Measurement 139 (2019) 370–379

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Learning deep representation of imbalanced SCADA data for fault detection of wind turbines

Longting Chen [a], Guanghua Xu [a,b,*], Qing Zhang [a], Xun Zhang [a]

[a] *School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, PR China*
[b] *State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, PR China*

ARTICLE INFO

Article history:
Received 18 July 2018
Received in revised form 25 February 2019
Accepted 10 March 2019
Available online 12 March 2019

Keywords:
Blades icing accretion fault
Deep learning
Imbalanced SCADA data
Triplet loss
Wind turbine

ABSTRACT

Numerous intelligent fault diagnosis models have been developed on supervisory control and data acquisition (SCADA) systems of wind turbines, so as to process massive SCADA data effectively and accurately. However, there is a problem ignored among these studies. That is, SCADA data distribution is imbalanced and the anomalous data mining is not sufficient. The amount of normal data is much more than that of abnormal data, which makes these models tend to be biased toward majority class, i.e., normal data, while accuracy of diagnosing fault is poor. Aimed at overcoming this problem, a novel intelligent fault diagnosis methodology is proposed based on exquisitely designed deep neural networks. The between-classes imbalance problem is handled by learning deep representation that can preserve within-class information and between-classes information based on triplet loss. The proposed method encourages a pair of data belonging to same class to be projected onto points as close as possible in new embedding space. It tries to enforce a margin between different class data. The effectiveness and generalization of the proposed method are validated on the SCADA data of two wind turbines containing blades icing accretion fault. The result demonstrates the proposed method outperforms the traditional normal behavior modeling method.

1. Introduction:

The rapid extension of wind farms has been drawn much attention with use of wind power that is regarded as a key source in the field of new and clean energy [1]. Naturally, many researchers have taken much consideration on wind turbines that are extremely crucial part of wind farms. Unexpected failures of wind turbines may result in serious accidents and huge economic loss. The fault diagnosis of wind turbines is very important. Blades icing accretion is one of common faults. It easily happens during winter in the north of China.

In general, detection methods of blades icing accretion can be classified into two categories. One is the directed, and the other is the indirect. The directed methods employ specific sensors that can detect physical properties change caused by icing accumulation, such as the microprocessor-based liquid sensor detecting temperature change on blades surface [2], the thermal infrared radiometry sensor probing emissivity change [3], and so on. The main disadvantages of them are that many sensors are not reliable to detect blades icing accretion in the early stage [4] and the extra cost limits their capacity of expansion. The indirect methods are mainly data-driven methods. They are explored on massive wind turbine operational data. They mainly contain three steps: condition monitoring data acquisition, feature extraction and fault classification. Supervisory control and data acquisition (SCADA) system is a mainstream and cost-effective approach to monitoring wind turbines [1,5,6]. It provides almost all-round information of wind turbines, such as environmental parameters, working status, control parameters, and so on. Almost all commercial wind farms have installed this system. Since blades icing accretion will result in power generation loss, many feature extraction approaches have been proposed to monitor dynamic power curves of wind turbines for detection of this kind of fault, such as support vector machine (SVM) [1], adaptive neuro-fuzzy-interference system (ANFIS) [7], neural network (NN) [7,8], and so on. Presently, the advanced feature extraction technology contained in the SCADA-based fault detection methods is the artificial neural networks (ANNs) [5,9–12]. It has proven to be effective in extracting fault information from large SCADA data sets. For example, A. Zaher et al. [5] use multilayer neural networks to build normal behavior model for the anomaly detection of gearbox, based on the normal SCADA data. Meik Schlechtingen et al. [13] apply same idea to predict

[*] Corresponding author at: School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, PR China.
E-mail address: ghxu@mail.xjtu.edu.cn (G. Xu).

https://doi.org/10.1016/j.measurement.2019.03.029
0263-2241/© 2019 Elsevier Ltd. All rights reserved.

expected power output on the basis of normal SCADA data. The prediction error between practical power output and expected model output is considered as an indicator for anomaly. The normal behavior modeling based on artificial neural networks (NBM-ANNs) is used for comparison in this paper.

From these cases studies, it can be understood that there is an embarrassing situation. That is, the phenomenon of data imbalance is ignored and the information of abnormal SCADA data is discarded in the process of feature extraction. This problem makes diagnosis results biased toward the majority class, while the ability of novelty detection is fair weak. After all, wind turbines work well in most of time. The amount of normal data is higher than that of abnormal data. Then, the aberrant status information is easily submerged by the normal. For this data imbalance problem, many methods have been proposed to solve it. They are divided into two categories. The one is data-level method; another is algorithm-level method. The former includes under-sampling [14], over-sampling [15] and synthetic minority oversampling technique (SMOTE) [16]. The goal of the first two methods is to balance data distribution through sampling polies, i.e., under-sampling the majority class or over-sampling the minority class. Essentially, they do not increase any extra data information. However, over-sampling is able to introduce undesirable noise with over-fitting risks and under-sampling may remove valuable information. SMOTE is a special sampling approach. It introduces new and non-repeating instances by interpolating neighboring minority class samples. Many variants of SMOTE have been proposed for improvements [17–19], but their extended decision boundaries are still error-prone by synthesizing noise and borderline samples. Algorithm-level method mainly contains bagging and boosting ensemble-based method [20,21] and cost-sensitive learning method [22,23]. The core of the former is to ensemble multiple 'weak' classifiers that are trained on many balanced subsets. These subsets are constructed by under-sampling the majority class without data information loss. Wu et al. [24] applied this kind of approach to handle multi-class imbalance problem of plant's fault prediction and achieved success. Cost-sensitive learning alternative avoids drawbacks of re-sampling means by assigning higher misclassification cost to the minority class than to the majority. It directly affects the learning process of models from optimization goals. However, different from sampling methods, it is not able to directly apply to all classifier algorithms. The aforementioned options have been well researched for the so called 'shallow' models, but their implications have not yet been systematically studied and verified for deep representation learning.

Aimed at above-mentioned problems, a feature extraction method for learning deep representation conditioned on the imbalanced data is proposed. It makes full use of two types data structure information, i.e., the normal data information and the abnormal data information. It is based on the deep neural networks (DNNs) which has achieved great success with highly non-linear learning capacity in many domains recently, such as image recognition [25], audio classification [26], text processing [27], as well as the fault diagnosis of some mechanical components [28,29]. These cases almost have an assumption that the distribution of target data is balanced. The imbalanced situation is considered rarely. In here, the proposed method works on the imbalanced data. It handles this problem from the inner intrinsic data structure and the DNNs structure. Unlike traditional fault classification methods whose optimization goal is the binary cross-entropy [28,30], the proposed method is intended to preserve within-class information and between-classes information of data simultaneously by learning deep representation on a *d* dimensional hypersphere. It does not directly take data label information into account. The loss function of the proposed DNNs is selected as the triplet loss [31]. It is aimed at preserving locality across same health condition data and discrimination between different health condition data. Meanwhile, the deep neural network for extracting features is exquisitely designed. The bypass structure is introduced into it for considering local features and global features of original SCADA data. There are few researches about learning deep feature representation based on imbalanced SCADA data.

The main contributions of this paper are summarized as follows: (1) the phenomenon of data imbalance is clearly pointed out in the process of wind turbine SCADA data acquisition. This phenomenon is ignored by many research cases. Its influence on fault diagnosis models performance is also analyzed theoretically. (2) Regarding above problem, a novel and advanced fault feature extraction method is proposed on the basis of well-designed DNNs and triplet loss. It can preserve within-class data information and between-classes data information at the same time by learning deep representation in hypersphere, unlike traditional DNN-based fault diagnosis methods directly exploiting class label information to build up discriminative models. (3) The proposed deep neural network is designed exquisitely by introducing bypass component. It can take local features and global features of SCADA data into consideration simultaneously. (4) The effectiveness and generalization ability of the proposed model are validated on SCADA data of two real wind turbines that operate on varying working condition. The proposed model is also compared with the traditionally classical SCADA-based method, named normal behavior modeling based on ANNs. The results indicate that the proposed method is superior to the NBM-ANNs.

The rest of paper is organized as follows: the effect of between-classes imbalance is analyzed and illustrated in Section 2. The intelligent feature extraction method for preserving data local information based on DNNs is introduced in Section 3. Section 4 presents evaluation metrics of fault diagnosis models in the imbalanced data situation. The effectiveness and generalization of the proposed method are validated by two real wind turbine SCADA data in Section 5. In addition, the proposed method is compared with the classical SCADA-based method, i.e. the anomaly detection method by modeling normal behavior based on ANNs. Conclusion is drawn in Section 6.

2. Analyzing the effect of between-classes imbalance

Data imbalance brings many obstacles to fault diagnosis of wind turbines when it comes to massive SCADA data. However, this problem is usually ignored for simplifying analysis. We will take binary classification problem in the two dimensional feature spaces as an example to simulate the effect of between-classes imbalance of SCADA data intuitively.

Considering that neural network is one of the most advanced algorithms applied in feature extraction of SCADA data, this method is also employed in this demonstration. Fig. 1(a) shows the distribution of simulation dataset with different coincidence degree in the negative data samples. The imbalanced situation is made by duplicating original negative samples in ten times. This way does not change the variance of original negative samples space. It simulates actual sample distribution contained in SCADA data to a certain degree. After all, wind turbines operate well in most time of their life cycles, so there is likely to be a lot of repetitive or similar data points in SCADA dataset. The simulated dataset is denoted as $\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$, $x^{(i)} \in R^2$, $y^{(i)} \in R^2$. The labels 0 and 1 denote normal class and fault class, respectively. Obviously, this dataset is non-linearly separable. The ratio between normal samples and abnormal samples is 1:1 in original training set. Due to negative samples being duplicated many times, the ratio between

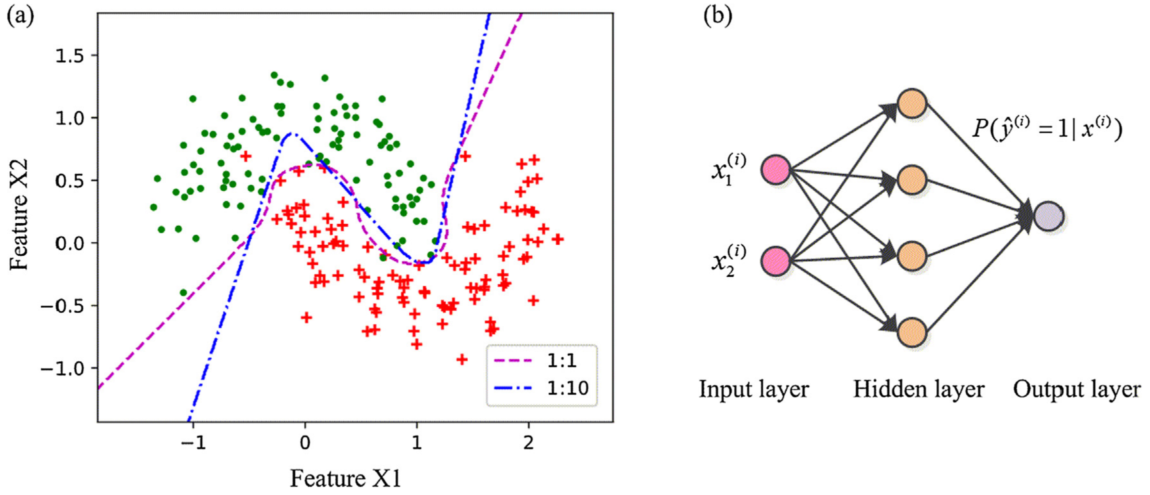

Fig. 1. (a) Simulation dataset where dot denotes positive samples and cross represents negative samples, and (b) structure of neural networks classifier.

positive and negative samples becomes 1:10 in new training set. The classifier, i.e., neural network, is depicted in Fig. 1(b). It has one input layer, one hidden layer with four hidden units, and one output layer. The loss function of this network is binary cross entropy which is used in almost all fault detection models based on classification [28,29]. Its mathematical expression is as follows:

$$J(\theta) = J(\theta)_{pos} + J(\theta)_{neg}$$
$$J(\theta)_{type} = -\frac{1}{m}\left[\sum_{i=1}^{m_{type}} y^{(i)}\log h_\theta(x^{(i)}) + (1-y^{(i)})log(1-h_\theta(x^{(i)}))\right], type \in \{pos, neg\} \tag{1}$$

This loss is derived from two parts, i.e., the normal class data and abnormal class data. The symbol θ represents weight parameters of the neural networks; m_{pos} and m_{neg} denote the number of positive samples and negative samples, respectively; $h_\theta(x^{(i)})$ is the conditional probability $P(y^{(i)} = 1|x^{(i)})$ that $i-th$ sample belongs to fault class.

When the model is trained over by back propagation (BP) algorithm [32], the decision boundaries can be obtained with regard to each dataset. They are plotted in Fig. 1(a). From this figure, it can be seen that data overlap makes great difference among these boundaries. As the repeat ratio increases, the discriminative curve is far away from the negative samples. More and more positive samples are misclassified. The area of the positive is invaded gradually. This phenomenon can be explained by the optimization goal and training process of neural networks. On the one hand, the update of weight parameters θ is driven by the average loss that defined in (1), that is:

$$\theta^{(l)} := \theta^{(l)} - r \cdot \frac{1}{m} \cdot \left[\sum_{i=1}^{m_{pos}} \frac{\partial}{\partial\theta^{(l)}} J(\theta; x^{(i)}, y^{(i)}) + \sum_{j=1}^{m_{neg}} \frac{\partial}{\partial\theta^{(l)}} J(\theta; x^{(j)}, y^{(j)})\right] \tag{2}$$

where the symbol r denotes learning rate and $J(\theta; x, y)$ represents single sample loss that is produced by the sample (x, y). When the number of repeated data points goes up gradually, the ratio between m_{neg} and m becomes smaller and smaller, which makes the average gradient that is derived from positive samples close to zero. Naturally, positive samples make little contribution to weight update. They go unlearned due to their severe underrepresentation. On the other hand, the cost function $J(\theta)$ is highly non-convex when the number of hidden layers goes up. The solution of it is prone to be located in the local extremum, which causes the neural networks to be learned incompletely. In other words, the training loss value is very likely to be non-zero, i.e., $J(\theta) = 0 + \textit{residual}$ in the end of training. The partial residual probably comes from these misclassified positive samples. It makes the anomaly detection worse and worse.

3. Preserving within-class and between-classes information through triplet loss

Different from traditional DNN-based fault diagnosis methods that directly exploit class label information to build up discriminative models, our target is to learn a Euclidean embedding $g(x)$ from a data sample x into a another feature space R^d based on DNNs, such that the embedded feature can preserve within-class and between-classes information concurrently with highly discriminative capacity. The embedding feature is limited on a d dimensional hypersphere. That is, $\| g(x) \|_2 = 1$. Afterwards, a simple k-nearest neighbor classifier is performed on these embedding features for the fault detection of wind turbines.

3.1. Triplet loss

In order to achieve above-mentioned goals, triplet pairs are made from the imbalanced SCADA data. It is motivated by the works of [31,33] in the context of the nearest-neighbor classification. Each element in one triplet is shown in Fig. 2, and is defined as follows:

1) x_a: an anchor.
2) x_p: one data sample having same class with anchor.
3) x_n: one data sample having converse class with anchor.

Then, the following relationship is encouraged to build up on each triplet pair in d dimensional feature space:

$$\| g(x_a^{(i)}) - g(x_p^{(i)}) \|_2^2 + \alpha < \| g(x_a^{(i)}) - g(x_n^{(i)}) \|_2^2 \forall (x_a^{(i)}, x_p^{(i)}, x_n^{(i)}) \in T \tag{3}$$

where the symbol T represents the training set, and α is the predefined margin between positive and negative samples. The distance between two feature embedding represents a measure of similarity. Thereby, the cost function of the fault diagnosis deep neural networks is expressed as below:

$$J(\theta) = \sum_{i=1}^{N} \max(0, \| g(x_a^{(i)}) - g(x_p^{(i)}) \|_2^2 - \| g(x_a^{(i)}) - g(x_n^{(i)}) \|_2^2 + \alpha) \tag{4}$$

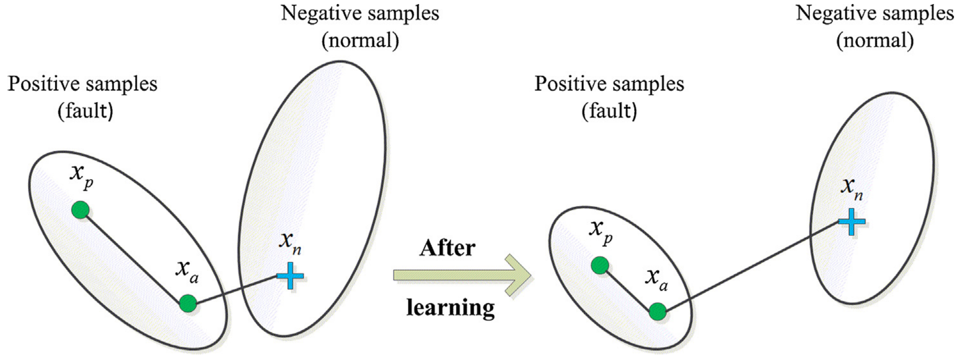

Fig. 2. Schematic diagram of triplet.

From (4), it is clear that this loss function makes samples to be in this kind of status: the distances between same classes of samples are much smaller than those of samples belonging to different classes. This loss enforces those same classes of samples to live on a manifold. It minimizes distance between anchor and positive, while the distance between anchor and negative is enlarged.

3.2. DNN-based intelligent fault diagnosis method

Based on DNNs, this study proposes a novel intelligent fault diagnosis method that can adaptively mine fault information from the massive SCADA data on the varying operation status. The above-mentioned embedding of each sample is extracted on the high-level layer of this network. Due to the SCADA containing multi-attributes information of wind turbines, such as wind parameters, energy conversion parameters, temperature parameters and so on, thereby this fault diagnosis problem can be formulated as anomaly detection issue aimed at structural data. That is, for each data sample $x^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}]$ having n variables or attributes, the conditional probability $P(y^{(i)} = 1|g(x^{(i)}))$ should be respectively calculated. Then, the network is designed for excavating the relationship among them.

The overall framework of the proposed DNNs is shown in Fig. 3. The main backbone is convolutional neural networks (CNNs) that are designed by obeying general design rule of CNNs [25,29,34], followed by an average pooling layer, a L_2 normalization layer and an output layer. Transformations for each convolution layer, pooling layer, and normalization layer are described respectively as follows:

$$\begin{aligned} X^{(l)} &= f(X^{(l-1)} * W^{(l-1)} + b^{(l-1)}) \\ P_k^{(l)} &= \{max(X_k^{(l-1)})|k \in M_{X^{(l-1)}}\} \\ L^{(l)} &= X^{(l-1)} / \| X^{(l-1)} \|_2 \end{aligned} \tag{5}$$

where the symbols $W^{(l-1)}$ and $b^{(l-1)}$ denote convolution kernel and bias respectively in the layer $l-1$; f is the activation function. The rectified linear units (ReLU) is employed, i.e., $Relu(z) = max(0, z)$; $M_{X^{(l-1)}}$ represents the selection of pooling regions in the input feature map $X^{(l-1)}$. The structural parameters of the proposed DNNs are provided in Table. 1. It helps to reproduce the method and results in this work. From this table, it can be seen that the convolution module with small kernel size 3 × 1 is the basic and core component of proposed DNNs. The general fully connected (FC) module that deals with structured data is not applied here. Actually, the FC module can be seen as the special case of convolution module. Fig. 4(a) shows this relationship intuitively. The convolution module evolves into FC module when the kernel size of it becomes big and is equal to input size. Further, the reason why small kernel size is adopted is that two cascaded small convolution modules are approximate to one large convolution module and the former can extract more local features [35]. Factorizing convolution can make neural networks deeper while keeping almost the same computation burden and parameter amount. Fig. 4(a) illustrates this transformation.

It is worth mentioning that there is a bypass component designed exquisitely and introduced to the whole network. This component contains the information flow of global features of each SCADA data sample. Fig. 4(b) shows the structure of it.

The computation of information flow in such block is as follows: for each bypass, the output is:

$$O_{bypass} = f(X^{(l)} * K + b) \odot V \tag{6}$$

where $X^{(l)}$, K, and b are the input features of bypass, convolution filters, and bias respectively. Note that the size of filter is equivalent to that of each feature map in $X^{(l)}$. V is the weight vector, i.e., $V = (v_1, v_2, \ldots, v_s)$, in which s is set to the number of feature maps

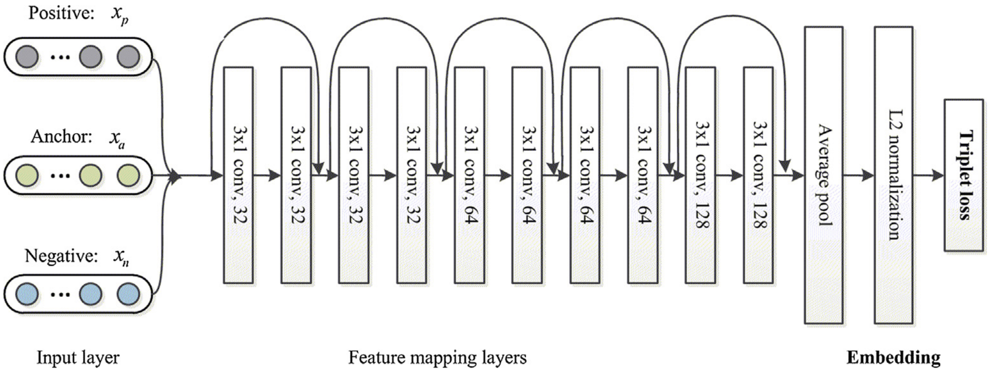

Fig. 3. Structure of the proposed DNNs fault diagnosis model.

Table 1
Architecture for the proposed DNNs fault diagnosis model.

Layer name	Output size	[Kernel size, channel]	Padding
Input layer	22×1	/	/
Block 1	22×1	$\begin{bmatrix} 3\times1, 32 \\ 3\times1, 32 \end{bmatrix} \times 2$	SAME
Block 2	14×1	$\begin{bmatrix} 3\times1, 64 \\ 3\times1, 64 \end{bmatrix} \times 2$	VALID
Block 3	10×1	$\begin{bmatrix} 3\times1, 128 \\ 3\times1, 128 \end{bmatrix} \times 1$	VALID
Average pool	640×1	Stride: 2	/
L2 normalization	640×1	/	/
triplet loss	1	/	/

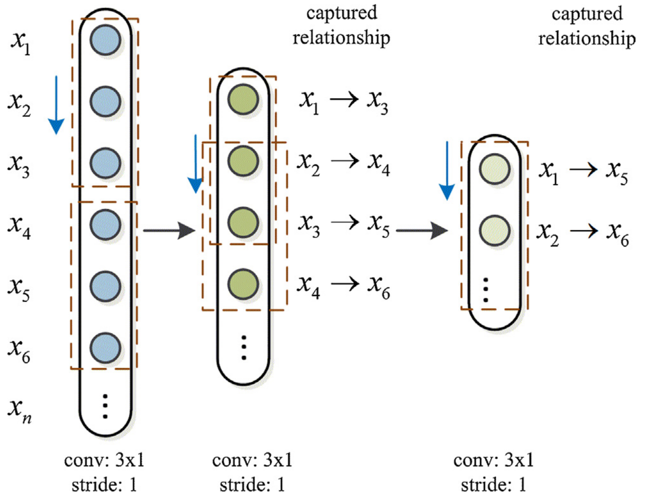

Fig. 5. Drawbacks of common CNNs in dealing with structural data.

in $X^{(l+2)}$. The symbol $\odot$ denotes element-wise multiplication. Then, combined with (5), the overall output of one block can be calculated as follows:

$$O_{block} = f(f(X^{(l)} * W^{(l)} + b^{(l)}) * W^{(l+1)} + O_{bypass}) \quad (7)$$

The reason why this design is added is as follows: on the one hand, the common CNNs is not good at handling structured data, like in SCADA data. That is, it is hard to catch relationship between two variables that are far apart in the lower layers, when used with small convolution kernels. Fig. 5 shows this kind of drawback of CNNs. The connection between x_1 and x_6 could not be captured in the layer one and layer two if the convolution kernels with size 3×1 is used. However, when the size of convolution kernels is expanded gradually, the CNNs start to degenerate to be dense-connected networks, which would result in surge in network parameters and over-fitting in data. By adding such block that contains bypass structure, the above drawbacks can be overcome. On the other hand, the proposed DNNs still takes local features and global features into account simultaneously. In addition, it can be seen that the global feature of data samples serves as bias from (7). In other words, the global feature calculated from bypass works like a gate. It controls the opening and closing of main road of information flow. Useless local features could be filtered out by it. Effectiveness of the proposed DNNs is validated in Section 5. For simplicity, this model is named as TL-Model.

4. Evaluation metric

Traditionally, with regard to one classifier, the most frequently-used evaluation metric is the accuracy. This indicator is also used for assessing many fault diagnosis models [1,28,29]. The fault class, i.e., the minority class is seen as the positive, while the normal class is considered as the negative in here. Then the performance of one classifier can be assessed by confusion matrix. Fig. 6 shows this matrix. The accuracy is calculated as below:

$$accuracy = \frac{(TP + TN)}{m} \quad (8)$$

However, this indicator is deceiving and sensitive to the changes in the imbalanced data situation. Supposing that there are 100 examples where the number of normal samples is 95, one classifier predicts all the samples to be the normal. Then, the accuracy of this classifier is 95%. This number is a nice value at a first glance. Nevertheless this classifier is really awful in practice, since it cannot detect any abnormal samples. Therefore, other comprehensive assessments metrics are introduced. They are precision (P), recall (R) and F-measure (F_β), which are calculated as:

$$\begin{aligned} P &= \frac{TP}{TP + FP} \\ R &= \frac{TP}{TP + FN} \\ F_\beta &= \frac{(1+\beta^2) \cdot P \cdot R}{\beta^2 \cdot P + R} \end{aligned} \quad (9)$$

The precision is considered as a measure of exactness and the recall is seen as a measure of completeness, intuitively. The metric F_β associates both of them for assessing the effectiveness of one

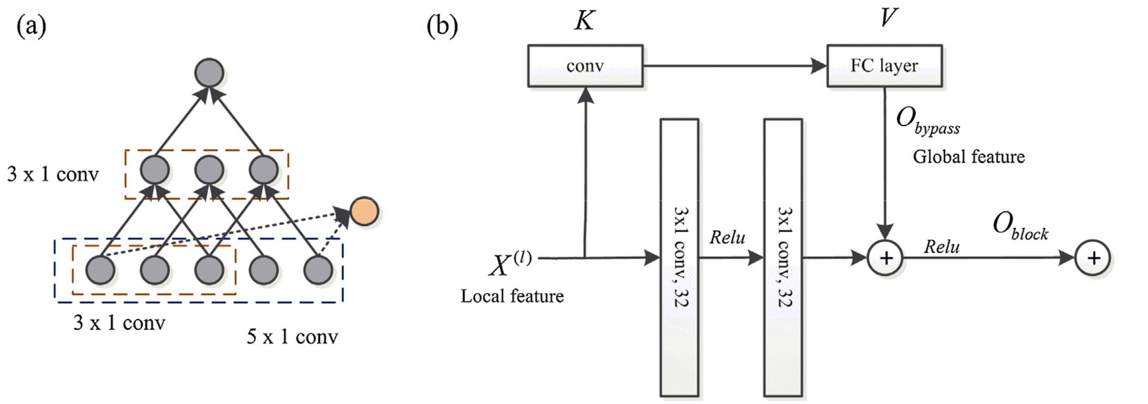

Fig. 4. (a) Approximation between FC module and convolution module and convolution Factorization, (b) structure of the bypass component.

True class

Hypothesis output

	Positive	Negative
Positive	True positive (TP)	False positive (FP)
Negative	False negative (FN)	True negative (TN)

Fig. 6. Confusion matrix.

classifier in terms of ratio β, which is used to weigh the importance between precision and recall. In here, β is set to 1, which indicates that both precision and recall are equally important.

5. Experimental verification

The data available for this research is collected from SCADA system of one wind farm in north China. As the designed rated power of wind turbines continues to increase, the tower heights of wind turbines are also growing. Therefore, blades of wind turbines are easy to freeze on a large area in winter. Then, the load of blades increases, which will cause a huge threat of blades broken to wind turbines. The proposed method is employed to detect the fault of blades icing accretion from imbalance SCADA data, and helps to start ice-breaking system in advance. There are three wind turbines considered, i.e., wind turbine 1, wind turbine 2, and wind turbine 3. All of them have individual pitch control systems (IPC), which can help them track wind power better. The IPC is composed of controller, servo driver, pitch motor, reducer, battery cabinet, battery and so on. The wind turbine 1 is used as training set, and the wind turbine 2 and 3 are used as testing set for validating the effectiveness and generalization ability of the proposed method.

5.1. SCADA data description

In this research, the data is collected at 7 s interval for each wind turbine, ranging from November 1, 2015 to January 1, 2016, in which the weather is fair cold. Fig. 7 shows partially environmental temperature curve of wind turbine 3. It can be seen that this temperature variation reaches atmospheric icing condition described in Ref. [4]. Red dot denotes normal samples while blue dot denotes abnormal samples in Fig. 7. Considering that blades icing accretion can lead to overloading and power generation loss, twenty two continuous variables related to fault information are

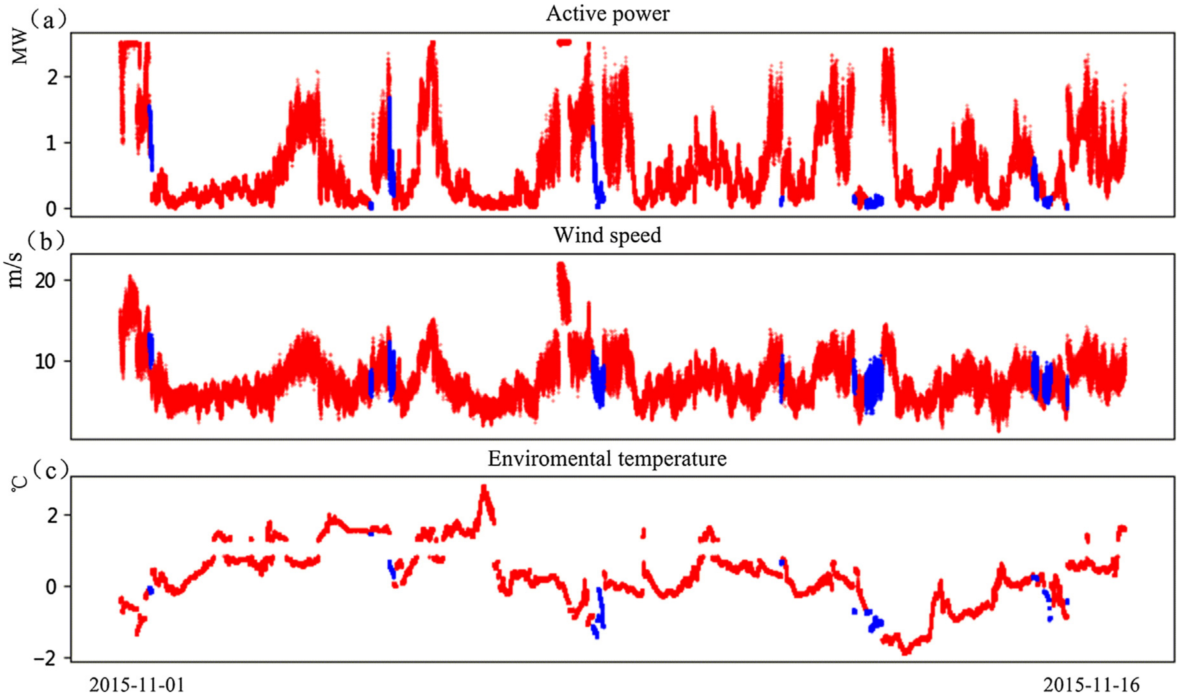

Fig. 7. Variation curve of three parameters: (a) active power, (b) wind speed, and (c) environmental temperature of wind turbine 3. Red dot denotes normal samples. Blue dot denotes abnormal samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Parameters related to fault information in SCADA data.

No.	Name	No.	Name
1	Wind speed w_s	12	Drive current of 1# pitch motor c_1
2	Wind direction in real time w_d	13	Drive current of 2# pitch motor c_2
3	Average wind angle in 25 s w_{d-25}	14	Drive current of 3# pitch motor c_3
4	Active power p_a	15	Environmental temperature t_{out}
5	Rotating speed of generator r	16	Nacelle temperature t_{in}
6	1# pitch angle a_1	17	Temperature of 1# pitch motor t_{m-1}
7	2# pitch angle a_2	18	Temperature of 2# pitch motor t_{m-2}
8	3# pitch angle a_3	19	Temperature of 3# pitch motor t_{m-3}
9	1# pitch speed s_1	20	Temperature of 1# battery cabinet t_{b-1}
10	2# pitch speed s_2	21	Temperature of 2# battery cabinet t_{b-2}
11	3# pitch speed s_3	22	Temperature of 3# battery cabinet t_{b-3}

Table 3
Proportion of the normal data and the abnormal data.

Name	Proportion of the normal (%)	Proportion of the abnormal (%)
Wind turbine 1	93.61	6.39
Wind turbine 2	94.08	5.92
Wind turbine 3	93.21	6.79

selected from SCADA. They can be grouped into three categories, i.e., (1) wind parameters: such as wind speed and wind direction. They are measured directly through anemometer and wind vane; (2) energy-related parameters: these parameters are closely related with power output of wind turbines, e.g., rotating speed of generator, pitch angle, pitch speed, drive current of pitch motor and so on; (3) temperature parameters: they reflect working condition and working status of wind turbines to a certain degree. They include environmental temperature, nacelle temperature, temperature of pitch motor, temperature of battery cabinet, and so on. Fig. 7(a–c) shows curves of some critical parameters, i.e., active power and wind speed. The detailed information is listed in Table 2. From this table, it can be seen that these parameters contain rich information of wind turbine operation condition and external environment. The collected SCADA data covers multiple operation stages of wind turbines, such as the stage of tracking wind energy, the stage of constant rotating speed, and the stage of constant power. Note that we discard the shutdown stage where the wind speed is bigger than cut-out speed. Due to wind turbines working well at the most of time, the phenomenon of imbalance is extremely severe, which can be seen from Table 3 that lists the proportions of normal data and abnormal data with regard to each wind turbine. The amount of normal data is much more than ten times that of the abnormal data.

5.2. Extraction and analysis of embedding features

Aimed at above-mentioned problems, the proposed method is applied to detect blades icing accretion fault of wind turbines from the imbalanced SCADA data. The feature extraction network is built up based on the principle that is stated in Section 3.2, and is programed by deep learning toolbox named Tensorflow. This network is optimized by the back propagation algorithm (BP) [32] and is trained on wind turbine 1 dataset. When the network is trained over, the embedding of each data sample can be calculated on the L_2 normalization layer. The dimension of it is up to 640. Because the learning process of proposed DNN is driven by the used of distance metric and original data points are projected onto high dimensional hyperplane, a simple k-nearest neighbor (kNN) classifier is selected to show the efficacy of the learned embedding features. It is trained on the extracted features of Turbine 1 with the parameter k being 20. This parameter is optimized by grid searching. Better performance is expected by the use of more elaborated classifiers. The result of blades icing accretion fault detection is shown in Table 4. From this table, it can be seen that both the precision and the F-measure are fair high on the tested wind turbine 2 and 3, while the recall (R) is relatively lower. The proposed method is able to differentiate the normal and abnormal status with a high accuracy.

In order to validate the effectiveness of extracted features further, the dimension reduction method named principle component analysis (PCA) is employed to visualize the discrimination and clustering of them. Fig. 8(a)–(d) shows the first three principle components of original features and embedding features as for wind turbine 2 and 3. It is clear that the points that belong to same class are clustered more closely and the points that represent different health conditions are separated far away, compared with the original features. Different health conditions of sample points are jagged with each other in the original feature space of reduced dimension. It is worth mentioning that the datasets of tested wind turbine 2 and 3 contain variable working condition information, such as the stage of tracking wind power, the stage of constant speed and the stage of constant power. It is ordinarily difficult for extracting and selecting features from such datasets. Nevertheless, Fig. 8(b) and (d) indicate that the proposed DNNs is still effective in adaptively mining the intrinsic features for characterizing health conditions of wind turbines.

In order to further validate the drawback of cross-entropy loss and the effectiveness of bypass components, two DNNs models are built. The one is constructed by replacing L_2 normalization layer of TL-Model with softmax classifier layer and uses cross-entropy loss as optimization goal; another is built by removing bypass component of TL-Model and still optimized by triplet loss. These two models are trained with BP algorithm. For convenience, they are named as CE-Model and TL-Model-Out, respectively. The blades icing accretion fault detection results of them are listed in Table 4. From this table, it is clear that the recall of CE-Model is much lower than that of TL-Model. There are many missed inspection points. Fault patterns go unlearned because they make little contribution to cross-entropy loss. Cross-entropy loss is sensitive to imbalanced data situation, which makes CE-Model's ability to detect faults very poor. With regard to TL-Model-Out, this model has comparable performance compared with TL-Model. By adding bypass component, TL-Model has a slight increase in precision. It can mine fault points more accurately.

For comparison, one of the advanced methods named normal behavior modeling (NBM) [13] is employed to deal with the same datasets. This method is usually applied for fault diagnosis of wind turbines. It is aimed at modeling the active power curve of normal condition by the use of ANNs, while the anomalous status data of wind turbines is ignored in the modeling process. This curve represents the relationship between power produced and the wind speed that is in the range from cut-in speed to cut-out speed. The shape of it indicates the health condition of a wind turbine [1]. Fig. 9(a), (b) shows the scatter plot of the relationship between power and wind speed for the test sets. Being same as the ANNs structure in [13], the ANNs used in here is constructed with two hidden layers. The number of hidden neurons is 9. The input features are also wind speed w_s, wind direction w_d, and environmental temperature t_{out}. The last two features make some contribution to the prediction of power curve, which has been validated by the [13]. The output is active power P_a of wind turbine. Then, the NBM-ANNs model is trained on the data of wind turbine 1 by BP

Table 4
Result of blades icing accretion fault detection.

Method	Wind turbine 2			Wind turbine 3		
	P	R	F_1	P	R	F_1
TL-Model	0.9085	0.6310	0.7447	0.8153	0.5960	0.6886
TL-Model-Out	0.8973	0.6389	0.7464	0.7984	0.5893	0.6781
CE-Model	0.7210	0.5415	0.6185	0.6225	0.4992	0.5541
NBM-ANNs	0.4239	0.5194	0.4668	0.4366	0.4893	0.4614

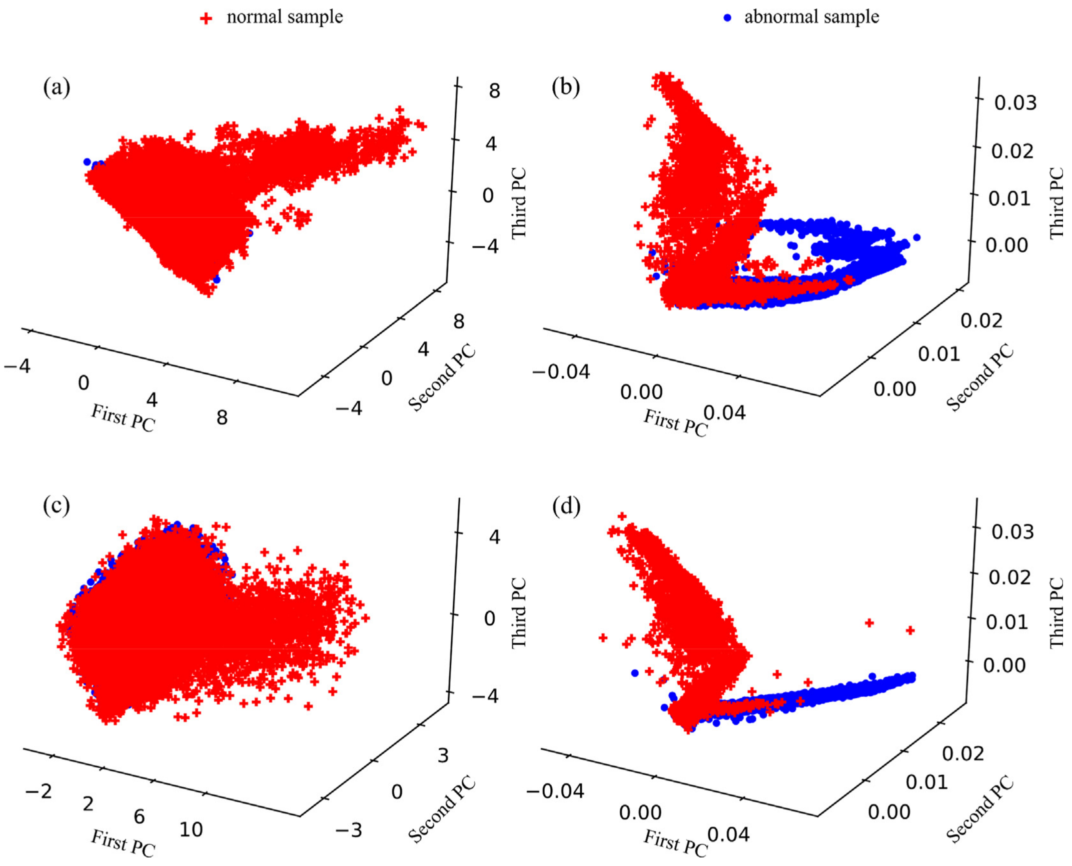

Fig. 8. Visualization of the first three principle components: (a) original features of wind turbine 2, (b) deep representation of wind turbine 2, (c) original features of wind turbine 3, and (d) deep representation of wind turbine 3.

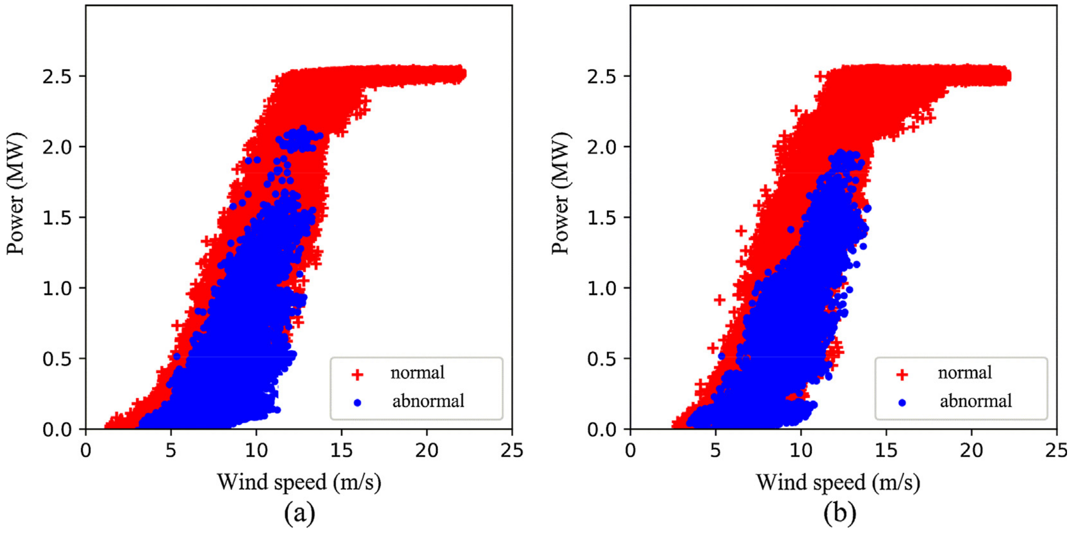

Fig. 9. Power curves of (a) wind turbine 2 and (b) wind turbine 3.

algorithm [32]. The prediction error between the model's output (expectation) and the real measurement can be used as an indicator for anomaly. Fig. 10 shows the predicted powers of the normal and abnormal data samples in tested datasets. From this figure, it is clear that there is a lot of overlap between the normal and the abnormal in the plane that defined by the actual power and its expectation, and the actual power of the most of abnormal samples is lower to their predicted power. Using the grid searching, we find that F_1 get the best value when the prediction error exceeds 0.38 MW. The fault diagnosis result of the NBM-ANNs is listed in Table 4. Both the precision and recall are very lower compared with the result that is obtained from the proposed method, which indicates that there exists a lot of missed inspection and false detection. When taking precision and recall into account simultaneously, the performance of the blade icing fault detection is improved more than 20% by the proposed method. Therefore, the

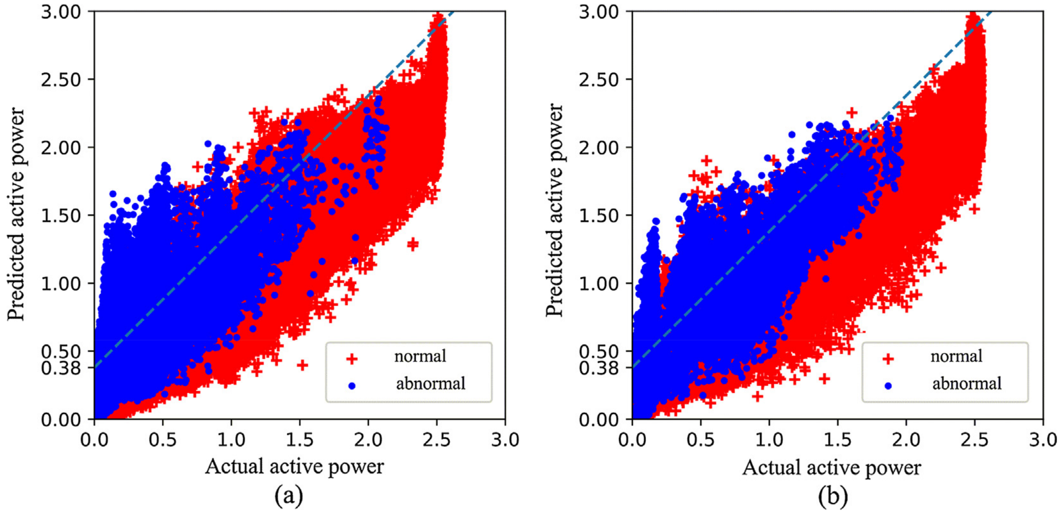

Fig. 10. Relationship between predicted and actual power in (a) wind turbine 2 and (b) wind turbine 3.

proposed method has a good behavior in adaptively mining the fault characteristics from the imbalanced massive SCADA data.

6. Conclusion

This paper proposes a DNN-based intelligent method for detection of blading icing accretion fault with approach of imbalanced SCADA data, in which the data imbalance problem is ignored by many studies of wind turbines in the process of data acquisition. Unlike traditional fault diagnosis methods that directly use data label information or drop the fault data information, the proposed method tries to learn a deep feature representation of each data sample in a hypersphere. This kind of feature representation can make the samples belonging to same type closer, while the data samples belonging to different types are separated far away. The network structure of the DNNs is also designed exquisitely by taking local features and global features simultaneously. The effectiveness of this method is verified by the use of three real SCADA dataset where the data samples contain different health conditions under varying operation status. The result shows that the proposed method highly outperforms the traditional method named NBM-ANNs in the detection of blade icing accretion. In the proposed method, only the between-class imbalanced is considered. It is interesting to consider both between-classes imbalance and inter-class imbalance simultaneously in the future study. After all, wind turbines work well in most of time, and the normal operation status contains many stages, such as the stage of maximal power point tracking, the stage of constant rotating speed, the stage of rated power output, and so on. It may improve the proposed model performance by taking different operation stages of wind turbines into consideration.

Acknowledgements

This research is supported by the Provincial Key Research Project of Shaanxi (Grant no. 2017ZDCXL-GY-01-03-03) and Ministry of Industry and Information Technology of China (Research and verification of remote operation and maintenance standards for TBM).

References

[1] A. Kusiak, W. Li, The prediction and diagnosis of wind turbine faults, Renew. Energy 36 (2011) 16–23.
[2] J. Maatuk, Microprocessor-based liquid sensor and ice detector, in, Google Patents, 1999.
[3] C.Q.G. Muñoz, F.P.G. Márquez, J.M.S. Tomás, Ice detection using thermal infrared radiometry on wind turbine blades, Measurement 93 (2016) 157–163.
[4] M.C. Homola, P.J. Nicklasson, P.A. Sundsbø, Ice sensors for wind turbines, Cold Reg. Sci. Technol. 46 (2006) 125–131.
[5] A. Zaher, S. McArthur, D. Infield, Y. Patel, Online wind turbine fault detection through automated SCADA data analysis, Wind Energy 12 (2009) 574–593.
[6] J. Tautz-Weinert, S.J. Watson, Using SCADA data for wind turbine condition monitoring–a review, IET Renew. Power Gener. 11 (2016) 382–394.
[7] M. Schlechtingen, I.F. Santos, S. Achiche, Wind turbine condition monitoring based on SCADA data using normal behavior models. Part 1: system description, Appl. Soft Comput. 13 (2013) 259–270.
[8] S. Li, D.C. Wunsch, E.A. O'Hair, M.G. Giesselmann, Using neural networks to estimate wind turbine power generation, IEEE Trans. Energy Convers. 16 (2001) 276–282.
[9] P. Bangalore, L.B. Tjernberg, An artificial neural network approach for early fault detection of gearbox bearings, IEEE Trans. Smart Grid 6 (2015) 980–987.
[10] M.C. Garcia, M.A. Sanz-Bobi, J. del Pico, SIMAP: intelligent system for predictive maintenance: application to the health condition monitoring of a windturbine gearbox, Comput. Ind. 57 (2006) 552–568.
[11] Y. Zhang, C. Zhang, Y. Zhao, S. Gao, Wind speed prediction with RBF neural network based on PCA and ICA, J. Electr. Eng. 69 (2018) 148–155.
[12] Y. Zhang, C. Zhang, J. Sun, J. Guo, Improved wind speed prediction using empirical mode decomposition, Adv. Electr. Comput. Eng. 18 (2018) 3–10.
[13] M. Schlechtingen, I.F. Santos, S. Achiche, Using data-mining approaches for wind turbine power curve monitoring: a comparative study, IEEE Trans. Sustain. Energy 4 (2013) 671–679.
[14] W.C. Lin, C.F. Tsai, Y.H. Hu, J.S. Jhang, Clustering-based undersampling in class-imbalanced data, Inform Sci. (2017).
[15] J. Xie, Z. Qiu, The effect of imbalanced data sets on LDA: a theoretical and empirical analysis, Pattern Recogn. 40 (2007) 557–562.
[16] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[17] H. Han, W.Y. Wang, B.H. Mao, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, International Conference on Advances in Intelligent, 2005.
[18] T. Maciejewski, J. Stefanowski, Local neighbourhood extension of SMOTE for mining imbalanced data, Comput. Intell. Data Min. (2011).
[19] Y. Zhang, X. Li, L. Gao, L. Wang, L. Wen, Imbalanced data fault diagnosis of rotating machinery using synthetic oversampling and feature learning, J. Manuf. Syst. (2018).
[20] C. Seiffert, T.M. Khoshgoftaar, J.V. Hulse, Improving software-quality predictions with data sampling and boosting, IEEE Trans. Syst. Man Cybern. – Part A 39 (2009) 1283–1294.
[21] N.V. Chawla, A. Lazarevic, L.O. Hall, K.W. Bowyer, SMOTEBoost: improving prediction of the minority class in boosting, Lect. Notes Comput. Sci. 2838 (2003) 107–119.
[22] C. Elkan, The foundations of cost-sensitive learning, in: Proceedings of the 17th International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers Inc., Seattle, WA, USA, 2001, pp. 973–978.
[23] M.T. Kai, An instance-weighting method to induce cost-sensitive trees, IEEE Trans. Knowl. Data Eng. 14 (2002) 659–665.
[24] Z. Wu, W. Lin, J. Yang, An Integrated Ensemble Learning Model for Imbalanced Fault Diagnostics and Prognostics, IEEE Access, 2018, pp. 8394–8402.
[25] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[26] H. Lee, P. Pham, Y. Largman, A.Y. Ng, Unsupervised feature learning for audio classification using convolutional deep belief networks, Adv. Neural Inf. Process Syst. (2009) 1096–1104.

[27] B. Shi, X. Bai, C. Yao, An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2016).

[28] F. Jia, Y. Lei, J. Lin, X. Zhou, N. Lu, Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data, Mech. Syst. Sig. Process. 72 (2016) 303–315.

[29] W. Zhang, C. Li, G. Peng, Y. Chen, Z. Zhang, A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load, Mech. Syst. Sig. Process. 100 (2018) 439–453.

[30] M. Gan, C. Wang, Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling element bearings, Mech. Syst. Sig. Process. 72 (2016) 92–104.

[31] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: a unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.

[32] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533–536.

[33] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (2009) 207–244.

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.