A Novel Vehicle Gearbox Fault Diagnosis Approach Based on Collective Anomaly Detection

Zhongfeng Hu[1*], Xiaodi Huang[1]

[1] School of Economic and Management, Hefei University, Hefei, 230601, China

Abstract. Targeting the problem of gearbox fault diagnosis, we proposed a novel semi-supervised approach based on collective anomaly detection. Based on the limited sample data, the principle of the approach is to detect whether a test dataset contains abnormal patterns by using data distribution as the metric. The sequence obeying unexpected distribution will be identified as collective anomaly, which may be generated by fault patterns. The approach consists of three steps. First, the mixture of multivariate Gaussian distribution is used to fit the structure of sample dataset and test dataset. Then, based on maximum likelihood estimate algorithm, we hope to search the optimal parameters which can fit the data distribution with the highest degree. Finally, the fixed point iteration algorithm is used to solve likelihood estimate functions. Experimental results demonstrate that the proposed approach can be used to find fault patterns of gearbox without the prior knowledge of their generated mechanisms.

1 Introduction

With the development of artificial intelligence technology, the research direction of fault diagnosis has changed to build an intelligent diagnosis system based on data driven and intelligent computing technologies. Most researches on gearbox fault diagnosis are based on the analysis of vibration signals. The data characteristics of vibration signal can be divided into two categories [1]: time domain and frequency domain, and the features in each of which are complex, changeable, and interactive. And they can be easily influenced by other vibration sources during the driving operation of the vehicle [2].

Currently, most research on fault diagnosis focus on semi-supervised anomaly detection approach [3], which detects unknown fault patterns according to a limited size of sample data. One of the common limitations of these existing approach is that they can hardly detect anomalies which are generated in the same background as normal data [4]. That is, how to accurately and sensitively detect the real time failures during the continuous operation of the gearbox. Although the data that makes up these real-time faults seems normal on their own, it is not normal for them to appear together as a set. If the detection only checks the single data of unmarked test dataset one by one, it is difficult to find such abnormal patterns. On the contrary, it may mistakenly identify some normal data falling in the low probability density range as abnormal.

To solve the problems mentioned above, we propose a semi-supervised collective anomaly detection approach based on data distribution similarity metric and apply it in the fault diagnosis of vehicle gearbox. This algorithm consists of three parts: 1) mixture of multivariate Gaussian distributions is used to fit the distributions of sample dataset and test dataset. 2) Based on MLE (maximum likelihood estimate) algorithm, search the optimal parameters which can fit the data distribution. 3) The fixed point iteration algorithm is used to solve likelihood estimate functions. When the distribution of a data pattern is significantly different from the sample data, it can be identified as a collective anomaly that may be generated by gearbox fault.

This paper is organized as follows. In section 2, we introduce the construction of the semi-supervised detection model in detail, and give the definition of the mathematical theory involved in it. Then, based on multivariate distribution fitting, maximum likelihood estimation and fixed point iteration, the solution process of the model is shown in detail in section 3. In section 4, we verify our detection approach by testing on actual operation data of gearbox in an automobile factory. Lastly, we summarize the research results in section 5.

2 Detection framework and related concept

2.1 Detection Framework

The framework of the proposed detection approach are conducted in two processes. Firstly, for a labeled normal sample dataset S_s, its data distribution can be denoted as *Eq.1*. The parameter f_s represents the data distribution function, θ_s represents parameters of the function.

$$D(S_s) = f_s(x|\theta_s) \quad (1)$$

Secondly, for an unlabeled test dataset S_t which may

*Corresponding author's e-mail: huzf@hfuu.edu.cn

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

including abnormal pattern S_a, the mixed multivariate Gaussian distribution function is used to fit the data structure as *Eq.2*. The parameter φ represents the portion of anomaly pattern, F represents the mixture distribution of the test dataset.

$$D(S_t) = F(x|\theta_t) = F[(1-\varphi)f_s(x;\theta_s); \varphi f_a(x;\theta_a)] \quad (2)$$

Based on our proposed collective anomaly detection approach, we hope to examine whether an unlabeled test contains fault patterns by similarity measures based on data distribution. For the realization of the goal, there are three parameters as θ_s, θ_a, and φ need to be estimated.

2.2 Collective Anomaly

The collective anomaly is a set of related data instances. When they appear together in a certain pattern, their overall behavior attribute will deviate significantly from the whole dataset, but the individual observation in the set may not be an anomaly.

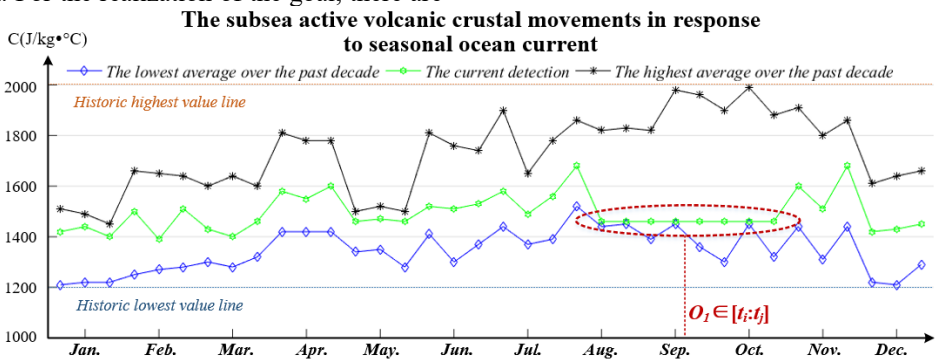

Figure 1. Collective anomaly O_1 in a Volcano supervision dataset

In Figure1, it shows a dataset of collecting the response of submarine active volcano movement to seasonal ocean currents which reflected by the metric of specific heat capacity (C, J/kg•°C). The yellow dotted line is the all-time high line, and the blue dotted line is the all-time low line. The yellow and blue solid line represent the highest and lowest averages in the past decade respectively, and the purple solid line represents the current situation. The data that constitute the subsequence O_1 denotes a collective anomaly, since the variation trend of the subsequence is obviously different from the historical curve during the time interval, even if each observation itself fluctuates within the normal range.

2.3 Fixed Point Iteration Algorithm

Fixed-point iteration is a successive approximation method with which represent the implicit equation by a set of explicit equations. In other word, the approximate value of the root is repeatedly corrected using an equation to make it convergence [5]. As an effective method for solving highly nonlinear numerical problems [6], due to its excellent mathematical properties and mature theorem proofs, fixed point iteration has been widely used for searching equation solution in many fields of engineering mathematics. Main concepts of the algorithm are shown as follows:

Definition 1. Suppose that X is a subset of R^n. If there is a specific $f(x) \in X$ corresponding to every point x in subset X, f is a self-mapping of X, denoted as f: $X \to X$.

Definition 2. Suppose that X is a nonempty set and f: $X \to X$ is its self-mapping. If there is a $x^* \in X$ satisfy $f(x^*) = x^*$, x^* is considered to be a precise fixed point of f.

Definition 3. Suppose that (X, ρ) is a metric space and T: $X \to X$ is a mapping. If there is an $L \in [0, 1)$ that enables $\rho(T(x), T(y)) \leq L\rho(x, y)$ for any $x, y \in X$, T is considered to be the contraction mapping on X.

Theorem 1. The *Banach* fixed-point theorem is also known as the contraction mapping theorem. Suppose that (X, ρ) is a nonempty perfect metric space and T: $X \to X$ is a contraction mapping, T has the only fixed point in X. The Banach fixed-point theorem determines the existence and uniqueness of the solution to equation $T(x) = x$.

Theorem 2. For any contraction mapping T: $X \to X$, suppose that X is a bounded discrete nonempty set, which means there is $a \leq x \leq b$ for any $x \in X$. If the following two conditions are satisfied: (1) there is $a \leq T(x) \leq b$ for any $x \in X$ and (2) there is a positive constant $L < 1$ that enables $|T(x) - T(y)| \ll L|x - y|$ for any x, $y \in X$, T has the only fixed point x^* within the bounded discrete nonempty set.

Definition 4. Approximate fixed point: suppose that ε is any positive constant and $|x - f(x)|$is the modulus of the vector $x - f(x)$ in n-dimensional Euclidean space Rn for the contraction mapping T: $X \to X$. If there is a point x^* satisfying$|x^* - f(x^*)| < \varepsilon$, x^* is an approximate fixed point.

The existence of a precise fixed point can be proven in many conditions, but the computation overhead is always too expensive to find it, besides, for the convenience of calculation, the precise value of fixed point is usually need to be approximated. Just like, the precise solution to $x^2 - 2 = 0$ is infinite which must be approximated to participate in the later calculation. Therefore, we introduced the concept of approximate fixed-point into our algorithm to solve this kind of problem. If the limited numerical value of precise fixed point was not found when the search reached the preset number of iterations, the approximate fixed point with the highest precision during the iteration will be taken as the result.

3 Algorithm construction

Our proposed detection approach consists of three part. Firstly, finite mixtures of multivariate Gaussian distributions are used to represent the distribution of labeled normal sample dataset (as shown in Eq.1) and unlabeled test dataset (as shown in *Eq.2*). Then, the MLE (maximum likelihood estimate) algorithm to estimate the parameters of the mixture distribution functions. Finally, fixed point iteration algorithm is carried out to solve the maximum likelihood estimate functions.

3.1 Mixture of Multivariate Gaussian Distributions

The multivariate mixture of Gaussian is adopted to represent the data distribution. In an *n*-dimensional Euclidean space R^n, the mixture multivariate Gaussian distributions of *K* components is defined as *Eq.3* and *Eq.4*. $N(x|\mu_k;\sum_k)$ represents the probability density function for the k Gaussian distribution with mean μ_k . The parameter $\sum$ is covariance matrix, which is symmetric and positive semi-definite, and the $|\sum_k|$ denotes matrix determinant. The λ_k is the mixing coefficient for the k Gaussian distribution, which satisfy $\lambda_k \geq 0$ and $\sum_{k=1}^{K}\lambda_k = 1$.

$$N(x|\mu_k;\sum_k) = \frac{1}{(2\pi)^{n/2}|\sum_k|^{1/2}} exp\left(-\frac{1}{2}(x-\mu_k)^T\sum_k^{-1}(x-\mu_k)\right) \quad (3)$$

$$D(S) = f(x|\theta) = \sum_{k=1}^{K}\lambda_k N(x|\mu_k;\sum_k) \quad (4)$$

For the labeled normal sample dataset, the data distribution function is defined as *Eq.5*. The parameter needs to be estimated is $\theta_s = (\lambda_j;\mu_j;\sum_j)$.

$$D(S_s) = f_s(x|\theta_s) = \sum_{j=1}^{J}\lambda_j N(x|\mu_j;\sum_j) \quad (5)$$

For the unlabeled test dataset, the data distribution function is defined as *Eq.6*. The parameter needs to be estimated is $\theta_t = (\theta_s;\varphi;\mu_q;\sum_q)$

$$D(S_t) = F(x|\theta_t) = (1-\varphi)f_s(x;\theta_s) + \varphi\sum_{q=j+1}^{J+Q}\lambda_q N(x|\mu_q;\sum_q) \quad (6)$$

3.2 Maximum Likelihood Estimate (MLE) algorithm

Maximum likelihood estimation algorithm is used in the case where the data distribution function is known but the function parameters are known. For a continuous sequence *S*, its probability density function is $f(x|\theta)$. If S_1=(X_1, X_2… X_n) is a sample of *S*, and the probability density function $p(X|\theta) = \prod_{i=1}^{n} f(x_i;\theta)$ is known. If the point $Y = (x_1, x_2, \cdots, x_n) \in S_1$, the probability that any random point in *S* falls on the adjacent side of *Y* can be approximately expressed as $p = \prod_{i=1}^{n} f(x_i;\theta)\,dx_i$. The likelihood function of sequence *S* can be calculated as *Eq.7*.

$$L(\theta) = L(x_1, x_2, \cdots, x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta) \quad (7)$$

The method of MLE algorithm is to find the parameter $\hat{\theta}$ which can make the probability $p = \prod_{i=1}^{n} f(x_i;\theta)\,dx_i$ reach maximum value, it can be defined as *Eq.8*.

$$L(x_1, x_2, \cdots, x_n;\hat{\theta}) = maxL(x_1, x_2, \cdots, x_n;\theta) \quad (8)$$

where $\hat{\theta}$ is associated with the selected point *Y*, the $\hat{\theta}(x_1, x_2, \cdots, x_n)$ represents the maximum likelihood estimation of the parameter θ of the probability density function $f(x|\theta)$.

Due to the function $lnL(\theta)$ is the increasing function of $L(\theta)$, both of them will reach their maximum value at the same point. Hence, it is usually to search the extreme point of $lnL(\theta)$ to replace $L(\theta)$, which can not only convert the multiplication to addition but also avoid the problem of floating point overflow. Thus, based on the *Eq.4 and Eq.6*, the likelihood function of labeled normal sample dataset can be calculated as *Eq.9*, and the likelihood function of unlabeled test dataset can be calculated as *Eq.10*.

$$lnL(x_1, x_2, \cdots, x_n;\hat{\theta}_s) = maxlnf_s(x|\theta_s) \quad (9)$$

$$lnL(x_1, x_2, \cdots, x_n;\hat{\theta}_t) = maxlnf_t(x|\theta_t) \quad (10)$$

3.3 Fixed point iteration algorithm

For *Eq.9* and *Eq.10*, the extreme value will appear at the inflection point of the function, it means if a point $Y_i = (y_{i1}, y_{i2}, \cdots, y_{in}) \in S$ satisfy the equation $lnL(\theta)' = 0$, the $\hat{\theta}(y_{i1}, y_{i2}, \cdots, y_{in})$ will represents the maximum likelihood estimation of the parameter θ. Based on the definition of fixed point iteration in former section, it can be used to search the maximum likelihood estimation of the parameter θ of the probability density function $f(x|\theta)$, the detailed steps are as follows.

(1) Construct the fixed-point iteration. The problem of searching function extreme value can be converted to seek the point that satisfying derivative $lnL(\theta)' = 0$.

(2) Transform the functional form of $lnL(\theta)' = 0$ into $\theta = f(\theta)$.

(3) Select an initial approximate solution $\theta_0(y_{01}, y_{02}, \cdots, y_{0n})$ and substitute it into the right side of $\theta = f(\theta)$, yielding $\theta_1 = f(\theta_0)$. This step was repeated according to the equations $\theta_k = f(\theta_{k-1})$, k is the size of the sequence.

(4) Before the process in step 3 reaching the max number of iterations, if there is a solution $\hat{\boldsymbol{\theta}}(\boldsymbol{y_{i1}}, \boldsymbol{y_{i2}}, \cdots, \boldsymbol{y_{in}})$ satisfying $\hat{\boldsymbol{\theta}} = \boldsymbol{f}(\hat{\boldsymbol{\theta}})$, it will be treated as the precise fixed point. According to the Definition 4, if the precise fixed point cannot be found, the point that satisfied $|\hat{\boldsymbol{\theta}} - \boldsymbol{f}(\hat{\boldsymbol{\theta}})| < \boldsymbol{\varepsilon}$ at the greatest extent will be taken as the approximate fixed point.

4 Experiment and analysis

The WLY·CVT25 stepless gearbox newly developed by an automobile manufacturer is selected as the experimental object, detailed product information is shown in Figure 2 (a). The experimental data is derived from the vibration signal collected by sensors under different working conditions of the gearbox, and the acquisition frequency is once every 5 seconds.

Product features

Rated torque: 250Nm
Dry weight: 84.5Kg
Axial length: 365mm
Center distance
(1-4 bearing): 197mm
Velocity ratio: 7.07

Figure 2(a). The detail information of WLY·CVT25

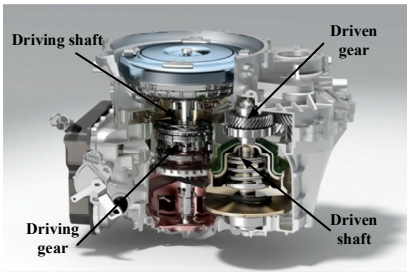

Figure 2(b). The test component of WLY·CVT25

4.1 Experimental dataset

The experimental dataset used in this section consists of three parts: normal sample dataset S_s, abnormal dataset S_a, and unknown test dataset S_t.

Normal sample dataset S_s: To avoid the data fluctuation caused by too long continuous working of one gearbox, three qualified gearboxes of the same model are selected to work continuously for 24 hours under the same load condition. The vibration datasets generated by the three gearboxes will be set as the normal dataset S_{normal}. The data collected within 10 independent hours are randomly selected from the normal dataset S_{normal} to form the normal sample dataset S_s.

Abnormal dataset S_a: As shown in Figure 2 (b), the driving shaft, driven shaft, driving gear, and driven gear are four most important components of the gearbox. Thus, we select them as the target of fault diagnosis. In the experiment, these four qualified parts will be replaced with cracked parts one by one. For each cracked part, under the same load condition, we collect 6 hours of vibration data as the abnormal dataset S_a.

In addition, compared with the cracked parts, the difference between the vibration signal generated by the worn old parts and the normal signal is not so obvious. In order to test the sensitivity of our algorithm, we also use the worn old parts to replace the qualified parts one by one, and collect the vibration data for 6 hours under the same load condition. To clearly represent these abnormal datasets, we label them according to Table 1 to avoid unnecessary troubles.

Table 1. Symbolic representation of all kinds of abnormal datasets

	Power take off shaft	Driven shaft	Driving gear	Driven gear
Crack	$S_{a1}(C)$	$S_{a2}(C)$	$S_{a3}(C)$	$S_{a4}(C)$
Abrasion	$S_{a1}(A)$	$S_{a2}(A)$	$S_{a3}(A)$	$S_{a4}(A)$

Unknown test dataset S_t: In the normal dataset, different kinds of abnormal dataset S_a are added one by one to form eight kinds of unknown test datasets S_t. Based on the previous assumption that the fault pattern only accounts for a small proportion of the entire dataset, the proportion of abnormal dataset is controlled below 5% of the normal dataset. The data of no more than 3 hours size are randomly select from all kinds of abnormal datasets to add to the normal dataset. Details are shown in Table 2.

Table 2. The results of fault diagnosis of all test datasets

	Sample dataset S_s	Normal dataset S_{normal}	Anomaly dataset S_a
$S_{t1}= S_{normal}+ S_{a1}(C)$	10h	72h	2.6h
$S_{t2}= S_{normal}+ S_{a2}(C)$	10h	72h	2.1h
$S_{t3}= S_{normal}+ S_{a3}(C)$	10h	72h	0.7h
$S_{t4}= S_{normal}+ S_{a4}(C)$	10h	72h	1.5h
$S_{t5}= S_{normal}+ S_{a1}(A)$	10h	72h	1.9h
$S_{t1}= S_{normal}+ S_{a2}(A)$	10h	72h	1.3h
$S_{t6}= S_{normal}+ S_{a3}(A)$	10h	72h	2.6h
$S_{t7}= S_{normal}+ S_{a4}(A)$	10h	72h	0.5h

4.2 Experimental results

For the sample dataset S_s and abnormal datasets S_a, data distributions are fitted based on our proposed algorithm. The parameters of the distribution function of the sample dataset will continue to participate in the subsequent analysis. However, the proportion and distribution function parameters of various abnormal datasets will be used as the real labels to compare with the detection results, rather than directly participate in the analysis on unknown test datasets S_t.

The parameters of the probability density function of sample dataset is $\boldsymbol{\theta_s} = (\boldsymbol{\mu = 0.0126; \sigma = 36.432})$. The detection results S_{t1}~S_{t8} are shown in Table 3, the proportion of collective anomaly detected by our proposed algorithm and the parameters of its probability density function are compared with its real labels. All the results were calculated to three decimal places.

Table 3. The results of fault diagnosis of all test datasets

		S_{t1}	S_{t2}	S_{t3}	S_{t4}	S_{t5}	S_{t6}	S_{t7}	S_{t8}
Detection result	Mean(μ)	0.1342	0.1027	0.0816	0.0685	0.0451	0.0268	0.0232	0.0198
	SD(σ)	233.839	174.385	101.519	86.392	82.391	58.746	66.288	43.727
	Proportion(λ)	3.168%	2.607%	1.024%	1.867%	2.435%	1.668%	3.271%	0.746%
Real label	Mean(μ)	0.1475	0.1099	0.0842	0.0742	0.0485	0.0295	0.0257	0.0206
	SD(σ)	256.966	183.950	106.080	93.201	88.492	60.742	71.847	43.397
	Proportion(λ)	3.485%	2.834%	0.963%	2.041%	2.571%	1.774%	3.485%	0.690%

From the experimental results in Table 3, it can be found that our proposed algorithm has reached more than 90% agreement in all detection indexes when detecting the fault pattern of each unknown test dataset, especially when identifying worn and old parts, it still shows high sensitivity, which can prove the effectiveness of our algorithm. In addition, based on the detection results, we can also draw the following conclusions: 1) the obvious degree of bearing fault is greater than that of gear. 2) The obvious degree of driving component fault is greater than that of driven component. 3) The obvious degree of crack component fault is greater than that of worn old component.

5 Conclusions

In this paper, we have presented a semi-supervised vehicle gearbox fault diagnosis approach based on collective anomaly detection. In the proposed algorithm, firstly the mixed Gaussian distribution was used to fit the vibration signal of the gearbox. Then, the parameter variation of the probability density function of the data distribution was taken as the measurement standard. Finally, based on the known normal sample dataset, the maximum likelihood method and fixed point iteration method were used to fit the distribution of the unknown test dataset. According to the fitting results of data distribution, data patterns that are subject to unknown or unexpected distributions will be identified as collective anomalies which may be generated by faults. For creditability verification, we have made the detection experiment on eight kinds of test datasets in which including different kinds of fault patterns. The experimental results show that, when detecting each test dataset, the proposed algorithm has achieved a fit of more than 90% on each parameter of the failure data distribution function, it still shows high sensitivity on identifying worn old parts. Therefore, it verifies that our proposed detection approach can be used to find fault patterns of vehicle gearbox without the prior knowledge of their generated mechanisms. Given the generality of the framework, it should be possible to find future applications also on other fields of science and technology.

Acknowledgments

This work was supported by the Humanities and Social Science Project of Anhui Provincial Education Department under Grant NO. SK2019A0696 and Research Fund Project of Hefei University under Grant NO. 16-17RC30.

References

1. Tao, J.F., Qin, C.J., Li, W. (2019) Intelligent Fault Diagnosis of Diesel Engines via Extreme Gradient Boosting and High-Accuracy Time-Frequency Information of Vibration Signals. Sensors, 19(15):7.
2. Feng, Z.P., Zuo, M.J. (2012) Vibration signal models for fault diagnosis of planetary gearboxes. J. Sound., Vib., 331(22):4919-4939.
3. Wen, L., Li, X. (2018) A New Convolutional Neural Network-Based Data-Driven Fault Diagnosis Method. IEEE Trans. Ind. Electron., Control Instrum., 65(7):5990-5998.
4. Zhao, M.B., Tian, Z.Y., Chow, T.W.S. (2019) Fault diagnosis on wireless sensor network using the neighborhood kernel density estimation. Neural Comput. Appl., 31(8):4019-4030.
5. Ahmed, M. (2018) Collective Anomaly Detection Techniques for Network Traffic Analysis. Ann. Data Sci., 23(3):1-16.
6. Hirstoaga, S.A. (2018) Iterative selection methods for common fixed point problems. J. Math. Anal., Appl., 324(2):1020-1035.