achieve 20% of U.S. electricity supplied from wind power by 2030 . Although the integration of wind power brings many advantages, high penetration of wind power provides a number of challenges in power system operations and planning, mainly due to its uncertain and intermittent nature.

In the electricity system the power supply must be equal to the power demand at all times. However, the variation of wind power output makes it difficult to maintain this balance. One of the possible solutions to the balance challenge is to improve the wind speed and power forecasting. Research in the area of forecasting wind speed or the power produced by wind farms has been devoted to the development of effective and reliable tools and many different approaches have been proposed and reviewed in . Accurate forecasting tools reduce operating costs and improve reliability associated with the integration of wind power into the existing electricity supply system. There are different users of wind speed and power forecasts. These users not only need point forecasts but also the uncertainty of the forecast is essential for determining the size of the operating reserves necessary to balance the generation with load.

The main objectives of wind speed and power forecasting is to estimate the wind speed and power as quickly and accurately as possible. Accurate forecasting tools reduce the financial risk and lead to improved scheduling and unit commitment plans. The statistical approaches provide good results in the majority of cases, including short-term, medium-term, and long-term forecasting. However, in the very short-term and short-term horizon, the influence of atmospheric dynamics becomes more important, so that the use of the physical approaches becomes necessary.

Advantage: This approach is able to not only improve the forecast accuracy, but also reduces the risk from extreme events. Ensemble forecasting models for probabilistic forecasting are used in order to obtain the expected spread of weather conditions and assess the probability of particular weather events.

4.PROBABILISTIC WIND SPEED FORECASTING

[4] presents a probabilistic forecasts of wind speed are becoming critical as interest grows in wind as a clean and renewable source of energy, in addition to a wide range of other uses, from aviation to recreational boating. Statistical approaches to wind forecasting offer two particular challenges: the distribution of wind speeds is highly skewed, and wind observations are reported to the nearest whole knot, a much coarser discretization than is seen in other weather quantities. The prevailing paradigm in weather forecasting is to issue deterministic forecasts based on numerical weather prediction models.

Uncertainty can then be accessed through ensemble forecasts, where multiple estimates of the current state of the atmosphere are used to generate a collection of deterministic predictions. Ensemble forecasts are often uncalibrated, however, and Bayesian model averaging (BMA) is a statistical way of post processing these forecast ensembles to create calibrated predictive probability density functions (PDFs).It represents the predictive PDF as a weighted average of PDFs centered on the individual bias-corrected forecasts, where the weights reflect the forecasts' relative contributions to predictive skill over a training period. In this paper we extend BMA to provide probabilistic forecasts of wind speed, taking account of the skewness of the predictive distributions and the discreteness of the observations.

Advantage: better calibration

5. MARKOV-SWITCHING AUTOREGRESSIVE MODELS

[5] presentsWind power production data at temporal resolutions of a few minutes exhibits successive periods with fluctuations of various dynamic nature and magnitude, which cannot be explained (so far) by the evolution of some explanatory variable. Our proposal is to capture this regime-switching behaviour with an approach relying on Markov-Switching Autoregressive (MSAR) models. An appropriate parameterization of the model coefficients is introduced, along with an adaptive estimation method allowing to accommodate long-term variations in the process characteristics. The objective criterion to be recursively optimized is based on penalized maximum-likelihood, with exponential forgetting of past observations. MSAR models are then employed for 1-step-ahead point forecasting of 10-minute resolution time-series of wind power at two large offshore wind farms. They are favourably compared against persistence and Autoregressive (AR) models.

The main objective of the present paper is to introduce a MSAR model whose coefficients are adaptively and recursively estimated, with application to the modelling and forecasting of offshore wind power fluctuations. The parameterization of the model coefficients employed here is inspired by those initially proposed .Adaptively in time is achieved with exponential forgetting of past observations. In addition, the formulation of the objective function to be minimized at each time-step includes a regularization term that permits to increase the generalization ability of estimated models, in addition to improving numerical stability of the recursive estimation procedure.

Advantage: Characterizing and modelling the power fluctuations for the specific case of offshore wind farms is a current challenge

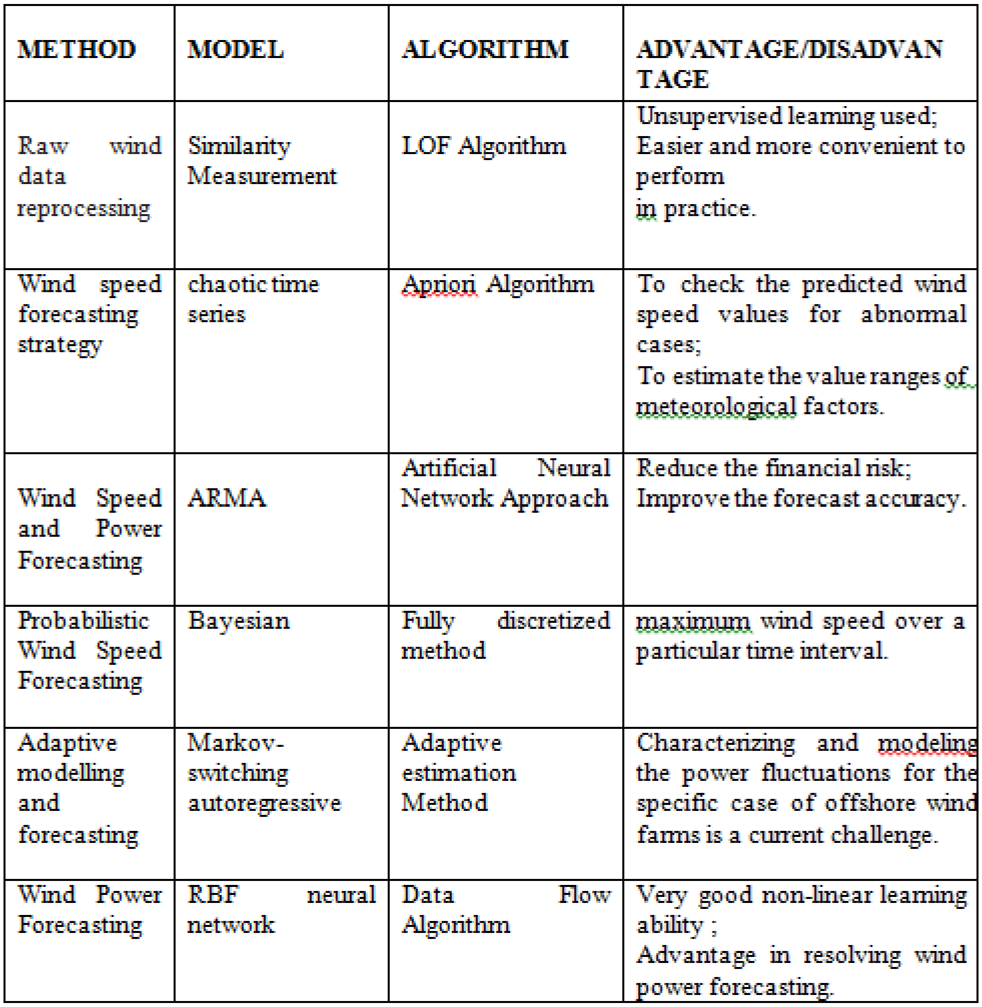
6. WIND POWER FORECASTING

[6]presents ARMA (q, p) model of time series to forecast wind speed and atmospheric pressure, and using the RBF neural network based on this to forecast wind power. Taking the data of measured wind speed and atmospheric pressure from a wind farm as example, to validate the method described above, and the result show that the method has a certain practicality. In this paper, to forecast wind speed by using the method of time series, which the data requirement is low and the cost used to forecast is also low, so it is suitable for the actual operation of businesses. Considering the high degree non-linear relationship showed between the wind speed data and the corresponded generation power, RBF neural network to be used to forecast generation power. And to verify the feasibility and effectiveness of the method in this paper through the experimental data from a wind farm.

The technical requirements (try out) of national grid wind farms accessing grid clearly pointed out the need for forecasting the wind farms power. The analysis of time series and the RBF neural network will be introduced into the wind power forecasting in this article. The forecasting of wind power has a great significance to the construction and operation of wind farm. The study

Conclusions of the above mentioned wind power forecasting system have: Time series model is a dynamic model, which has a very good extension to dynamic data, thereby it could avoid the impact of the directly adding "Window" when we strike the statistical properties of dynamic data. For the random and dynamic of wind speed, the method of time series ARMA reflects a larger advantage.

Advantage: Very good non-linear learning ability; Advantage in resolving wind power forecasting



METHOD	MODEL	ALGORITHM	ADVANTAGE/DISADVANTAGE
Raw wind data reprocessing	Similarity Measurement	LOF Algorithm	Unsupervised learning used; Easier and more convenient to perform in practice.
Wind speed forecasting strategy	chaotic time series	Apriori Algorithm	To check the predicted wind speed values for abnormal cases; To estimate the value ranges of meteorological factors.
Wind Speed and Power Forecasting	ARMA	Artificial Neural Network Approach	Reduce the financial risk; Improve the forecast accuracy.
Probabilistic Wind Speed Forecasting	Bayesian	Fully discretized method	maximum wind speed over a particular time interval.
Adaptive modelling and forecasting	Markov-switching autoregressive	Adaptive estimation Method	Characterizing and modeling the power fluctuations for the specific case of offshore wind farms is a current challenge.
Wind Power Forecasting	RBF neural network	Data Flow Algorithm	Very good non-linear learning ability ; Advantage in resolving wind power forecasting.

RESULT ANALYSIS

Raw wind data pre-processing, it is a type of unsupervised learning algorithm. Therefore, it can detect and classify the raw data using solely the attributes of the data themselves. It is easier and more convenient to perform in practice. Wind speed forecasting strategy is to check the predicted wind speed values for abnormal cases and also to estimate the value ranges of meteorological factors. Wind Speed and Power Forecasting reduce the financial risk and improve the forecast accuracy. Probabilistic Wind Speed Forecasting gives maximum wind speed over a particular time interval. Adaptive modelling and forecasting is used for characterizing and modelling the power fluctuations for the specific case of offshore wind farms is a current challenge. Wind Power Forecasting has very good non-linear learning ability. Here it is understood that raw wind data pre-processing is more better than other papers as it is unsupervised learning .And also it identifies and elimates the outliers.

CONCLUSION

In this paper, raw wind data properties were analyzed. Invalid data can be categorized into five types. A wind data pre-processing methodology has been proposed. Because identifying the unnatural and the irrational data is challenging, this paper treats them 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 effectiveness of the algorithm and the similarity measurement. The performance evaluation of the algorithm has also been discussed. One of the greatest advantages of the proposed methodology is that it is a type of unsupervised learning algorithm. Therefore, it can detect and classify the raw data using solely the attributes 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.

REFERENCES

1. Raw Wind Data Preprocessing: A Data-Mining Approach -Le Zheng, Wei Hu, and Yong Minieee TRANSACTIONS ON SUSTAINABLE ENERGY, VOL. 6, NO. 1, JANUARY 2015
2. A new wind speed forecasting strategy based on the chaotic time series Modelling technique and the Apriori algorithm-Zhenhai Guo a,⇑, Dezhong Chi , Jie Wu, Wenyu Zhang cEnergy Conversion and Management 84 (2014) 140–151
3. Current Status and Future Advances for Wind Speed and Power Forecasting -Jaesung Junga*, Robert P. Broadwatera
4. Probabilistic Wind Speed Forecasting using Ensembles and Bayesian Model Averaging-J. McLean Sloughter, Tilmann Gneiting, and Adrian E. Raftery
5. Adaptive modelling and forecasting of offshore wind power fluctuations with Markov-switching autoregressive models -Pierre Pinson, Henrik Madsen
6. Wind Power Forecasting Based on Time Series and Neural Network -Lingling Li1,2 , Minghui Wang2 , Fenfen Zhu2, and Chengshan Wang*
7. From Data Mining to Knowledge Discovery in Databases -Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth
8. Using Neural Networks to Estimate Wind Turbine Power Generation -Shuhui Li, Member, IEEE, Donald C. Wunsch, Senior Member, IEEE, Edgar A. O’Hair, and Michael G. Giesselmann, Senior Member, IEEE
9. Z. Q. Liu,W. Z. Gao, Y. H.Wan, and E. Muljadi, “Wind power plant prediction by using neural networks,” in Proc. IEEE Energy Convers. Congr.Expo., 2012, pp. 3154–3160.
10. M. Ali, I. Ilie, J. V. Milanovic, and G. Chicco, “Wind farm model aggregation using probabilistic clustering,” IEEE Trans. Power Syst., vol. 28,no. 1, pp. 309–316, Feb. 2013.

11. M. Schlechtingen, I. F. Santos, and S. Achiche, “Using data-mining approaches for wind turbine power curve monitoring: A comparative study,” IEEE Trans. Sustain. Energy, vol. 4, no. 3, pp. 671–679, Jul.2013
12. A. Kusiak, H. Y. Zheng, and Z. Song, “Short-term prediction of wind farm power: A data mining approach,” IEEE Trans. Energy Convers., vol. 24,no. 1, pp. 125–136, Mar. 2009.

AUTHORS

Mahima susan abraham received her Bachelor of Engineering in Computer Science and Engineering from Anna University, Chennai in 2012 .She worked as an Android Developer for 1 year. She is currently doing her Master of Technology in Computer and information Science at Cochin University of Science and Technology. Her area of interest includes Data Mining. E-Mail: mahimasusan1990@gmail.com

Jiby j.puthiyidam received his Bachelor of Engineering in Computer Science and Engineering from Madras University in 1998 and Master of Technology in Computer and information Science from Cochin University of Science and Technology in 2008. He is currently working as Assistant Professor, Department of Computer Science and Engineering, College of Engineering Poonjar, Kerala. He is a life member of Indian Society of Technical Education (ISTE). He has presented many papers in National and International conferences. His area of interest includes Wireless Sensor Networks and Data Mining. E- mail id: jibyjp@gmail.com