J Inf Process Syst, Vol.16, No.4, pp.870~881, August 2020
https://doi.org/10.3745/JIPS.04.0183

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Research on Fault Diagnosis of Wind Power Generator Blade Based on SC-SMOTE and kNN

Cheng Peng*, Qing Chen*, Longxin Zhang*, Lanjun Wan*, and Xinpan Yuan*

Abstract

Because SCADA monitoring data of wind turbines are large and fast changing, the unbalanced proportion of data in various working conditions makes it difficult to process fault feature data. The existing methods mainly introduce new and non-repeating instances by interpolating adjacent minority samples. In order to overcome the shortcomings of these methods which does not consider boundary conditions in balancing data, an improved over-sampling balancing algorithm SC-SMOTE (safe circle synthetic minority oversampling technology) is proposed to optimize data sets. Then, for the balanced data sets, a fault diagnosis method based on improved k-nearest neighbors (kNN) classification for wind turbine blade icing is adopted. Compared with the SMOTE algorithm, the experimental results show that the method is effective in the diagnosis of fan blade icing fault and improves the accuracy of diagnosis.

Keywords

Fault Diagnosis, kNN Algorithm, SCADA dataset, SC-SMOTE Algorithm

1. Introduction

With the economic progress brought by the two industrial revolutions [1], a large number of traditional energy resources have been exhausted in recent years. The destruction of ecological environment, such as global warming, soil erosion, land desertification and other issues, has been alerting people to adjust the use of energy and explore new energy resources, each country attaches great importance to the research and application of new energy technologies, and China has placed high expectations on the wind power industry. As a clean, safe and renewable green energy, wind energy [2] has become one of the new energy sources with great development potential, which promotes the improvement of modern energy structure and provides important support for the sustainable development of human society.

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 [3] data acquisition. This phenomenon is ignored by many research cases. Although the existing SMOTE method [4] considers the imbalance phenomenon, it does not deal with the attribution problem of the boundary samples.

(2) Regarding above problem, a novel and improved balanced algorithm SC-SMOTE (safe circle

Manuscript received May 31, 2019; first revision November 4, 2019; accepted December 21, 2019.

Corresponding Author: Qing Chen (Qinchen@hut.edu.cn)

* School of Computer Science, Hunan University of Technology, Zhuzhou, China (chengpeng@csu.edu.cn, Qinchen@hut.edu.cn, 292702368@qq.com, 2330955108@qq.com, 1762048027@qq.com)

synthetic minority oversampling technology) considering boundary sample assignment problem is proposed. It focus on regional and edge confusion caused by new samples, avoiding new samples and further marginalizing data sets and sample mixing, unlike traditional SMOTE does not consider positive class sample marinating and the distribution interval of newly generated sample sets is too single.

(3) Balanced data is classified using the k-nearest neighbors (kNN) classification [5] algorithm, which can improve the timeliness of wind turbine blade icing fault diagnosis and reduce the computational complexity.

(4) The effectiveness and generalization ability of the proposed model are validated on SCADA data of the wind turbine that operate on varying working condition. The proposed model is also compared with the traditionally classical SCADA-based method, named SMOTE. The results indicate that the proposed method is superior to the SMOTE.

This paper is organized as follows: Section 2 discusses the existing research status. In Section 3, we propose the improved algorithm of SC-SMOTE for balancing data sets and kNN [6] for fault diagnosing. In Section 4, we analyze and compare the existing and proposed algorithm in terms of accuracy. Section 5 concludes this paper.

2. Research Status and Analysis

At present, most wind power plants in China use manual experience to judge whether the ice blower needs shutdown maintenance or not, and the scientific research on it is still a relatively frontier field. There are few references on ice fault diagnosis and prediction of wind turbine blades at home and abroad, which is still in the stage of research and development, and the existing research, has only appeared in recent years.

Wang et al. [7] pointed out that a large number of wind turbines are installed in high-altitude windy areas and there are often blade icing failure. Sun et al. [8] studied the various effects on the control system when fan blades were covered with ice. Dong et al. [9] summarized the phenomenon of fan blade icing and some current icing fault monitoring methods, pointing out that fan blade icing has always been a common problem for wind turbine workers worldwide. Puyals et al. [10] described some technical routes based on the deviation between the SCADA monitoring data and the expected data to warn the fan operating condition.

For multiclass monitoring data, Li et al. [11] proposed a method to judge the icing failure based on the experience and setting the specified threshold. The conclusion given in the paper is: air humidity is greater than 75%, the fan power curve drops by more than 20%, the horizontal acceleration of the fan vibration is greater than 50% of normal operation", the early warning accuracy reach 100%. However, the conclusion given in the paper is not absolutely universal. The accuracy mentioned in the paper is actually the precision in machine learning two-category evaluation index, which is the correct proportion only in the forecasted frozen samples, as long as the rules are set more harsh, the threshold to trigger alarm is rarer, and trying to rule out the state of suspected icing and only predicting the situation with great certainty.

Wang et al. [12] summarized that there have been studies to deal with this problem by additionally arranging ultrasound, infrared and other sensors in recent years. Although these measures can play a role, they will require additional customization, affecting balance and operating costs of original wind turbines,

and the popularization rate in practical application is also very low. Yang et al. [13] reviewed the research status and important research results of wind turbines state detection in recent years from the aspects of wind turbine health state detection and performance state detection, discussed the current problems faced by wind turbines state detection, and solved the facing problems from the aspects of state detection equipment, software integration, state detection intelligence and standardization, etc. The article points out that fault mechanism analysis, multi-state detection fusion technology and unified platform comprehensive health detection and evaluation system are the new trends in the development of wind turbine state detection. Guo et al. [14] developed a technology of monitoring fan icing and deicing based on FBG (filter Bragg grating) sensing and grapheme film.

For the imbalance problem of datasets, many methods have been adopted to solve it. They are divided into two categories. The first one is data-level method; another is model-based method. The former includes under sampling, oversampling, and synthesis of a few oversampling techniques. The goal of the first two methods is to balance the distribution of data by sampling strategies, that is, to under sample most of the categories or to sample too few. In essence, they do not add any additional data information. However, oversampling introduces unwanted noise and there is a risk of over fitting, and oversampling may delete valuable information. SMOTE is a special sampling method. It introduces new and non-repeating instances by interpolating adjacent minority samples. However, by combining noise and boundary samples, their extended decision boundaries are still prone to errors.

Aimed at above-mentioned problems, a data balanced method for solving the problem of boundary sample attribution by learning deep representation conditioned on the imbalanced data is proposed. In here, the proposed method works on the imbalanced data. It handles this problem from the restricted sample generating area, the probability of the region where the boundary sample is located is judged by triangular region through density analysis of a few samples. Unlike traditional balance methods whose optimization goal is the linear partition, the proposed method is intended to preserve positive sample and negative sample information of data simultaneously by learning deep representation on two circular areas.

3. Proposed Fault Diagnosis Scheme of Fan Blade

At present, most of the data sets collected by power plants are segmented discrete data sets due to various reasons such as shutdown, take SCADA as an example. Based on the real data sets, an improved diagnosis method of fan blade icing is proposed in this paper. The specific diagnosis process is shown in Fig. 1.

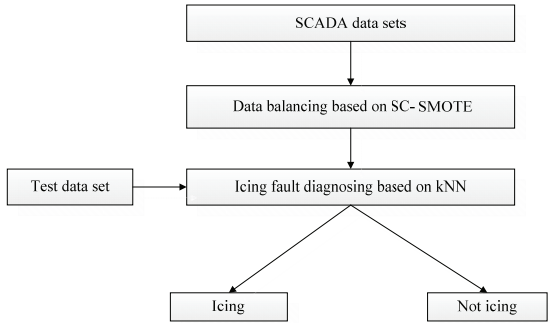

Fig. 1. Flow chart of icing fault diagnosis for fan blades.

3.1 Balancing Data Sets based on SC-SMOTE

There are great differences in the proportion of positive and negative samples in the data collected from the monitoring fan operation process. This data imbalance will have a great impact on the classification algorithm. There are three common processing methods: under-sampling, over-sampling, adding penalty items in the loss function of the model and model integration. In order to preserve the original true information as much as possible, after eliminating the monitoring variables with very low correlation degree with fan blade icing, this paper adopts over-sampling method for a few key feature data.

At present, most algorithm optimization research focus on the weight of the positive sample and the misjudgment penalty, and the research on the data set distribution are insufficient. The boundary of different classes is the most error-prone areas of classification and discrimination algorithms. This article starts with limiting the sample generation area, defining that if the other class samples around a minority sample are denser, the more difficult it is to distinguish the classification algorithm, the more "dangerous" it is, otherwise the more "safe", as shown in Fig. 2. Based on this definition, an improved SMOTE algorithm is proposed to improve the representation of over-sampled samples while reducing the mis-classification by kNN classification algorithm because of the newly generated samples marginalization.

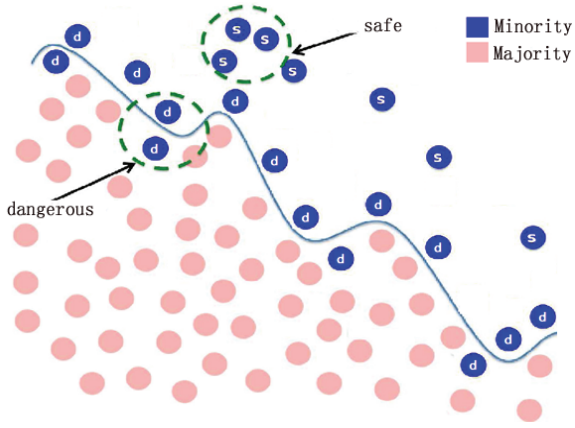

Fig. 2. Simple illustration of unbalanced data.

In order to solve the problem that the SMOTE [15] algorithm has not taken into account the blindness of the marginalization of the positive sample and the too single distribution interval of the newly generated sample set when the new sample is generated, this paper proposes the SC-SMOTE algorithm. The SC-SMOTE algorithm focuses on the problem of the confusion of region and edge generated by new samples to avoid new samples to further marginalize the data set distribution and sample confounding.

Firstly, input m positive sample sets $X_{positive}$ and n negative sample sets $X_{negative}$. Then, separately selecting the two types of monitoring variables, the yaw position and the cabin temperature, which have the highest correlation with blade icing, and naming it $X_i = \{X_{i1}, X_{i2}\}$ for subsequent calculation. Next divide positive sample set into three groups, and randomly select a positive sample X_i and name it X_{j1} from the positive sample set. Then select two positive samples X_{j+1}, X_{j+2} which are closest to them and name them X_{j2}, X_{j3} (If X_{j1}, X_{j2}, X_{j3} are collinear, then select the next positive sample X_{j+3} to join this group, and so on). Take these three positive samples as a group, and record this group of samples into the

loop table C , then find the next positive sample X_i that is not included in the group. Repeat the above process until $\lfloor m/3 \rfloor$ groups are divided. Finally, according to the geometric principle, a circumscribed circle can be uniquely determined by three non-collinear points, and the center $O(u,v)$ of the circle uniquely determined by three samples can be calculated:

$$\begin{cases}(X_{11}-u)^2+(X_{12}-v)^2=(X_{21}-u)^2(X_{22}-v)^2\\(X_{11}-u)^2+(X_{12}-v)^2=(X_{31}-u)^2(X_{32}-v)^2\end{cases} \tag{1}$$

This is simplified:

$$\begin{cases}(X_{11}-X_{21})u+(X_{12}-X_{22})v=(X_{11}^2-X_{21}^2)-(X_{22}^2-X_{11}^2)/2\\(X_{11}-X_{31})u+(X_{32}-X_{21})v=(X_{11}^2-X_{31}^2)-(X_{32}^2-X_{11}^2)/2\end{cases} \tag{2}$$

According to Cramer's rule, the center coordinate $O(u,v)$is:

$$\begin{cases}u=\frac{(X_{21}^2-X_{11}^2+X_{22}^2-X_{12}^2)(X_{12}-X_{32})/2-(X_{11}^2-X_{31}^2+X_{12}^2-X_{32}^2)(X_{12}-X_{22})/2}{(X_{11}-X_{21})(X_{12}-X_{32})-(X_{11}-X_{31})(X_{12}-X_{22})}\\v=\frac{(X_{11}^2-X_{31}^2+X_{12}^2-X_{32}^2)(X_{11}-X_{21})/2-(X_{11}^2-X_{21}^2+X_{12}^2-X_{22}^2)(X_{11}-X_{21})/2}{(X_{11}-X_{21})(X_{12}-X_{32})-(X_{11}-X_{31})(X_{12}-X_{22})}\end{cases} \tag{3}$$

Expand the center *x* to the same dimensions as the other samples in the current data set. Therefore, the center *O* is expanded to the same-dimensional data except for the angle values of three blades in the multidimensional information recorded with each time stamp of the SCADA monitoring set, that is, 23 dimensional:

$$O=\left(u,v,\frac{1}{3}\sum_{i=1}^{3}X_{i3},\frac{1}{3}\sum_{i=1}^{3}X_{i4},\ldots,\frac{1}{3}\sum_{i=1}^{3}X_{i23}\right) \tag{4}$$

Then, according to the order in the loop table C, three new positive samples are randomly produced according to the following formula by connecting the three points on the plane of a group of positive samples with the corresponding straight line of the center in the circle.

$$X_{new}=X_{jk}+rand(0,1)\times\left(O-X_{jk}\right) \tag{5}$$

where k=1, 2, 3 and $rand(0,1)$ represents a random number from 0 to 1. X_{new} represents the newly generated sample. Each time positive samples in a set of circles are produced, it is judged once whether the number of positive samples is equal to negative sample. If they are not equal, the next positive sample in the distance table is selected and three new positive samples are produced in the circle determined by the corresponding group. Then judge again until the number of positive samples is no longer less than the number of negative samples. Most of the new samples produced by this method will aggregate with similar samples better than those generated only on the two-sample lines and reduce the problem of data over-fitting because the generated region is single. It not only improves representative of the positive sample generation interval, but also widens the generation space of new sample, avoids the over-fitting problem of subsequent training model caused by data bunching, and provides clearer division conditions for the subsequent classifier algorithm.

The flow of the SC-SMOTE algorithm is shown in Fig. 3.

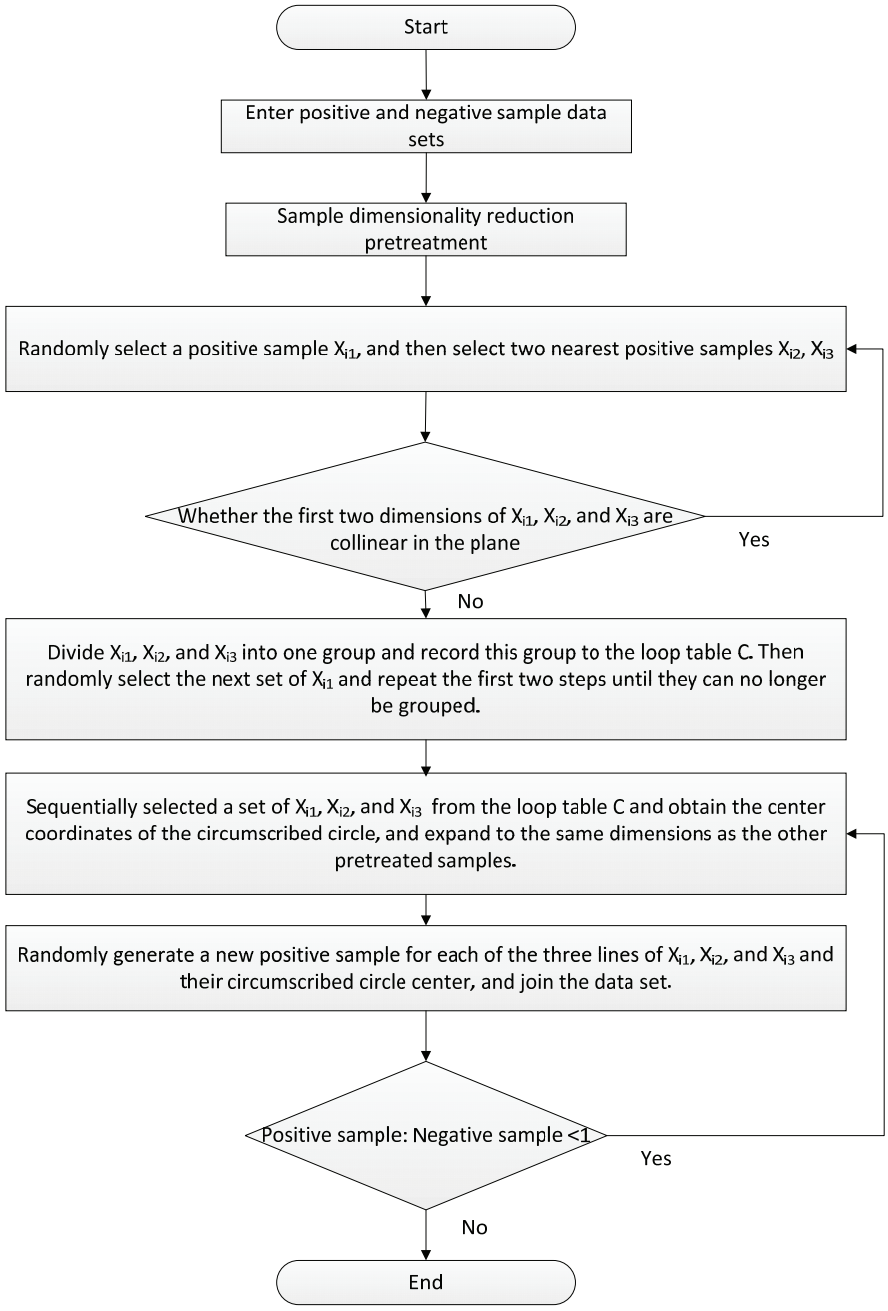

Fig. 3. The flow of the SC-SMOTE algorithm.

3.2 The Improved kNN Algorithm

The core idea of the kNN algorithm is: in the feature space, if the majority of *k* nearest neighbors of a sample belongs to a certain category, then the sample also belongs to this category and has the characteristic of this class sample. The algorithm is also particularly well suited for handling multi-label classification problems, the steps of the kNN algorithm are as follows:

$$d(x_i, x_j) = \sqrt{\sum_{i=1}^{n}(x_{in} - x_{jn})^2} \tag{6}$$

$$F(x_q) = arg\ max \sum_{i=1}^{k} \delta(v, f(x_i)) \tag{7}$$

(1) Construct a training sample set *T*.
(2) Set the initial value of *k*.
(3) Select the sample most similar to the test samples in the training sample set. The similar standard here is defined as the Euclidean distance. Assume that all samples belong to the n-dimensional space R^n and get sample $x_i = (x_{i1}, x_{i2}, \dots, x_{in},) \in R^n$ arbitrarily. Where x_{in} is the i^{th} feature value of the i^{th} sample. Define the dissimilarity degree between the samples x_i and x_j as $d(x_i, x_j)$.
(4) For the test sample $x_q, x_1, \dots, x_k$ are the k samples closest to x_q, set the discrete point objective function as $F: R^n \to v^i$, v^i is the i^{th} class label, the tag set is defined as $V = \{v^1, \dots v^s\}$. When $a = b, \delta(a, b) = 1$; otherwise $\delta(a, b) = 0$.
(5) Use majority vote as the predicted value of the sample to be tested.

In order to improve timeliness of data-driven fan blade icing fault diagnosis and reduce the computational complexity, this paper proposes a classification algorithm based on improved kNN, the diagnosis process is shown in Fig. 4.

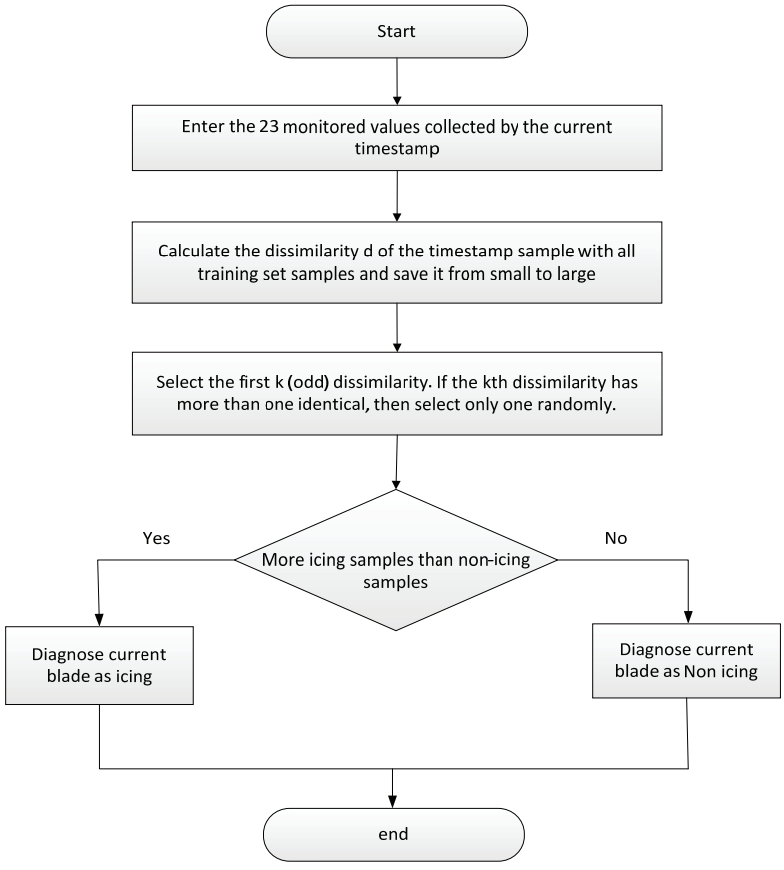

Fig. 4. Classification flow chart of improved kNN.

4. Experiments and Results Analysis

4.1 Experimental Data

In this paper, from the effective data set of this fan, odd-day two types of 24D data of odd days collected by the SCADA sensor system [16] and marked with or without icing are selected from 1:30:37 on December 1, 2017 to 0:01:01 on January 1, 2018 as the original data set. Then in icing condition data set of even days during this period, 10 pieces of data are randomly selected as the test data set, and sorted by time as shown in Table 1.

The data sets used for the specific experiments are: the original data set, the data sets of icing and non-icing conditions reached 1:1 by SMOTE algorithm and SC-SMOTE algorithm, respectively.

Table 1. Test data set

Number	Time stamp
1	2017/12/4 21:53:55
2	2017/12/4 21:55:12
3	2017/12/4 22:28:35
4	2017/12/14 3:35:46
5	2017/12/14 3:41:42
6	2017/12/14 4:07:12
7	2017/12/14 7:19:52
8	2017/12/14 7:31:13
9	2017/12/20 10:20:31
10	2017/12/20 11:36:10

4.2 Experimental Results and Analysis

Firstly, the initial data without over-sampling equilibration [17] is subjected to an icing fault diagnosis test by using diagnostic method based on kNN idea. Because classification effect based on kNN is closely related to the value of k, this paper refers to practical application conclusion and experience of kNN algorithm. In order to avoid the equal number of the nearest two types of data, the value of k generally takes an odd number, that is, the values of k are taken as 1, 3, 5, 7, 9, and 11, respectively. An experiment was conducted on each pair of test data sets, and the experimental result is shown in Table 2.

Table 2. Original test data set

Data set	Correct number	Misjudged number	Correct rate (%)
$k = 1$	3	7	30
$k = 3$	2	8	20
$k = 5$	1	9	10
$k = 7$	1	9	10
$k = 9$	0	10	0
$k = 11$	0	10	0

It can be seen that when the data set is seriously unbalanced, the accuracy of classification diagnosis based on kNN idea is generally low on the secondary data set.

Next, the original data set was subjected to 100 oversampling equilibrations by SMOTE and SC-SMOTE. On each of the balanced data sets, the values of k were also taken as 1, 3, 5, 7, 9, and 11,

respectively, and each of the data in the test data set was subjected to an icing fault diagnosis, and a total of 12,000 tests were performed. The experiment found that the best effect is achieved when k is 3, and the performance of the test data set in each of the balanced data sets is shown in Fig. 5.

The specific indicators of this round of experiment result are shown in Table 3.

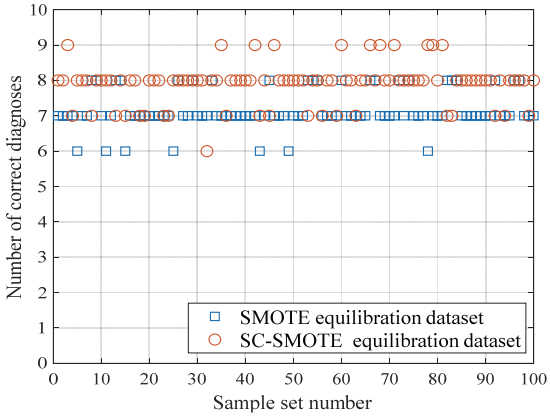

Fig. 5. Comparison of experimental effect when $k = 3$.

Table 3. Comparison of experimental results of two hundred equilibrated data sets

		SMOTE	SC-SMOTE
$k = 3$	Single maximum correct number	8	9
	Single minimum correct number	6	6
	The number of correct times is 6	7	1
	The number of correct times is 7	71	20
	The number of correct times is 8	22	68
	The number of correct times is 9	0	11
	Average correct rate (%)	71.5	78.9

The average accuracy for different values of k is shown in Fig. 6.

It is not difficult to find through experiments that the SC-SMOTE algorithm first randomly selects the positive class 1, and then finds the two nearest samples that are closest to them, and divides them into a group. Then, from the unclassified positive class samples, randomly select the positive class 2, and also find the two ungrouped similar samples that are closest to them, and divide them into a group. The grouping is stopped until there are less than three ungrouped samples. Then, according to the coordinates of the three selected samples of each group, the center coordinates of the uniquely determined circumscribed circle are calculated, and then randomly generated on the straight line connecting the three positive samples in each group with the corresponding circumscribed circle center to generate new positive class sample.

The new samples generated by this method generally have a tendency to move closer to the "safe area" with high sample density, and reduce the degree of blurring of the edge areas of the positive and negative samples. At the same time, as the selectable region for generating new samples increases, the new positive samples are generated too much because the SMOTE algorithm is always on the direct connection of a single two positive samples, resulting in distribution of a newly generated positive class samples is too monotonous and dense, which is likely to cause subsequent training of the sample set after this

oversampling balance, and the potential impact of over-fitting occurs, in order to improve the accuracy of the classification algorithm on the balanced sample set.

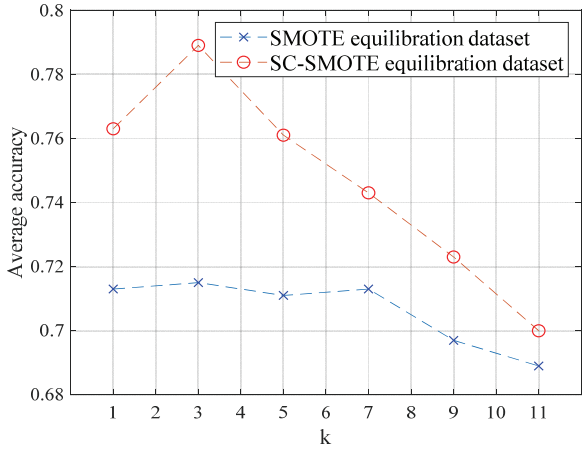

Fig. 6. Comparison of experimental effect of different *k* value.

According to the above experiment, it can be concluded that when *k* is 3, the diagnostic effects of both types of over-sampling algorithms are the highest. At the same time, the improvement of SMOTE algorithm makes the accuracy of classification diagnosis generally higher than the original over-sampling algorithm, which is effective for improving the subsequent classification diagnosis. The data based on SC-SMOTE algorithm equilibration has a certain improvement in the diagnosis of icing faults.

5. Conclusion

In view of the shortcomings of SMOTE algorithm in generating new samples in the edge region, such as edge indistinct, single interval of new samples, affecting the original distribution, an improved SMOTE over-sampling algorithm is proposed innovatively in this paper, aiming at improving the distribution interval of new samples and reducing the blindness of the interval. The algorithm is combined with kNN algorithm and applied in the wind turbine blade icing fault diagnosis experiment. A large number of experiments show that the proposed method can diagnose the ice fault of wind turbine blades very well, and compared with SMOTE algorithm; the ice fault diagnosis accuracy of SC-SMOTE algorithm on the balance set is higher, they are 78.9% and 71.5%, respectively.

However, there are still many improvements can be done in the future: the data from the industrial site has low quality. Therefore, the data preprocessing is very important. And it may improve the proposed model performance by taking different operation stages of wind turbines into consideration.

Acknowledgement

This paper is supported by Natural Science Foundation of China (No. 61871432, 61771492), the Natural Science Foundation of Hunan Province (No. 2020JJ4275, 2017JJ3065, 2019JJ6008, 2019JJ

60054), 2018 China Scholarship Council higher education teaching method research project, and 2017 Zhuzhou Science and Technology project.

References

[1] J. Butler-Adam, "The fourth industrial revolution and education," *South African Journal of Science*, vol. 114, no. 5-6, article no. a0271, 2018.

[2] L. Alhmoud, "Reliability improvement for a high-power IGBT in wind energy applications," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 9, pp. 7129-7137, 2018.

[3] Z. Bo, Z. Yanan, and C. Changzheng, "Acoustic emission detection of fatigue cracks in wind turbine blades based on blind deconvolution separation," *Fatigue & Fracture of Engineering Materials & Structures*, vol. 40, no. 6, pp. 959-970, 2017.

[4] Y. S. Jeong and J. H. Park, "Advanced big data analysis, artificial intelligence & communication systems," *Journal of Information Processing Systems*, vol. 15, no. 1, pp. 1-6, 2019.

[5] Y. S, Jeong and J. H. Park, "Artificial intelligence for the fourth industrial revolution," *Journal of Information Processing Systems,* vol. 14, no. 6, pp. 1301-1306, 2018.

[6] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, "Learning k for KNN classification," *ACM Transactions on Intelligent Systems and Technology*, vol. 8, no. 3, article no. 43, 2017.

[7] Y. Wang, X. Zhou, L. Liang, M. Zhang, Q. Zhang, and Z. Niu, "Short-term wind speed forecast based on least squares support vector machine," *Journal of Information Processing Systems*, vol. 14, no. 6, pp. 1385-1397, 2018.

[8] S. Sun, H. Xu, P. Fu, and J. Cai, "Impact of blade icing on wind turbines," *Wind Energy*, vol. 2014, no. 9, pp. 100-103, 2014.

[9] Q. Dong, Z. Jin, and Z. Yang, "A review of icing effect on horizontal axis wind turbine," *Machinery Design and Manufacture*, vol. 2014, no. 10, pp. 269-272, 2014.

[10] C. Puyals, S. Dunnett, and S. Zhang, "Research on the causes of performance degradation of wind farm and fault warning of wind turbine components by SCADA data analysis," *Wind Energy Industry*, vol. 2018, no. 8, pp. 65-68, 2018.

[11] N. Li, T. Yan, N. Li, D. Kong, Q. Liu, and Y. Lei, "Ice detection method by using SCADA data on wind turbine blades," *Refrigeration Air Conditioning & Electric Power Machinery*, vol. 2018, no. 1, pp. 58-62, 2018.

[12] S. Wang, Y. Li, K. Tagawa, and F. Feng, "A wind tunnel experimental study on icing distribution of rotating blade," *Journal of Engineering Thermophysics*, vol. 38, no. 6, pp. 1229-1236, 2017.

[13] B. Yang, R. Liu, and X. Chen, "Sparse time-frequency representation for incipient fault diagnosis of wind turbine drive train," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 11, pp. 2616-2627, 2018.

[14] Y. Guo, X. Chen, S. Wang, R. Sun, and Z. Zhao, "Wind turbine diagnosis under variable speed conditions using a single sensor based on the synchrosqueezing transform method," *Sensors*, vol. 17, article no. 1149, 2017.

[15] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Information Sciences*, vol. 465, pp. 1-20, 2018.

[16] J. H. Park, "Advances in multimedia computing and security and the introduction of new senior Editors," *Journal of Information Processing Systems*, vol. 12, no. 4, pp. 499-504, 2016.

[17] S. S. Kang, "Word similarity calculation by using the edit distance metrics with consonant normalization," *Journal of Information Processing Systems*, vol. 11, no. 4, pp. 573-582, 2015.

Cheng Peng **https://orcid.org/0000-0002-1920-7488**

He received M.S. and PhD. degrees in School of Information Science and Engineering from Central South University in 2010 and 2013, respectively. Since July 2013, he is with the School of Computer Science from Hunan University of Technology as a teacher and he is associate professor. At present, he is working as a post-doctor in the automation and control major of Central South University; his current research interests include industry big data analysis and industry equipment health management.

Qing Chen **https://orcid.org/0000-0001-9254-0837**

She is a Professor in the School of Computer, Hunan University of Technology. She received the Ph.D. in School of Information Science and Engineering, Central South University Changsha, China. Her research interests include industry big data analysis, equipment health state evaluation.

Longxin Zhang **https://orcid.org/0000-0002-4413-9974**

He received M.S. and Ph.D. degrees in School of information science and Engineering from Hunan University in 2010 and 2015, respectively. Since July 2015, he is with the School of Computer Science from Hunan University of Technology as a teacher. His current research interests include industry big data analysis and high-performance computation.

Lanjun Wan **https://orcid.org/0000-0001-7236-3589**

He received Ph.D. degrees in School of information science and Engineering from Hunan University 2015. Since July 2015, he is with the School of Computer Science from Hunan University of Technology as a teacher. His current research interests include industry big data analysis and high-performance computation.

Xinpan Yuan **https://orcid.org/0000-0001-9509-0755**

He received M.S. and Ph.D. degrees in School of information science and Engineering from Hunan University in 2008 and 2012, respectively. Since July 2014, he is with the School of Computer Science from Hunan University of Technology as a teacher. His current research interests include industry big data analysis and text analysis.