Journal of Intelligent Manufacturing (2020) 31:1467–1481
https://doi.org/10.1007/s10845-019-01522-8

Imbalanced fault diagnosis of rotating machinery via multi-domain feature extraction and cost-sensitive learning

Qifa Xu[1,2] · Shixiang Lu[1] · Weiyin Jia[3] · Cuixia Jiang[1]

Received: 2 August 2019 / Accepted: 7 December 2019 / Published online: 14 December 2019

Abstract

Fault diagnosis plays an essential role in rotating machinery manufacturing systems to reduce their maintenance costs. How to improve diagnosis accuracy remains an open issue. To this end, we develop a novel framework through combined use of multi-domain vibration feature extraction, feature selection and cost-sensitive learning method. First, we extract time-domain, frequency-domain, and time-frequency-domain features to make full use of vibration signals. Second, a feature selection technique is employed to obtain a feature subset with good generalization properties, by simultaneously measuring the relevance and redundancy of features. Third, a cost-sensitive learning method is designed for a classifier to effectively learn the discriminating boundaries, with an extremely imbalanced distribution of fault instances. For illustration, a real-world dataset of rotating machinery collected from an oil refinery in China is utilized. The extensive experiments have demonstrated that our multi-domain feature extraction and feature selection can significantly improve the diagnosis accuracy. Meanwhile, our cost-sensitive learning method consistently outperforms the traditional classifiers such as support vector machine (SVM), gradient boosting decision tree (GBDT), etc., and even better than the classification method calibrated by six popular imbalanced data resampling algorithms, such as the Synthetic Minority Over-sampling Technique (SMOTE) and the Adaptive Synthetic sampling method (ADASYN), in terms of decreasing missed alarms and reducing the average cost. Owing to its high evaluation scores and low average misclassification cost, cost-sensitive GBDT (CS-GBDT) is preferred for imbalanced fault diagnosis in practice.

Keywords Rotating machinery · Fault diagnosis · Imbalanced classification · Feature extraction · Cost-sensitive learning

Introduction

The breakthrough development of AI technology and millions of machineries equipped with smart sensors are accelerating the transformation from traditional manufacturing industry towards smart manufacturing (Dou et al. 2018; Liu et al. 2018b). Rotating machinery fault diagnosis plays an indispensable role in smart manufacturing, and the demand for effective diagnosis of its operation condition is increasing rapidly. With a large number of high quality and reliable real-time equipment operation data collected, it enables to construct an automatic monitoring, intelligent diagnosis and prognosis system of rotating machinery, which can reduce the maintenance cost significantly (Tao et al. 2018; Wu et al. 2019a; Sánchez et al. 2018).

There are numerous useful data acquisition techniques adopted in the fault diagnosis of rotating machinery, including vibration analysis, oil analysis, acoustic emission method, temperature monitoring, and microwave flaw detection. In

✉ Cuixia Jiang
jiangcuixia@hfut.edu.cn

Qifa Xu
xuqifa@hfut.edu.cn

Shixiang Lu
lushixiang@mail.hfut.edu.cn

Weiyin Jia
weiyin.jia@ronds.com.cn

[1] School of Management, Hefei University of Technology, Hefei 230009, Anhui, People's Republic of China

[2] Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, Hefei 230009, Anhui, People's Republic of China

[3] Anhui Ronds Science & Technology Incorporated Company, Hefei 230088, People's Republic of China

practice, vibration analysis is popular for its solid theoretical foundation and mature measurement tool (Ben Ali et al. 2018; Zhao and Lin 2018; Gan et al. 2018). However, the high frequency dynamic signal of rotating machinery is usually a superposition of multiple components of different amplitudes, and presents non-stationarity and non-linearity (Zhao et al. 2019; Amrhein et al. 2016). Thus, it calls for more effective and robust methods to extract features from vibration signals.

Apart from feature extraction, designing an irreplaceable fault diagnosis method is another crucial step in rotating machinery diagnosis. Fault diagnosis methods can be generally categorized into two types: mechanism-based methods and data-driven methods (Ren et al. 2018; Han et al. 2019a). Mechanism-based methods are employed only when accurate mathematical models of the failure can be built (Wu et al. 2019b). In fact, they have been greatly limited in the diagnosis of rotating machinery, owing to the complexity of their internal structure and the diversity of the external operating environment. On the other hand, data-driven methods are especially powerful for the complex industrial processes, since they directly use condition monitoring data to infer mechanical failure without any assumption on the underlying failure mechanism (Kang 2018; Tidriri et al. 2016; Wang et al. 2019). Although data-driven methods are effective in fault diagnosis, most studies are far from the actual operating conditions of rotating machinery, because the balanced datasets they used are too few to get (Santos et al. 2015).

In practical industrial applications, the machinery works in a normal condition throughout the whole operation cycle, and there are seldom faults happening in its operating phases (Jia et al. 2018), which means that fault instances are seriously insufficient (Han et al. 2019a; Liu et al. 2018a; Jiang et al. 2019; Sun et al. 2007). The first China Industrial Big Data Innovation Competition in 2017 (http://www.industrial-bigdata.com) makes them more specific, e.g., the open access data contains 300,000 instances and the ratio of fault/normal is about 1/10, which is a typical imbalanced classification problem. Under the assumption of equal misclassification cost or balanced class distribution, traditional classifiers lose their ability to deal with rare classes (Jia et al. 2018; Zhang et al. 2019; Li et al. 2020). In the fault diagnosis of rotating machinery, missing the detection of a failure condition, a rare class in mechanical operation may cause a catastrophic accident, whereas misclassifying a normal condition as a failure can be verified by manual checking easily (Zhang et al. 2018). This reminds us the misclassification cost differs significantly between two different types of errors. Much more attention should be paid to missed alarm for its greater cost of misclassification than false alarm (Kuo et al. 2018; Han et al. 2019b; Zan et al. 2019).

In this paper, we propose a novel data-driven framework for rotating machinery diagnosis, comprising comprehensively extracting and selecting of a series of features from collected vibration signals, and effectively diagnosing the imbalanced operating condition of rotating machinery via cost-sensitive learning. For this purpose, multiple features are extracted through the combination of time-domain, frequency-domain, and time-frequency-domain methods to discover useful information from vibration signals. Then, feature selection reduces the extracted feature set to a more compact one. Additionally, the cost-sensitive learning method takes misclassification costs into consideration to calibrate the classification results, which aims to minimize the average cost. To illustrate the efficacy of our method, extensive experiments are conducted on a real-world dataset of rotating machinery collected from an oil refinery in China. With the same feature selection technique, our multi-domain feature extraction method is compared with several previous feature extraction studies and it has been proved that our method can significantly improve the accuracy of fault diagnosis. Meanwhile, our cost-sensitive learning shows its great advantages in reducing average misclassification cost, compared with the cost-insensitive learning (classifiers treat each type misclassification cost equally) and the classification methods calibrated by imbalanced data resampling. Moreover, a sensitivity experiment shows that the classification performance of our cost-sensitive learning method will be affected by different cost matrices. It is noteworthy that CS-GBDT (cost-sensitive gradient boosting decision tree) is preferred in imbalanced fault diagnosis of rotating machinery, for its relatively high evaluation scores and low average misclassification cost across different cost matrices.

The remainder of this paper proceeds as follows. “Related works” section briefly reviews some related works. “The framework” section presents the proposed framework including multi-domain feature extraction, feature selection and cost-sensitive learning method. In “The real-world application” section, we apply the novel framework to a real-world application and illustrate its superiority through extensive comparisons. “Conclusions” section summarizes and concludes the paper.

Related works

The fault diagnosis procedure can be roughly divided into two steps: extracting robust features and diagnosing the condition of the machine (Song et al. 2018). In the step of feature extraction, after the properties of the vibration signals being well-understood, researchers could employ their domain knowledge in signal processing to design suitable features. In the second step, some new classification techniques have been developed to effectively diagnose the fault type of rotating machinery with imbalanced data.

Feature extraction from vibration signals

Vibration signals are an important information source for feature extraction in fault diagnosis because they contain high frequency and time-varying information. Specifically, feature extraction from vibration signals can be characterized into three domains: time-domain, frequency-domain, and time-frequency-domain.

First, in the time-domain, different statistical features are extracted to represent how signal amplitude varies with respect to time. However, vibration signals produced by a machine contain many components and are often difficult to be observed in the time-domain. Therefore, it is unlikely that a fault will be detected by a simple visual inspection. Second, frequency-domain analysis is to extract features in a given frequency band, and fast Fourier transform (FFT) is considered as one of the most commonly used feature extraction technique (Seera et al. 2014). In addition, amplitude spectrum (Larsson et al. 2002), power spectrum (Jiang et al. 2018) and cepstrum (Hwang et al. 2009) are often used in some frequency-domain scenarios as well. Unfortunately, the transition from time-domain to frequency-domain is based on the hypothesis of stationarity, which is often violated in the stage of mechanical failure (Tao et al. 2018). Third, time-frequency-domain analysis jointly representing the information from both time-domain and frequency-domain, is good at dealing with non-stationary signals. The two most powerful and extensively used techniques are wavelet transform (WT) (Wang et al. 2017) and empirical mode decomposition (EMD) (Georgoulas et al. 2013). The remarkable merit of WT is that it has a good localization property in both time-domain and frequency-domain. It can automatically adjust the scale of the frequency components, so as to observe and analyze the arbitrary details of the signals. EDM depicts the oscillation structure and frequency component of each part of the signals by decomposing the signal into a set of orthogonal complete intrinsic mode functions (IMFs). However, their common drawback is the lack of a translation-invariant property in the vibration signal processing, thus generating some false components (Ciabattoni et al. 2018).

Moreover, considering that different domains have their own advantages and disadvantages, one starts to combine multiple domain features for effective fault diagnosis. For example, Zhang et al. (2018) and Ragab et al. (2019) utilize time-domain and time-frequency-domain features to classify multiple failure modes of rotating machinery. Zhang et al. (2012) extracts time-domain and frequency-domain features to diagnose different failure stages of a wind turbine gearbox. However, the combination only involves two domains and is relatively subjective without any convincing explanation (Ben Ali et al. 2018). Seldom literatures extract features from three domains to make full use of the meaningful information contained in vibration signals (Wu et al. 2019b).

Classification for imbalanced data

According to the different stages of introducing data handling techniques, the methods for imbalanced data classification can be grouped into three categories: prior training, during training, and after training methods.

The first type of methods deal with imbalanced data though changing the distribution of training sample, such as manually rebalancing training sample by over-sampling minority instances or under-sampling majority instances (Zhang et al. 2019; Xie et al. 2019), and adjusting the training sample by the misclassification cost of the instances. These methods are often used to convert the arbitrary cost-insensitive classifiers into a cost-sensitive equivalence, without modifying the underlying learning algorithm. Unlike the first type, the second one directly modifies the learning procedure to improve the sensitivity of the classifier toward minority classes (Mathew et al. 2018; Correa Bahnsen et al. 2015). However, this type of methods designing different algorithms according to different classifiers are neither efficient nor flexible in practice (Lee et al. 2012). The third type concerns only the estimating membership class probabilities generated by cost-insensitive classifiers, assigning each instance to its lowest risk class based on Bayes risk theory, such as MetaCost (Domingos 1999) and Bayes minimum risk (Khan et al. 2018). Although these methods have been applied to many classifiers achieving the bias towards the minority but costly classes, they are restricted and require the cost-insensitive classifiers that can produce accurate posteriori probability estimations of the training instances (Lee et al. 2012).

To sum up, owing to their flexibility and not limited by the posterior probability of the classifier, the first type of methods are more suitable for industrial equipment fault diagnosis (Zhang et al. 2018). There are two different strategies for changing the distribution of training sample: (1) converting the original imbalanced training sample into a balanced one. The random under-sampling, random over-sampling, synthetic minority over-sampling technique (SMOTE) (Chawla et al. 2002), SMOTE-borderline1, SMOTE-borderline2 (Mathew et al. 2018; Han et al. 2005) and the adaptive synthetic sampling method (ADASYN) (Haibo et al. 2008) are classical resampling methods. However, these methods do not take into account the unequal costs imposed by the imbalanced distributions and may have difficulty in achieving unbiased results (Castro and Braga 2013). (2) Generating new training sample according to the unequal misclassification costs of instances. Zadrozny et al. (2003) designs two different sampling methods: sampling-with-replacement and cost-proportionate rejection sampling.

However, the proportional sampling-with-replacement produces duplicated instances in the training process, resulting in over-fitting in model building. On the contrary, “rejection sampling” is able to avoid duplication. Despite cost-proportionate rejection sampling has shown its strength in product recommendations (Lee et al. 2012), credit card fraud detection (Correa Bahnsen et al. 2015), and biomedical engineering (Gardner and Xiong 2009), its application is very rare or absent in the manufacturing domain to the best of our knowledge.

The framework

The purpose of this paper is to develop a new framework for fault diagnosis of rotating machinery. As shown in Fig. 1, the whole framework involves three parts: (1) multi-domain feature extraction for maximizing the information value of vibration data, (2) feature selection for eliminating redundant features and identifying effective ones, and (3) cost-sensitive learning for minimizing misclassification costs. All of them play an indispensable role in actual fault diagnosis.

Feature extraction

To make full use of the information in the raw vibration signals, we decompose the high-frequency dynamic signals into time-domain, frequency-domain, and time-frequency-domain features, respectively.

Time-domain feature extraction

Time-domain analysis characterizes the vibration change of rotating machinery when a fault occurs by directly constructing statistics from the vibration waveform. In this paper, we produce five widely used statistical features, namely, mean value (MV), root-mean-square value (RMSV), skewness coefficient (SC), kurtosis coefficient (KC), and shape factor (SF). The detailed calculations are shown in Table 1, where $z(t)$ is the signal at time t and N_z is the sample size.

Table 1 Time-domain statistical features

Statistical feature	Definition
MV	$MV_z = \frac{1}{N_z}\sum_{t=1}^{N_z} z(t)$
RMSV	$RMSV_z = \sqrt{\frac{1}{N_z}\sum_{t=1}^{N_z} z^2(t)}$
SC	$SC_z = \frac{1}{RMSV_z^3}\sum_{t=1}^{N_z} (z(t) - MV_z)^3$
KC	$KC_z = \frac{1}{RMSV_z^4}\sum_{t=1}^{N_z} (z(t) - MV_z)^4$
SF	$SF_z = \frac{RMSV_z}{\frac{1}{N_z}\sum_{t=1}^{N_z} \lvert z(t)\rvert}$

Frequency-domain feature extraction

In frequency-domain analysis, spectrum analysis and envelope spectrum analysis are two widely used methods. Spectrum analysis is to transform the vibration signals of rotating machinery from time-domain to frequency-domain, so that it can detect the corresponding fault characteristic of frequency components. FFT technique plays an important role in extracting spectral features, which is defined as

$$z(f) = \int_{-\infty}^{+\infty} z(t)e^{-j2\pi ft}dt, \tag{1}$$

where $z(t)$ denotes time-domain signal at time t, $z(f)$ denotes frequency-domain signal at frequency f.

Envelope spectrum is a curve formed by connecting the peak points of amplitudes at different frequencies during a time period. As acclaimed in Jiang et al. (2016), when the local damages or defects of the rotating machinery happen, the process of the load will produce mutational decaying shock pulse force and high frequency natural vibration. Thus, the final vibration waveform of rotating machinery is represented as a complex amplitude modulation wave. Envelope demodulation can be realized by Hilbert transform (HT) and then FFT. The HT of a signal is defined as

$$H[z(t)] = \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{z(\tau)}{t-\tau}d\tau. \tag{2}$$

We combine $z(t)$ and $H[z(t)]$ to form a new analytic signal

$$g(t) = z(t) + jH[z(t)]. \tag{3}$$

The amplitude $A(t)$ of $g(t)$ is then obtained by

$$A(t) = \sqrt{z(t)^2 + H[z(t)]^2}. \tag{4}$$

After HT, envelope spectra could be obtained by FFT to extract the feature of defect frequencies.

Through the above spectrum analysis and envelope spectrum analysis, we extract 56 frequency-domain features, such as root-mean-square frequency (RMSF), root variance frequency (RVF), spectral kurtosis (SK). We do not present all of them here to save limited space.

Time-frequency-domain feature extraction

In time-frequency-domain, EMD is considered as one of the most powerful techniques to extract features of rotating machinery. It decomposes the original non-stationary signals into a series of stationary signals indicating the natural oscillatory mode, which is termed as intrinsic mode functions

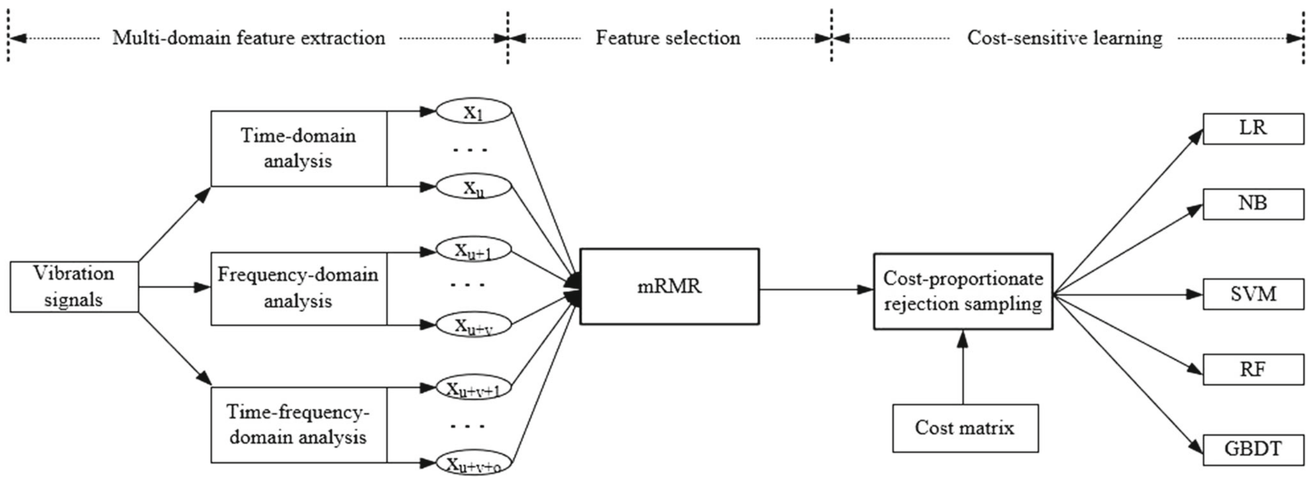

Fig. 1 Framework for fault diagnosis of rotating machinery

(IMFs). The procedure of intrinsic energy feature extraction from IMFs is described below.

Step 1, identify all local maxima and minima of the signal $z(t)$ separately, and then connect all the local maxima by a cubic spline line to form the upper envelope. Repeat the same procedure for the local minima points to form the lower envelope.

Step 2, denote the mean of the upper and lower envelope value as μ_1, and calculate the difference between the signal $z(t)$ and μ_1 as

$$\eta_1(t) = z(t) - \mu_1. \tag{5}$$

Step 3, determine whether $\eta_1(t)$ satisfies: (1) the number of extrema and the number of zero crossings are the same or at most one difference; and (2) its upper and lower envelopes are locally symmetric with respect to zero. If the two requirements satisfied, $\eta_1(t)$ becomes the first IMF component of signal $z(t)$ as

$$IMF_1(t) = \eta_1(t). \tag{6}$$

Otherwise, $\eta_1(t)$ will be regarded as the raw signal $z(t)$, and return to step 1.

Step 4, calculate the residual $r_1(t)$ with the form

$$r_1(t) = z(t) - IMF_1(t). \tag{7}$$

Stop decomposition if $r_1(t)$ is a monotone function. Otherwise, $r_1(t)$ is treated as the raw signal $z(t)$.

Step 5, repeat steps 1 to 4 and obtain all IMFs: $IMF_1(t)$, $IMF_2(t), \ldots, IMF_M(t)$, as well as the final residual $r_M(t)$. In this regard, the raw signal can be expressed as

$$z(t) = \sum_{j=1}^{M} IMF_j + r_M(t), \tag{8}$$

where M denotes the number of IMFs.

Step 6, define the intrinsic energy features of rotating machinery as

$$E_j = \frac{1}{L}\sum_{i=1}^{L}\left[IMF_j\left(t_i\right)\right]^2, \tag{9}$$

where L represents the number of instances in each IMF.

In general, the fault information of rotating machinery is mainly in high frequency band. As a result, the fault characteristic information can be represented by the first few IMF components. In this paper, we extract the first four IMFs' energy to represent time-frequency-domain features.

Feature selection

From a practical fault diagnosis point of view, some extracted features are redundant or non-significant (Ragab et al. 2019). Directly using all the extracted features in the model may lead to computation inefficiency, over-fitting, high maintenance workload, and difficulty of model interpretation. To address these issues, an effective subset that contains informative features should be selected.

There are quantities of feature selection methods that can reduce the original feature set to a more compact one. Such methods can be generally categorized into three types: filter methods, wrapper methods, and embedded methods (Han et al. 2019a; Ding and Peng 2005). In this paper, a filter method called min-redundancy and max-relevance (mRMR) (Peng et al. 2005) is employed. The advantages of the mRMR over the other feature selection methods can be illustrated from two aspects. First, as a filtering method, mRMR has the advantage of computational efficiency and the ability to generalize different machine learning models. Out of filter nature, the motivation to adopt mRMR is that it can effectively reduce the redundant features while keeping the relevant features in the model, not perturbing or hiding the physical meaning of the features. Therefore, it keeps the interpretability of the proposed diagnosis model, which is very important for the decision maker.

Actually, mRMR is a mutual-information-based feature selection method that simultaneously measures the relevance and redundancy of features, to obtain a feature subset with good generalization properties. The mutual information used to measure the dependency between two discrete random variables $X^{(1)}$ and $X^{(2)}$, can be expressed as

$$I(X^{(1)}, X^{(2)}) = \sum_{i,j} p(x_i^{(1)}, x_j^{(2)}) \log \frac{p(x_i^{(1)}, x_j^{(2)})}{p(x_i^{(1)})p(x_j^{(2)})}, \quad (10)$$

where $p(x_i^{(1)}, x_j^{(2)})$ denotes the joint probability density of two random variables $X^{(1)}$ and $X^{(2)}$. $p(x_i^{(1)})$ and $p(x_j^{(2)})$ denote the marginal probability density of $X^{(1)}$ and $X^{(2)}$, respectively.

For features variables, mutual information is used to measure the level of "similarity" between the extracted features. The principle of minimum redundancy is to select the features so that they are mutually maximally dissimilar. The minimum redundancy function can be expressed as

$$\min W_I = \frac{1}{|S|^2} \sum_{(g_i, g_j \in S)} I\left(g_i, g_j\right), \quad (11)$$

where S denotes the feature subset we are seeking, $|S|$ is the number of features in S, g_i is the ith feature in S. In addition, mutual information is also used to measure the discriminant ability of features to target variable y. $I\left(y, g_i\right)$ quantifies the relevance of g_i for classification tasks. Thus the maximum relevance function can be expressed as

$$\max V_I = \frac{1}{|S|} \sum_{(g_i \in S)} I\left(y, g_i\right). \quad (12)$$

In this paper, we adopt the strategy to combine the two functions into a single criterion one, named the mutual information quotient (MIQ), which maximizes the function $\max\left(V_I / W_I\right)$. Based on this criterion, a small subset is selected from the extracted features using the vibration signal. As we will see later, the 21 selected features show better diagnostic performance.

Imbalanced cost-sensitive classification

This subsection designs a cost-sensitive learning method for fault diagnosis of rotating machinery. It consists of a cost matrix that defines the cost of misclassification, and a cost-sensitive classification construction that calibrates imbalanced classification results.

Cost matrix determination

The effectiveness of cost-sensitive learning method depends strongly on the given cost matrix. Improperly initializing costs are not conducive to the learning process (Zhang and Hu 2014). In other words, too low costs are not enough to adjust the classification boundary, while too high costs may lead to poor generalization capacity of the classifier on other classes. In this paper, a handcrafted cost matrix based on expert knowledge is recommended to the trade-off.

In practice, the costs caused by misclassification of different fault conditions of rotating machinery are heterogeneous. For example, the cost of misclassification of fault condition into normal condition is much greater than a false alarm. Moreover, the misclassification faults of different components are also different. This usually requires access to industry experts who have the ability to assess the most realistic cost values. The cost of misclassification given by experts may differ under different fault-tolerant criteria. In order to popularize our methodology, we abstract two basic rules for the cost matrix.

Rule 1: all costs are nonnegative. The entry of cost matrix C is described as

$$C = \begin{cases} C_{p,q} = 0, & p = q \\ C_{p,q} \in \mathbb{N}^+, & p \neq q \end{cases} \quad (13)$$

where $C_{p,q}$ represents the cost of misclassifying the class p into the class q. In practice, there are four different situations.

- For $C_{p,q}$ $(p = q)$, it stands for correct classification. In this paper, we define $C_{p,q} = 0$ $(p = q)$.
- For $C_{p,q}$ $(p = 0, q = 1, 2, \ldots k)$, it represents a false alarm, which means classifying a normal condition as a fault condition.
- For $C_{p,q}$ $(p = 1, 2, \ldots k, q = 0)$, it represents a missed alarm, which means classifying a fault condition as a normal condition.
- For $C_{p,q}$ $(p \neq q, p \neq 0, q \neq 0)$, it means classifying one fault condition as another one.

Rule 2: the cost of misclassification satisfies

$$C_{0,q} < C_{p,q} < C_{p,0}, \quad (14)$$

where $p \neq q \neq 0$, and the subscript 0 stands for a normal condition. This rule presents the idea that the costs of different misclassifications are significantly different from each other, i.e., the cost of false alarm or type I error $(C_{0,q})$ is small, as only a confirmation is required by operator. The cost of missed alarm or type II error $(C_{p,0})$ is large, as it may lead to serious damage to equipment or even catastrophe. In addition, misclassifying one fault condition as another one can alert the operator with a failure signal, but it is difficult to make an accurate diagnosis. Therefore, $C_{p,q}$ is between $C_{0,q}$ and $C_{p,0}$.

Cost-sensitive learning by cost-proportionate rejection sampling

Cost-proportionate rejection sampling processes the training sample through proportional sampling and adjusts the outcome class distribution of the sampling instances. Then a cost-insensitive learning algorithm can be directly applied to the sampled instances. Here we're just going to give a brief overview of cost-proportionate rejection sampling, and the full details about this algorithm are well presented in Zadrozny et al. (2003).

In cost-sensitive learning phase, the objective is to learn a classifier $s : X \rightarrow Y$ to minimize the expected cost,

$$E_{x,y,c\sim D}[c \cdot I(s(x) \neq y)], \tag{15}$$

where (x, y, c) is the form of given training data, D denotes the distribution with domain $X \times Y \times C$, X denotes the input space, Y denotes the output space, C denotes the misclassification costs and $I(\cdot)$ is the indicator function.

Assuming that there exists a constant $N_c = E_{x,y,c\sim D}\lceil c \rceil$, the goal of minimizing the expected cost can be transformed into minimizing the ratio of errors under $\hat{D}$,

$$\begin{aligned} &E_{x,y,c\sim D}[c \cdot I(s(x) \neq y)] \\ &\quad= \sum_{x,y,c} D(x, y, c) \cdot c \cdot I(s(x) \neq y) \\ &\quad= N_c \cdot \sum_{x,y,c} \hat{D}(x, y, c) \cdot I(s(x) \neq y), \\ &\quad= N_c \cdot E_{x,y,c\sim \hat{D}}[I(s(x) \neq y)] \end{aligned} \tag{16}$$

where $\hat{D}(x, y, c) = \frac{c}{N_c} D(x, y, c)$. Specifically, the translation theorem in Eq. (16) indicates that each instance in the original training sample (S) is drawn independently once, and accepted into the sample $\left(S'\right)$ with the probability c/Z_c, where Z_c is a predefined constant. In this paper, the constant Z_c is chosen as the maximal misclassification cost. To determine whether to keep an instance, the algorithm first generates a random number r_v from the uniform distribution $U(0, 1)$, and compares it with the acceptance probability c_i/Z_c of the instance under evaluation. The instance i is accepted when r_v doesn't exceed its acceptance probability and rejected otherwise. Eventually, we obtain a new set $\left(S'\right)$ which is generally smaller than (S).

This method is very suitable for industrial practice owing to its two distinguished advantages: (1) it comes with a theoretical guarantee. The use of the sampled set $\left(S'\right)$ guarantees a cost-minimizing classifier, assuming the underlying learning algorithm can achieve an approximate minimization of classification errors (Zadrozny et al. 2003). (2) It is general and flexible in practice. Instances are selected from the original training sample according to their misclassification costs, and then used to construct a classifier with an adequate classification learning technique, which results in cost-sensitive learning (Lee et al. 2012).

We should note that cost proportionate rejection sampling is designed especially for binary cost-sensitive classification. However, our task is to distinguish multi-class data. The one-against-rest approach is adopted to solve our multiclassification problem (Beygelzimer et al. 2005). Meanwhile, cost-sensitive learning by cost-proportionate rejection sampling only concerns adjusting the distribution of training sample by misclassification costs, which implies that it follows the same procedure of training model and fine-tuning parameters as traditional cost-insensitive learning methods.

As mentioned above, this cost-sensitive learning is a general method, which can improve classification performance through combined use with traditional or cost-insensitive classifiers. In this paper, we consider five classical classifiers, logistic regression (LR), naive Bayes (NB), support vector machine (SVM), random forest (RF), and gradient boost decision tree (GBDT), to identify the operational status of rotating machinery. Moreover, we combine them with the cost proportional rejection sampling method to construct five cost-sensitive learning models: cost-sensitive logistic regression (CS-LR), cost-sensitive naive Bayes (CS-NB), cost-sensitive support vector machine (CS-SVM), cost-sensitive random forest (RF), and cost-sensitive gradient boost decision tree (CS-GBDT).

The real-world application

In this section, we apply the proposed framework to the fault diagnosis of rotating machinery by using the acquired vibration data from an oil refining in China. Its superiority has been demonstrated from two aspects. First, our multi-domain feature extraction has shown its advantages in vibration signal analysis when compared with some previous works. Second, the cost-sensitive learning method can effectively reduce the average cost of misclassification and performs well in fault diagnosis with imbalanced data. Meanwhile, a sensitivity analysis is conducted under different misclassification costs,

Table 2 Description of the fault type

No.	Fault type	Number of fault instances	Imbalanced ratio (normal/fault)
1	Bearing cage fault	36	8.25
2	Bearing inner race fault	35	8.49
3	Bearing outer race fault	27	11.00
4	Bearing rolling element fault	25	11.88
5	Coupling fault	24	12.38
6	Component loosening	54	5.50
7	Other working conditions with faults (e.g., insufficient lubrication, seal leakage, etc.)	35	8.49

and CS-GBDT is recommended for its better performance and more robust results.

Data description

The raw vibration data used here is collected from an oil refinery in Zibo, Shandong Province, China. We conduct feature extraction from vibration signals using time-domain, frequency-domain and time-frequency-domain analyses, and obtain 65 features eventually. In monitoring the pump with vibration signals for up to two years, we have gathered 7 fault types including bearing cage fault, bearing inner race fault, bearing outer race fault, bearing rolling element fault, coupling fault, component loosening and other working conditions with faults, as shown in Table 2. The imbalanced ratio (IR) in Table 2 is defined as the ratio of the number of normal condition (297) to the number of each fault type. It is obvious that we have to face a multi-classification problem with imbalanced data.

Multiclass performance metrics

The main purpose of this paper is to pursue the minimum cost of misclassification without reducing the classification performance of minority and majority classes. The widely used evaluation metrics in binary classification settings are accuracy, precision, recall, F1-score, and G-mean. Here, these evaluation metrics should be extended to meet the multi-classification need. Macro-averaging, generalizes the concept of these evaluation metrics to multiple dimensions, by averaging the metric values of each pair of classes (Santos et al. 2015). The specific metrics are expanded to:

$$\text{MA-A} = \frac{\sum_{i=0}^{k} \text{ True Positive }_i}{\sum_{i=0}^{k} (\text{ True Positive }_i + \text{ False Positive }_i)}, \quad (17)$$

$$\text{MA-P} = \frac{1}{k+1}\sum_{i=0}^{k} \frac{\text{True Positive}_i}{\text{True Positive}_i + \text{False Positive}_i}, \quad (18)$$

$$\text{MA-R} = \frac{1}{k+1}\sum_{i=0}^{k} \frac{\text{True Positive}_i}{\text{True Positive}_i + \text{False Negative}_i}, \quad (19)$$

$$\text{MA-F} = \frac{2}{1/\text{MA-P} + 1/\text{MA-R}}, \quad (20)$$

$$\text{MA-G} = \sqrt[k+1]{\prod_{i=0}^{k} \frac{\text{True Positive}_i}{\text{True Positive}_i + \text{False Negative}_i}}, \quad (21)$$

where $k + 1$ denotes the total number of classes. MA-A, MA-P, MA-R, MA-F, and MA-G denote macro-averaging accuracy, precision, recall, F1 and G-mean, respectively.

Although the above five metrics are widely used to evaluate the classification performance, the actual cost of each type of misclassification is not taken into account. To address this issue, the average cost (A-cost), joint consideration for the number and cost of misclassification error in different conditions, is the highlight of imbalanced classification metric (Zhou and Liu 2006). In general, it is defined as

$$\text{A-cost} = \frac{1}{N_s}\sum_{p=0}^{k}\sum_{q=0}^{k} C_{p,q} \cdot N_{p,q}, \quad (22)$$

where $N_{p,q}$ represents the number of misclassifying the class p into the class q, N_s represents the total number of instances.

Comparison with other feature extraction methods

To illustrate the effectiveness of our multi-domain feature extraction, it is necessary to compare the extracted features with those used in previous works. We introduce three feature sets widely used in fault diagnosis of rotating machinery: feature set 1 (FS1), feature set 2 (FS2), and feature set 3 (FS3), and then compare them with our proposed feature set (PFS).

(1) FS1: Time-domain and time-frequency-domain features.
(2) FS2: Time-domain and frequency-domain features.

Table 3 The specified subset of the hyper-parameter space

Model	Hyperparameter	Search area
LR	L2-norm parameter	$[0.1, 0.2, 0.3, \ldots, 0.8, 0.9, 1.0]$
NB	Smoothing parameter	$[0, 0.1, 0.2, \ldots, 1.8, 1.9, 2.0]$
SVM	Penalty parameter	$\left[10^{-4}, 10^{-3}, 10^{-2}, \ldots, 10^{2}, 10^{3}, 10^{4}\right]$
	Kernel parameter	$[0.1, 0.2, 0.3, \ldots, 1.8, 1.9, 2.0]$
RF	Number of estimators	$[10, 20, 30, \ldots, 80, 90, 100]$
	Maximum depth	$[None, 2, 4, \ldots, 16, 18, 20]$
GBDT	Number of estimators	$[20, 30, 40, \ldots, 70, 80, 90]$
	Maximum depth	$[2, 4, 6, \ldots, 12, 14, 16]$
	Minimum samples in the leaf node	$[10, 20, 30, \ldots, 80, 90, 100]$

(3) FS3: Frequency-domain and time-frequency-domain features.

(4) PFS: Time-domain, frequency-domain, and time-frequency-domain features.

As mentioned in “Feature selection” section, directly using each group of all features may lead to over-fitting and affect its classification performance. To overcome the interference of redundancy and irrelevant features, we use mRMR to select important features in each group and feed them into the five classical classifiers introduced in “Cost-sensitive learning by cost-proportionate rejection sampling” section. To better show the role of our feature selection, we investigate the impact of different numbers of features on the experimental results. In each group of experiments, the sample data is randomly partitioned into training set and test set by 6:4. Based on the training set, the parameters of each classifier are tuned by five-fold cross-validation (80 % for training and 20% for testing at each CV iteration) independently. We adopt the grid search approach, an exhaustive searching through a manually specified subset of the hyper-parameter space of a learning algorithm, to select the optimal hyper-parameters. The main hyper-parameter search space for each model is listed in Table 3.

The diagnosis performance results are shown in Fig. 2, where the x-axis represents the number of important features selected using mRMR. It is clear that as the number of important features increases, the classification performance improves significantly in the early stage, and then tends to be stable. In addition, we should note that when too many features are added, the performance of some classifiers begins to decline gradually. For example, when the number of selected features in PFS exceeds 31, there is an obvious over-fitting phenomenon in SVM, which verifies the importance of feature selection. Based on the overall classification performance of each classifier, 21 important features are finally selected from the extracted features using the vibration signals for later use.

By comparing four different feature sets, we observe that all the classifiers have achieved relatively good prediction performance using PFS in terms of higher scores. In other words, PFS extracted from three different domains can better represent the operational status of rotating machinery. Meanwhile, GBDT has higher scores on all five metrics of each feature set and is less likely to be disturbed by noise or irrelevant features, compared with the other four classical classifiers. It illustrates the accuracy and robustness of GBDT in this practical classification.

Comparison with traditional data-driven methods

To verify that our proposed cost-sensitive learning method is superior to the classical cost-insensitive learning and other imbalanced learning methods in fault diagnosis, two comparative experiments are conducted. The important features selected by mRMR are chosen as input variables, whose effectiveness has been well demonstrated by the experiments in “Comparison with other feature extraction methods” section.

Comparison with classical cost-insensitive learning methods

We investigate the classification performance of cost-sensitive learning based on a given cost matrix. Specifically, we use the five cost-sensitive models proposed in “Cost-sensitive learning by cost-proportionate rejection sampling” section, namely, CS-LR, CS-NB, CS-SVM, CS-RF, and CS-GBDT. Accordingly, to illustrate the calibration effect of cost-sensitive learning based on misclassification costs, the five classical cost-insensitive classification models (LR, NB, SVM, RF, and GBDT) are selected as a benchmark. Additionally, based on the determination rules of cost matrix in “Cost matrix determination” section, we define a cost matrix with entries

$$C_{i0} = 10w, \quad C_{0j} = w, \quad \text{and} \quad C_{ij} = 3w \quad \text{for } i \neq j \neq 0, \tag{23}$$

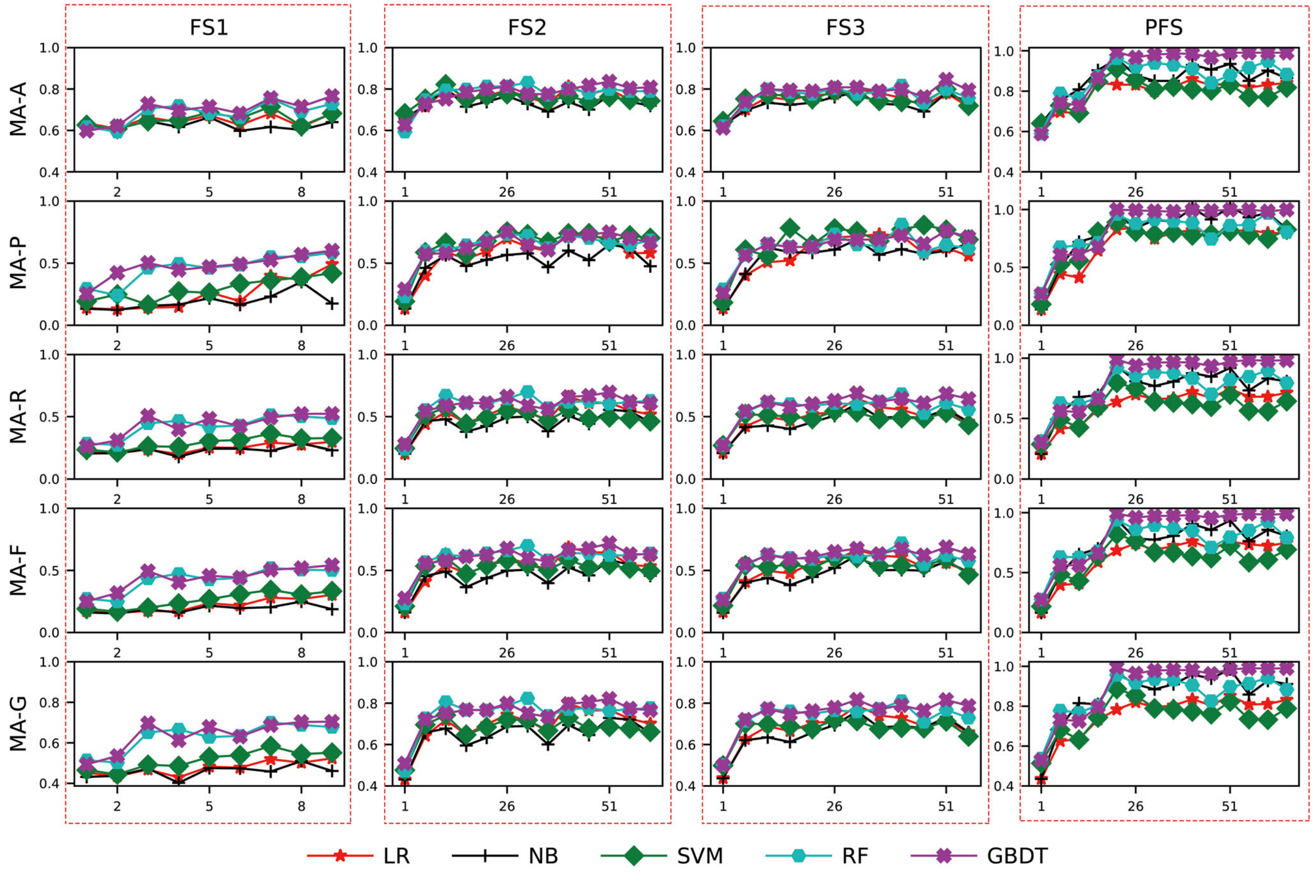

Fig. 2 The comparison of classification performance on different feature sets

where w denotes the minimum misclassification cost unit. This definition indicates that the cost of missed alarm is ten times higher than that of false alarm, while the cost of misclassifying one fault as another one is much smaller. It highlights the dangers and seriousness of missed alarm in fault diagnosis.

In this experiment, the sample data is randomly divided into training set and test set by 6:4. Based on training set, grid search and five-fold cross-validation are used for each model to select the optimal hyper-parameters. The specific hyper-parameter spaces of each model are listed in Table 3. The average results with standard deviations in parentheses on test set across 50 independent experiments are listed in Table 4.

From the results in Table 4, some interesting findings emerge. First, cost-sensitive learning method outperforms cost-insensitive learning one, which validates the need for cost-sensitive learning in fault diagnosis. Taking LR for example, cost-sensitive learning method increases the MA-A metric from 0.8326 to 0.8458, the MA-R metric from 0.6736 to 0.7602, the MA-F metric from 0.6959 to 0.7321, and the MA-G metric from 0.8063 to 0.8618. Meanwhile, the average misclassification cost (A-cost) is halved. Second, CS-GBDT is superior to the other cost-sensitive methods in terms of highest gains and lowest loss. As for A-cost, CS-GBDT reduces it to 0.0266, which is only 1/17 to 1/3 of those results from CS-LR, CS-NB, CS-SVM, and CS-RF. Third, we find that GBDT without cost-sensitive learning calibration also performs much better than CS-LR, CS-NB, CS-SVM and CS-RF.

The results show that the fault diagnosis with imbalanced data are not only highly dependent on cost-sensitive learning, but also vary with different classifiers. Thus, it calls for a framework with flexibility and strong generalization ability, to choose the best cost-sensitive model from different classifiers in practical applications. Our designed framework meets the demand quite well, since it is convenient for operators of industrial equipment diagnosis to calibrate the classification results given on a cost matrix, without spending a lot of time and resources in reconstructing the training and fine-tuning procedures of classifiers.

In addition to reporting the average misclassification cost across 50 experiments in Table 4, we further test the significant improvement in prediction ability of each model under cost-sensitive calibration. To this end, we use the Diebold-Mariano (DM) test (Diebold and Mariano 1995) to compare the average misclassification cost of cost-insensitive learning (model 1) with cost-sensitive learning (model 2). It is note-

Table 4 Experimental results of cost-insensitive learning and cost-sensitive learning

Metrics	Cost-insensitive					Cost-sensitive				
	LR	NB	SVM	RF	GBDT	CS-LR	CS-NB	CS-SVM	CS-RF	CS-GBDT
MA-A	0.8326	0.9672	0.8541	0.9378	0.9901	**0.8458**	**0.9677**	0.8505	**0.9695**	**0.9921**
	(.0230)	(.0138)	(.0223)	(.0242)	(.0087)	**(.0206)**	**(.0131)**	(.0220)	**(.0168)**	**(.0090)**
MA-P	0.7756	0.9536	0.8021	0.9056	0.9965	0.7394	**0.9600**	0.7578	**0.9649**	0.9907
	(.0502)	(.0206)	(.0280)	(.0366)	(.0040)	(.0362)	**(.0120)**	(.0378)	**(.0237)**	(.0108)
MA-R	0.6736	0.9375	0.7069	0.8840	0.9791	**0.7602**	**0.9412**	**0.7467**	**0.9433**	**0.9837**
	(.0420)	(.0251)	(.0405)	(.0459)	(.0170)	**(.0334)**	**(.0243)**	**(.0378)**	**(.0318)**	**(.0172)**
MA-F	0.6959	0.9423	0.7175	0.8869	0.9868	**0.7321**	**0.9437**	**0.7271**	**0.9511**	0.9860
	(.0405)	(.0231)	(.0389)	(.0436)	(.0109)	**(.0344)**	**(.0217)**	**(.0380)**	**(.0278)**	(.0152)
MA-G	0.8063	0.9655	0.8309	0.9352	0.9881	**0.8618**	**0.9680**	**0.8546**	**0.9682**	**0.9912**
	(.0270)	(.0140)	(.0254)	(.0261)	(.0098)	**(.0200)**	**(.0133)**	**(.0230)**	**(.0180)**	**(.0093)**
A-cost	1.0412	0.1508	0.5720	0.2610	0.0932	**0.4863**	**0.0908**	**0.4566**	**0.1746**	**0.0266**
	(.2000)	(.0725)	(.1311)	(.1270)	(.0848)	**(.0852)**	**(.0399)**	**(.0788)**	**(.1139)**	**(.0337)**

Boldface indicates the performance of cost-sensitive learning over the corresponding cost-insensitive learning, i.e., LR versus CS-LR, NB versus CS-NB, SVM versus CS-SVM, RF versus CS-RF, GBDT versus CS-GBDT

Table 5 DM significance test of misclassification cost (A-cost)

Model comparison	DM test	
(Model 2 vs. model 1)	Statistic	*p*-value
CS-LR versus LR	15.6150	2.2e−16***
CS-NB versus NB	6.4440	4.823e−08***
CS-SVM versus SVM	5.7616	5.44e−07***
CS-RF versus RF	3.0786	0.0034***
CS-GBDT versus GBDT	4.7952	1.556e−05***

***Statistical significance at the 1% level

worthy that the argument ‘alternative’ in R function ‘dm.test’ is set to be ‘two.sided’. In other words, the alternative hypothesis is that the average misclassification cost of model 1 is not equivalent to that of model 2. If p-value is less than 1%, we accept the alternative hypothesis and deem that model 2 is statistically superior to model 1 when the DM statistic is positive, while model 1 is statistically superior to model 2 when the DM statistic is negative. Otherwise, we have to accept the null hypothesis that there is no difference between model 1 and model 2.

We report in Table 5 the DM test results and find that all statistics are positive and significant at the level of 1%. This confirms that the feasibility of our cost-sensitive method to reducing the misclassification cost in fault diagnosis of rotating machinery.

Comparison with other imbalanced learning methods

We have proved the advantages of cost-sensitive learning methods in fault diagnosis over the traditional cost-insensitive classification methods through the above experiments. In order to further verify the superiority of our cost-sensitive method in dealing with imbalanced data, we conduct another experiment by comparing our cost-sensitive method with a series of rebalancing methods. To this end, we select six commonly used rebalancing methods, including random under-sampling, random over-sampling, SMOTE, SMOTE-borderline1, SMOTE-borderline2, and ADASYN. They are applied to convert an imbalanced data into a balanced one, which will be used to train the five basic classifiers. For fair comparisons, each classifier takes the same procedure using grid search and five-fold cross-validation to select the optimal hyper-parameters, and repeats 50 independent experiments. The specific hyper-parameter spaces of each classifier are listed in Table 3.

We use the cost-sensitive learning method as the benchmark and report in Table 6 the classification results of each classifier with different rebalancing methods. The results show that the overall performance of each classifier based on cost-sensitive calibration is better than the six rebalancing methods, except for only a few metrics. For example, SVM with Random over-sampling can obtain the maximum average precision. Taking a closer look at the results of SMOTE-borderline2, we find that each classifier trained using the synthesized data has poor performance, especially for NB. The possible reason is that the generation of synthetic instances in SMOTE-borderline2 will be hampered by the skewed distribution of the “danger instance”. Meanwhile, the poor performance of random under-sampling may be caused by losing quantities of instances information.

Table 6 Experimental results of other imbalanced learning methods

Metrics	Models	Cost-sensitive	Random under-sampling	Random over-sampling	SMOTE	SMOTE-borderline1	SMOTE-borderline2	ADASYN
MA-A	LR	**0.8458 ± .0206**	0.7911 ± .0419	0.8328 ± .0270	0.8354 ± .0235	0.8342 ± .0250	0.8120 ± .0287	0.8293 ± .0228
	NB	0.9677 ± .0131	0.9349 ± .0196	**0.9678 ± .0140**	0.9663 ± .0134	0.9668 ± .0138	0.5783 ± .1769	0.9660 ± .0158
	SVM	**0.8505 ± .0220**	0.7390 ± .0410	0.8293 ± .0237	0.8173 ± .0249	0.8133 ± .0240	0.8165 ± .0250	0.8154 ± .0253
	RF	**0.9695 ± .0168**	0.8412 ± .0525	0.9450 ± .0211	0.9473 ± .0196	0.9393 ± .0226	0.9409 ± .0200	0.9495 ± .0176
	GBDT	**0.9921 ± .0090**	0.9307 ± .0511	0.9919 ± .0074	0.9900 ± .0108	0.9913 ± .0088	0.9803 ± .0167	0.9908 ± .0081
MA-P	LR	0.7394 ± .0362	0.6921 ± .0341	**0.7408 ± .0368**	0.7391 ± .0348	0.7375 ± .0402	0.6912 ± .0451	0.7269 ± .0335
	NB	**0.9600 ± .0120**	0.9021 ± .0312	0.9553 ± .0199	0.9538 ± .0196	0.9547 ± .0200	0.6314 ± .0716	0.9533 ± .0202
	SVM	**0.7578 ± .0378**	0.6473 ± .0443	0.7536 ± .0456	0.7498 ± .0467	0.7447 ± .0430	0.7198 ± .0463	0.7484 ± .0432
	RF	**0.9649 ± .0237**	0.7828 ± .0523	0.9348 ± .0304	0.9285 ± .0276	0.9203 ± .0337	0.9015 ± .0362	0.9374 ± .0268
	GBDT	**0.9907 ± .0108**	0.9231 ± .0464	0.9863 ± .0122	0.9844 ± .0173	0.9863 ± .0138	0.9672 ± .0252	0.9855 ± .0135
MA-R	LR	0.7602 ± .0334	0.7432 ± .0379	0.7907 ± .0398	**0.7739 ± .0373**	0.7630 ± .0417	0.7503 ± .0454	0.7675 ± .0377
	NB	**0.9412 ± .0243**	0.9127 ± .0321	0.9385 ± .0297	0.9359 ± .0292	0.9368 ± .0298	0.6579 ± .0722	0.9354 ± .0293
	SVM	0.7467 ± .0378	0.6836 ± .0438	**0.7700 ± .0462**	0.7386 ± .0478	0.7271 ± .0472	0.7417 ± .0444	0.7370 ± .0491
	RF	**0.9433 ± .0318**	0.8530 ± .0445	0.8987 ± .0394	0.9074 ± .0391	0.8941 ± .0392	0.8912 ± .0385	0.9125 ± .0348
	GBDT	**0.9837 ± .0172**	0.9582 ± .0281	0.9809 ± .0181	0.9773 ± .0261	0.9809 ± .0170	0.9662 ± .0264	0.9789 ± .0196
MA-F	LR	0.7321 ± .0344	0.6903 ± .0330	**0.7490 ± .0389**	0.7410 ± .0356	0.7340 ± .0405	0.7030 ± .0449	0.7329 ± .0344
	NB	**0.9437 ± .0217**	0.8987 ± .0341	0.9427 ± .0275	0.9406 ± .0270	0.9414 ± .0277	0.5570 ± .0960	0.9400 ± .0273
	SVM	0.7271 ± .0380	0.6629 ± .0443	**0.7386 ± .0456**	0.7174 ± .0475	0.7054 ± .0457	0.7069 ± .0447	0.7149 ± .0475
	RF	**0.9511 ± .0278**	0.7972 ± .0510	0.9111 ± .0357	0.9122 ± .0342	0.9001 ± .0376	0.8883 ± .0381	0.9199 ± .0313
	GBDT	0.9860 ± .0152	0.9287 ± .0448	0.9819 ± .0170	0.9791 ± .0227	0.9820 ± .0172	0.9640 ± .0290	0.9803 ± .0185
MA-G	LR	0.8618 ± .0200	0.8488 ± .0234	**0.8781 ± .0238**	0.8687 ± .0223	0.8621 ± .0252	0.8542 ± .0275	0.8648 ± .0226
	NB	**0.9680 ± .0133**	0.9507 ± .0178	0.9661 ± .0164	0.9646 ± .0162	0.9651 ± .0165	0.7873 ± .0518	0.9644 ± .0162
	SVM	0.8546 ± .0230	0.8112 ± .0276	**0.8665 ± .0272**	0.8477 ± .0284	0.8408 ± .0283	0.8496 ± .0268	0.8468 ± .0292
	RF	**0.9682 ± .0180**	0.9125 ± .0267	0.9428 ± .0226	0.9480 ± .0222	0.9401 ± .0227	0.9397 ± .0215	0.9505 ± .0197
	GBDT	**0.9912 ± .0093**	0.9741 ± .0171	0.9898 ± .0097	0.9879 ± .0141	0.9897 ± .0092	0.9816 ± .0146	0.9887 ± .0105
A-cost	LR	**0.4863 ± .0853**	0.5833 ± .1283	0.5081 ± .1071	0.5231 ± .1055	0.5619 ± .1246	0.5257 ± .1050	0.5253 ± .1084
	NB	**0.0908 ± .0399**	0.1875 ± .0556	0.1414 ± .0664	0.1450 ± .0652	0.1436 ± .0653	0.7206 ± .2071	0.1456 ± .0651
	SVM	**0.4566 ± .0788**	0.6378 ± .1215	0.4596 ± .0914	0.5028 ± .0937	0.5322 ± .0994	0.4939 ± .0838	0.4993 ± .1031
	RF	**0.1746 ± .1139**	0.4021 ± .1573	0.3121 ± .1292	0.2511 ± .1270	0.3171 ± .1475	0.2075 ± .0867	0.2749 ± .1169
	GBDT	**0.0266 ± .0337**	0.1284 ± .0918	0.0451 ± .0304	0.0350 ± .0254	0.0465 ± .0322	0.0703 ± .0449	0.0467 ± .0304

(1) Boldface indicates the best performance on each row. (2) The figure following "±" is the standard deviation across 50 experiments

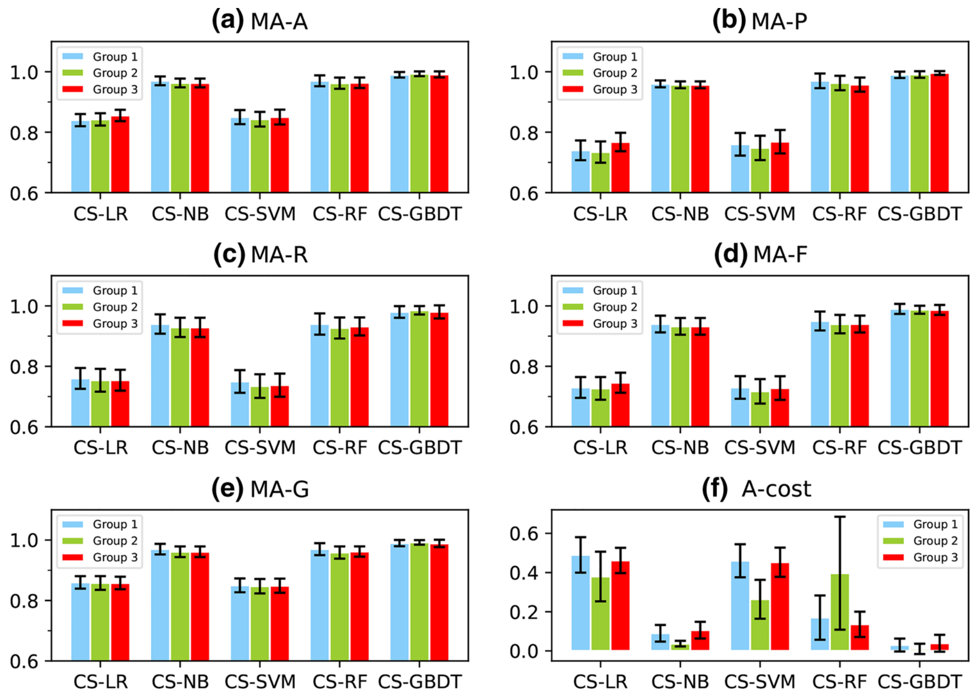

Fig. 3 The comparison of cost-sensitive classification performance on different cost matrices

Sensitivity analysis of cost-sensitive learning with different cost matrices

In this subsection, we do sensitivity analysis of our cost-sensitive learning method in fault diagnosis. To this end, we consider three groups of cost matrices:

Group 1: $C_{i0} = 10w$, $C_{0j} = w$, and $C_{ij} = 3w$ for $i \neq 0$, $j \neq 0$, and $i \neq j$.

Group 2: $C_{i0} = 15w$, $C_{0j} = w$, and $C_{ij} = 2w$ for $i \neq 0$, $j \neq 0$, and $i \neq j$.

Group 3: $C_{i0} = 7w$, $C_{0j} = w$, and $C_{ij} = 4w$ for $i \neq 0$, $j \neq 0$, and $i \neq j$.

The details of group 1 have been covered in “Comparison with classical cost-insensitive learning methods” section and serves as a benchmark. In group 2, we increase the cost of missed alarm and reduce the cost of misjudging one fault for another one. In contrast, the cost of missed alarm in group 3 is reduced, whereas the cost of misjudging one fault as another one is increased. Taking the same procedure as “Comparison with classical cost-insensitive learning methods” section, we show the experimental results for three groups in Fig. 3, where the error bars represent the standard deviations.

From the results in Fig. 3, an intuitive conclusion is that different misclassification costs will lead to different classification performances of each model. First, by comparing the sensitivity of classification performance in different groups, there are certain fluctuations in CS-GBDT and CS-NB. Conversely, CS-SVM fluctuates significantly with the change of misclassification costs. Second, by comparing the sensitivity of classification performance in the same group, the standard deviation of CS-RF classification performance in 50 independent experiments is greater than those of CS-GBDT and CS-NB.

Apart from the above two points, we also notice that the higher evaluation scores are not necessarily accompanied by lower misclassification costs. For example, in terms of evaluation scores, CS-NB, CS-RF and CS-GBDT are superior to CS-LR and CS-SVM for all cost settings, while the average misclassification cost of CS-RF is greater than that of CS-SVM in group 2. This result validates the necessity of introducing A-cost metric, which is also one of the distinctive contributions in this paper. Fortunately, CS-GBDT can achieve better performance and more robust results than the other four models in terms of higher evaluation scores and lower misclassification costs across different cost matrices. Thus, we recommend it as a suitable method for practical use.

Conclusions

The purpose of this paper is to develop a novel framework to diagnose the imbalanced operation condition of rotating machinery through combined use of multi-domain feature extraction, feature selection and cost-sensitive learning method. We first extract 65 features from the perspectives of time-domain, frequency-domain and time-frequency-domain. The multi-domain features can comprehensively reflect the operational status of a rotating machinery. Second, the mRMR feature selection technique is used to reduce the entire feature set to a more compact one. Third, we design the cost-sensitive learning method to improve the performance of fault diagnosis with imbalanced data. This is done by imposing different misclassification costs on false alarm and missed alarm respectively.

Our framework is evaluated on the acquired vibration data of rotating machinery from an oil refinery in Zibo, Shandong Province, China. From extensive experimental comparisons, the results illustrate that our multi-domain feature extraction is valuable and the extracted features have the ability to achieve a higher classification performance than previous works. By comparison with traditional cost-insensitive methods as well as other imbalanced learning methods, our cost-sensitive learning method performs better in imbalanced fault classification. Meanwhile, a sensitivity experiment demonstrates that different misclassification costs will lead to different classification performance. However, it is worth mentioning that CS-GBDT is more robust and is preferred for its high evaluation scores and low average misclassification cost.

In the future, this research can be further promoted in two directions. First, we can consider multi-source data fusion in fault diagnosis. Owing to the limitation of the data acquired in this paper, the features are only extracted from the vibration signals, while the effects of other signals, such as electrical signals and temperature, on the diagnostic results of rotating machinery are ignored. With the new paradigm of Industry 4.0, more and more multimode sensors make it possible to fuse multi-source data and conduct more comprehensive fault diagnosis. Second, we could consider the severity of each fault condition. We have roughly set the cost matrix for different fault conditions without considering the severity of each fault condition. It is very useful to divide different levels of severity and specify the cost accordingly. In this regard, the cost-sensitive learning method can play a greater role in fault diagnose and produce more accurate results for scientific decision-making. We leave this topic for future research.

Acknowledgements The authors are grateful to the Editor-in-Chief, the Associate Editor, and two anonymous referees for their helpful comments and constructive guidance. This work is supported by the National Natural Science Foundation of China (71671056), the Humanity and Social Science Foundation of the Ministry of Education of China (19YJA790035), and the National Statistical Science Research Projects of China (2019LD05). Special thanks to data support from the industrial partner RONDS.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Amrhein, W., Gruber, W., Bauer, W., & Reisinger, M. (2016). Magnetic levitation systems for cost-sensitive applications-some design aspects. *IEEE Transactions on Industry Applications*, *52*(5), 3739–3752.

Ben Ali, J., Saidi, L., Harrath, S., Bechhoefer, E., & Benbouzid, M. (2018). Online automatic diagnosis of wind turbine bearings progressive degradations under real experimental conditions based on unsupervised machine learning. *Applied Acoustics*, *132*, 167–181.

Beygelzimer, A., Dani, V., Hayes, T., Langford, J., & Zadrozny, B. (2005). Error limiting reductions between classification tasks. In *Proceedings of the 22nd international conference on machine learning* (pp. 49–56).

Castro, C. L., & Braga, A. P. (2013). Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, *24*(6), 888–899.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Ciabattoni, L., Ferracuti, F., Freddi, A., & Monteriú, A. (2018). Statistical spectral analysis for fault diagnosis of rotating machines. *IEEE Transactions on Industrial Electronics*, *65*(5), 4301–4310.

Correa Bahnsen, A., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, *42*(19), 6609–6619.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*(3), 253–263.

Ding, C., & Peng, H. (2005). Minmum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, *3*(2), 185–205.

Domingos, P. (1999). MetaCost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 155–164).

Dou, R., He, Z., & Hsu, C. (2018). Foreword: Smart manufacturing, innovative product and service design to empower industry 4.0. *Computers & Industrial Engineering*, *125*, 514–516.

Gan, M., Wang, C., & Zhu, C. (2018). Fault feature enhancement for rotating machinery based on quality factor analysis and manifold learning. *Journal of Intelligent Manufacturing*, *29*(2), 463–480.

Gardner, J., & Xiong, L. (2009). An integrated framework for de-identifying unstructured medical data. *Data & Knowledge Engineering*, *68*(12), 1441–1451.

Georgoulas, G., Loutas, T., Stylios, C. D., & Kostopoulos, V. (2013). Bearing fault detection based on hybrid ensemble detector and empirical mode decomposition. *Mechanical Systems and Signal Processing*, *41*(1–2), 510–525.

Haibo, H., Yang, B., Garcia, E. A., & Shutao, L. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 155–164).

Han, H., Wang, W., & Mao, B. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Proceedings of advances in intelligent computing* (pp. 878–887).

Han, S., Choi, H., Choi, S., & Oh, J. (2019a). Fault diagnosis of planetary gear carrier packs: A class imbalance and multiclass classification problem. *International Journal of Precision Engineering and Manufacturing*, *20*(2), 167–179.

Han, T., Liu, C., Yang, W., & Jiang, D. (2019b). Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application. *ISA Transactions,* In press.

Hwang, Y., Jen, K., & Shen, Y. (2009). Application of cepstrum and neural network to bearing fault detection. *Journal of Mechanical Science and Technology*, *23*(10), 2730–2737.

Jia, F., Lei, Y., Lu, N., & Xing, S. (2018). Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization. *Mechanical Systems and Signal Processing*, *110*, 349–367.

Jiang, G., He, H., Yan, J., & Xie, P. (2019). Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox. *IEEE Transactions on Industrial Electronics*, *66*(4), 3196–3207.

Jiang, Q., Shen, Y., Li, H., & Xu, F. (2018). New fault recognition method for rotary machinery based on information entropy and a probabilistic neural network. *Sensors*, *18*(2), 337–349.

Jiang, W., Spurgeon, S. K., Twiddle, J. A., Schlindwein, F. S., Feng, Y., & Thanagasundram, S. (2016). A wavelet cluster-based band-pass filtering and envelope demodulation approach with application to fault diagnosis in a dry vacuum pump. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *221*(11), 1279–1286.

Kang, S. (2018). Joint modeling of classification and regression for improving faulty wafer detection in semiconductor manufacturing. *Journal of Intelligent Manufacturing*,. https://doi.org/10.1007/s10845-018-1447-2.

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., & Togneri, R. (2018). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(8), 3573–3587.

Kuo, R. J., Su, P. Y., Zulvia, Ferani E., & Lin, C. C. (2018). Integrating cluster analysis with granular computing for imbalanced data classification problem—a case study on prostate cancer prognosis. *Computers & Industrial Engineering*, *125*, 319–332.

Larsson, E. G., Stoica, P., & Jian, L. (2002). Amplitude spectrum estimation for two-dimensional gapped data. *IEEE Transactions on Signal Processing*, *50*(6), 1343–1354.

Lee, Y., Hu, P. J., Cheng, T., & Hsieh, Y. (2012). A cost-sensitive technique for positive-example learning supporting content-based product recommendations in B-to-C e-commerce. *Decision Support Systems*, *53*(1), 245–256.

Li, P., Hu, W., Hu, R., & Chen, Z. (2020). Imbalance fault detection based on the integrated analysis strategy for variable-speed wind turbines. *International Journal of Electrical Power & Energy Systems,116*, In press.

Liu, J., An, Y., Dou, R., Ji, H., & Liu, Y. (2018a). Helical fault diagnosis model based on data-driven incremental mergence. *Computers & Industrial Engineering*, *125*, 517–532.

Liu, R., Yang, B., Zio, E., & Chen, X. (2018b). Artificial intelligence for fault diagnosis of rotating machinery: A review. *Mechanical Systems and Signal Processing*, *108*, 33–47.

Mathew, J., Pang, C. K., Luo, M., & Leong, W. H. (2018). Classification of imbalanced data by oversampling in kernel space of support vector machines. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(9), 4065–4076.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238.

Ragab, A., Yacout, S., Ouali, M., & Osman, H. (2019). Prognostics of multiple failure modes in rotating machinery using a pattern-based classifier and cumulative incidence functions. *Journal of Intelligent Manufacturing*, *30*(1), 255–274.

Ren, L., Sun, Y., Cui, J., & Zhang, L. (2018). Bearing remaining useful life prediction based on deep autoencoder and deep neural networks. *Journal of Manufacturing Systems*, *48*, 71–77.

Sánchez, R., Lucero, P., Vásquez, R. E., Cerrada, M., Macancela, J., & Cabrera, D. (2018). Feature ranking for multi-fault diagnosis of rotating machinery by using random forest and KNN. *Journal of Intelligent & Fuzzy Systems*, *34*(6), 3463–3473.

Santos, P., Maudes, J., & Bustillo, A. (2015). Identifying maximum imbalance in datasets for fault diagnosis of gearboxes. *Journal of Intelligent Manufacturing*, *29*(2), 333–351.

Seera, M., Lim, C. P., & Loo, C. K. (2014). Motor fault detection and diagnosis using a hybrid FMM-CART model with online learning. *Journal of Intelligent Manufacturing*, *27*(6), 1273–1285.

Song, L., Wang, H., & Chen, P. (2018). Vibration-based intelligent fault diagnosis for roller bearings in low-speed rotating machinery. *IEEE Transactions on Instrumentation and Measurement*, *67*(8), 1887–1899.

Sun, Y., Kamel, M. S., Wong, A. K. C., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, *40*(12), 3358–3378.

Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. *Journal of Manufacturing Systems*, *48*, 157–169.

Tidriri, K., Chatti, N., Verron, S., & Tiplica, T. (2016). Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges. *Annual Reviews in Control*, *42*, 63–81.

Wang, P., Ananya, Yan, R., & Gao, R. X. (2017). Virtualization and deep recognition for system fault classification. *Journal of Manufacturing Systems,44*, 310–316.

Wang, X., Zhang, X., Li, Z., & Wu, J. (2019). Ensemble extreme learning machines for compound-fault diagnosis of rotating machinery. *Knowledge-Based Systems,* In press.

Wu, C., Jiang, P., Ding, C., Feng, F., & Chen, T. (2019a). Intelligent fault diagnosis of rotating machinery based on one-dimensional convolutional neural network. *Computers in Industry*, *108*, 53–61.

Wu, J., Wu, C., Cao, S., Or, S. W., Deng, C., & Shao, X. (2019b). Degradation data-driven time-to-failure prognostics approach for rolling element bearings in electrical machines. *IEEE Transactions on Industrial Electronics*, *66*(1), 529–539.

Xie, Y., Peng, L., Chen, Z., Yang, B., Zhang, H., & Zhang, H. (2019). Generative learning for imbalanced data using the gaussian mixed model. *Applied Soft Computing*, *79*, 439–451.

Zadrozny, B. Langford, J., & Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. In *Proceedings—IEEE international conference on data mining* (pp. 435–442).

Zan, T., Liu, Z., Wang, H., Wang, M., & Gao, X. (2019). Control chart pattern recognition using the convolutional neural network. *Journal of Intelligent Manufacturing,* In press.

Zhang, X., & Hu, B. (2014). A new strategy of cost-free learning in the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, *26*(12), 2872–2885.

Zhang, Y., Li, X., Gao, L., Wang, L., & Wen, L. (2018). Imbalanced data fault diagnosis of rotating machinery using synthetic oversampling and feature learning. *Journal of Manufacturing Systems*, *48*, 34–50.

Zhang, C., Tan, K. C., Li, H., & Hong, G. S. (2019). A cost-sensitive deep belief network for imbalanced classification. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(1), 109–122.

Zhang, Z., Verma, A., & Kusiak, A. (2012). Fault analysis and condition monitoring of the wind turbine gearbox. *IEEE Transactions on Energy Conversion*, *27*(2), 526–535.

Zhao, M., Jiao, J., & Lin, J. (2019). A data-driven monitoring scheme for rotating machinery via self-comparison approach. *IEEE Transactions on Industrial Informatics*, *15*(4), 2435–2445.

Zhao, M., & Lin, J. (2018). Health assessment of rotating machinery using a rotary encoder. *IEEE Transactions on Industrial Electronics*, *65*(3), 2548–2556.

Zhou, Z., & Liu, X. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, *18*(1), 63–77.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.