Bearing fault diagnosis based on active learning and random forest

Jiayu Chen[1], Chen Lu[2], Hang Yuan[3]
School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China
Science and Technology on Reliability and Environmental Engineering Laboratory, Beijing 100191, China
[3]Corresponding author
E-mail: [1]*chenjiayu@buaa.edu.cn*, [2]*luchen@buaa.edu.cn*, [3]*buaayuanhang@163.com*

(Accepted 14 September 2015)

Abstract. Bearing plays an important role in rotating machineries and has received increasing attention in diagnosis of its faults accurately. This paper proposes a fault diagnosis approach exploiting active learning (AL) based on random forest (RF), which can perform accurate bearing fault diagnosis with most valuable samples. First, feature vectors are obtained by empirical mode decomposition (EMD) process for original vibration signals and selected as input of the system. Second, samples with highest uncertainty are selected through AL and added to the training set to train RF classifier. Finally, trained RF is employed to perform classification for bearing faults with testing set. Experimental results demonstrate that the proposed approach can effectively and accurately identify typical bearing faults.

Keywords: bearing fault diagnosis, active learning, random forest, uncertainty sampling.

1. Introduction

For rotary machine, bearing is one of the most important elements, which is used to support rotating components. The performance of bearing has a great influence on the whole machine, whose failure may cause machinery breakdown or even casualties. To prevent these unexpected cases, bearing has received sustained attention for many years in Prognostics and Health Management (PHM) field [1]. To prevent severe machinery failures, it is significant important to diagnose bearing faults at their early developing stages.

The information of bearing faults is diagnosed mainly by vibration analysis [2]. Vibration signals measured from bearings are complex multi-component signals. To analyze vibration signals and extract features, many methods are developed and used extensively such as short-time Fourier transform wavelet analysis [3] and empirical mode decomposition (EMD). EMD is a traditional approach to decompose nonlinear signal, which is adopted in this paper to get feature vectors. With these feature vectors, numerous methods are utilized to perform pattern classification of bearing fault such as artificial neural networks (ANNs), support vector machines (SVMs), rule-based induction, case-based reasoning, etc. However, in many pattern classification tasks, labels are often expensive or time consuming to obtain while a vast amount of unlabeled data are easily available. Simultaneously, redundant samples are often in the training set, which slows down the training process of the classifier without improving classification results. Active learning (AL) techniques are proposed to select the most valuable samples for manually labeling to train a classifier [4, 5]. AL has evolved as a key concept to reduce annotation costs and generally refers to systems where the learning algorithm receives some control over the selection of additional training data during several iterations [6, 7]. Uncertainty sampling strategy is to query the labels for samples with high uncertainty, which can be measured by the posterior probabilities [8] and marginal entropy. Therefore, the selected informative samples are manually labeled and added to the training to train the classifier. As a promising classifier in rotating machinery fault diagnosis [9], random forest (RF) is a general term for ensemble methods using tree-type classifiers, which is induced by Breiman [10]. RF builds a large number of decision trees [11, 12] out of a sub-dataset from a unique original training set by using bagging, which is a meta-algorithm to improve classification, and regression models according to stability and classification accuracy. Variance is reduced through bagging, which helps to avoid over-fitting synchronously.

This procedure extracts cases randomly from original training data set and the bootstrap sets are used to construct each of the decision trees in the RF. Each tree classifier is named a component predictor. The RF makes decisions by counting the votes of component predictors on each class and then selecting the winner class in terms of the number of votes to it [13].

Large numbers of AL algorithms are based on SVM and regression classifier, while there is little work about AL using RF classifier. DeBarr et al. have made an exploration in RF AL [14]. In this paper, an AL based on RF method is presented to select samples with high uncertainty for manually labeling and exploit the most valuable samples to train RF classifier to increase the classification accuracy. Therefore, the paper is organized as follows. Section 2 introduces the details of the proposed method, consisting of EMD, RF and AL; Experimental results are shown in Section 3; Section 4 concludes this paper.

2. Methodology

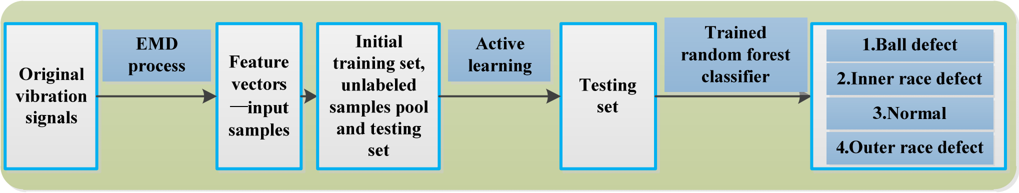
Original vibration signals
EMD process
Feature vectors —input samples
Initial training set, unlabeled samples pool and testing set
Active learning
Testing set
Trained random forest classifier
1.Ball defect
2.Inner race defect
3.Normal
4.Outer race defect

Fig. 1. Framework of the proposed approach

The framework of the proposed approach is shown as Fig. 1. To begin with, feature vectors are obtained through EMD process for original vibration as input of system, which are divided into initial training set, unlabeled samples pool and testing set. Next, AL selects the most valuable samples from unlabeled samples pool for manually labeling, which are added to the training set. Simultaneously, with updating the training set, RF classifier is trained. Finally, the fault patterns of roller bearings are identified by trained RF classifier.

2.1. EMD process

There are numerous methods to extract bearing fault features of vibration signals, such as short-time Fourier transform, wavelet analysis and Empirical Mode Decomposition (EMD). In this paper, EMD is chosen to decompose the original vibration signals into a series of Intrinsic Mode Functions (IMF) from high frequencies to low frequencies. Then, the features of vibration signals are extracted by calculating EMD energy entropies of each IMF. Therefore, the energy feature vectors are obtained and conducted as input samples.

2.2. Random forest

An RF is a classifier consisting of a collection of tree structured classifiers $\{C(X,\theta_k), k=1,\dots\}$ where the θ_k is independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input X.

A general RF procedure is shown as Fig. 2 and described as follows:

Step 1: k samples are selected by using bootstrap sampling from training set and the sample size of each selected sample is the same as the training sets.

Step 2: k decision tree models are built for k samples and k classification results are obtained by these decision tree models.

Step 3: According to k classification results, the final classification result is decided by voting on each record.

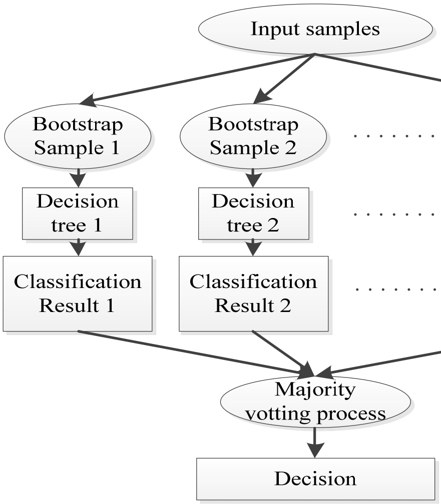

Fig. 2. Framework of RF

RF increases the differences among classification models by building different training sets, therefore extrapolation forecasting ability of ensemble classification model is enhanced. After k times training, a classification model series $\{h_1(X), h_2(X), \dots, h_k(X)\}$ is obtained, which is utilized to structure a multi-classification model system. The final classification result of the system is simple majority voting and the final classification decision is as Eq. 1:

$$H(X) = \operatorname{argmax}_{Y} \sum_{i=1}^{k} I(h_i(x) = Y), \tag{1}$$

where $H(x)$ is the ensemble classification model, h_i is a single decision tree classification model, Y is the objective output, I is an indicative function, Eq. 1 explains the final classification is decided by majority voting.

2.3. Active learning based on random forest

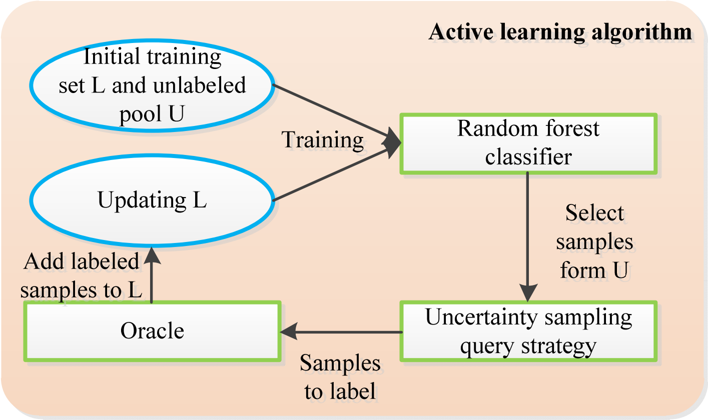

Fig. 3. Framework of the proposed AL based on RF

A general AL procedure is as follows:

Step 1: select several samples to construct an initial training set L to train a classifier.

Step 2: according to some query strategy, select a set of samples from unlabeled pool U for manually labeling.

Step 3: selected samples are added to L and the classifier is retrained by updated training set.

Step 4: repeat 2 and 3 until a stop criterion is satisfied.

The key part in AL is to select a set of samples from unlabeled pool U in Step 2. This paper adopts uncertainty sampling query strategy to select samples with maximum uncertainty to improve RF classifier. In the following, the way in this paper to select samples with maximum uncertainty is introduced.

Given a dataset by $D = L \cup U$ that is a mixture of an labeled set $L = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and an unlabeled set $U = \{x_n + 1, x_n + 2, \ldots, x_{n+m}\}$. Here x denotes a data sample of d dimension vector, y denotes its class label and $y \in \{1,2,3, \ldots, c\}$.

First, RF is trained by the labeled set L and performs classification on unlabeled data U. Therefore, the class posterior probability $p(y = j|x_i, D)$ ($i = n + 1, n + 2, \ldots, n + m$; $j = 1, 2, 3, \ldots, c$) can be provided by RF.

Second, the predictive marginal entropy H for the data sample x_i is given by Eq. 2:

$$H[y|x_i, D] = -\sum_{j=1}^{c} p(y = j|x_i, D) \times \log\big(p = (y = j|x_i, D)\big). \tag{2}$$

Third, the maximum uncertainty sample is selected by Eq. 3:

$$x_u = \operatorname{argmax}(H[y|x_i, D]), \tag{3}$$

where $u \in \{n + 1, n + 2, \ldots, n + m\}$.

Fourth, the selected unlabeled sample x_u is manually labeled and added to the labeled set L to retrain RF.

Then, repeat above five steps until some criterion is satisfied. In this way, RF classifier is trained by the most valuable and least samples.

3. Case study

An experiment was conducted to illustrate the effectiveness of the proposed method. The data used in the experiment comes from the bearing data center, which provides access to the ball bearing test data for normal and faulty bearings. A 2 hp Reliance Electric motor was applied in the experiment and acceleration data was measured at locations near to and remote from the motor bearings respectively. The motor speed is 1750 r/min and the sampling rate is 120 kHz. Faults ranging from 0.007 inches in diameter to 0.040 inches in diameter were introduced separately at the inner raceway, rolling element (i.e. ball) and outer raceway. In this experiment, three common types of bearing fault were set: inner race defect, ball defect and outer race defect.

To demonstrate the effectiveness of the proposed algorithm, it was compared with Random Forest method, which conducted classification with traditional RF method. 400 feature vectors, consisting of 100 feature vectors of normal, inner race defect, ball defect and outer race defect bearing respectively, were obtained through EMD preprocessing for original vibration signals. 40 feature vectors were randomly selected from 400 feature vectors severed as initial training set L, which including 10 feature vectors from each class. 80 feature vectors were selected randomly as testing set T and the remaining feature vectors were set as unlabeled training set U (see Table 1).

Table 1. Composition of input samples

Method	Input samples	Initial training set (L)	Unlabeled training set (U)	Testing set (T)
AL based on RF	400	40	280	80
RF	400	40	280	80

For AL based on RF, at each iteration, 1 sample from U was queried for manually labeling and added to L. The number of iteration is respectively 10, 20, 30, 40, 50 and 60. Correspondingly, the capacity of updated training set was 50, 60, 70, 80, 90 and 100. For RF, 10, 20, 30, 40, 50 and 60 samples were selected randomly from U and labeled to sever as training set.

To emphasize classification accuracy, it is supposed that the classification is effective and successful when the forecast classification probability of the testing sample is more than 90 %. For example, forecast probabilities of a sample for four patterns are 0.2 %, 0.3 %, 95 % and 4.5 % and the sample exactly belongs to the third class, therefore the classification is supposed to be effective. However, when the probabilities are 3 %, 5 %, 88 % and 4 %, the classification is considered to be failed though the majority votes are right.

The experiment is repeated for 20 times and the average classification accuracy is shown in Fig. 4. The classification accuracy of both methods increases with the increment of training samples. However, the classification accuracy of AL based on RF exceeds 4 % than RF's averagely. Hence, it is obvious that the classification accuracy of the proposed method outperforms RF method. Moreover, when the number of iteration is over 10, the classification accuracy of the proposed method maintains more than 95 %, which is a high classification accuracy considering the forecast probability of each iteration is more than 90 %.

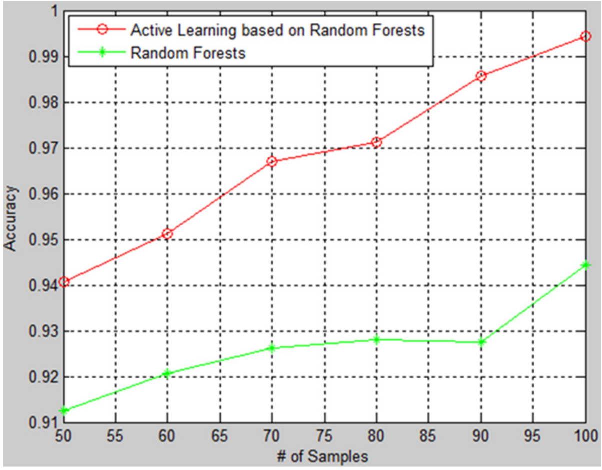

Fig. 4. Classification accuracy of AL based on RF and RF

4. Conclusion

This paper presents a method for fault diagnosis of bearing by using AL and RF. The advantage of the proposed method is to select the most valuable samples through AL, which queries samples by uncertainty sampling query strategy and select samples with maximum uncertainty for manually labeling. Moreover, RF is introduced into AL to serve as a classifier and trained by the updating training samples. Finally, AL based on RF method can perform an accurate fault diagnosis of roller bearing, which is beneficial to maintenance. An experiment was conducted to demonstrate the feasibility and effectiveness of the proposed approach. Future works are expected to be conducted to expended machinery product and component, such as hydraulic pump.

Acknowledgements

This research was supported by the Technology Foundation Program of National Defense (Grant No. Z132013B002).

References

[1] **Lu Chen, Yuan Hang, Tang Youning** Bearing performance degradation assessment and prediction based on EMD and PCA-SOM. Journal of Vibroengineering, Vol. 16, Issue 3, 2014, p. 1387-1396.

[2] **Muruganatham B., et al.** Roller element bearing fault diagnosis using singular spectrum analysis. Mechanical Systems and Signal Processing, Vol. 35, Issues 1-2, 2013, p. 150-166.

[3] **Pan Y., Chen J., Li X.** Bearing performance degradation assessment based on lifting wavelet packet decomposition and fuzzy c-means. Mechanical Systems and Signal Processing, Vol. 24, Issue 2, 2010, p. 559-566.

[4] **Cohn D. A., Ghahramani Z., Jordan M. I.** Active learning with statistical models. Journal of Artificial Intelligence Research, Vol. 4, 1996, p. 129-145.

[5] **Settles B.** Active Learning Literature Survey. University of Wisconsin, Madison, 2010.

[6] **Cohn D.** Active Learning. Encyclopedia of Machine Learning. Springer-Verlag, New York, 2011, p. 10-14.

[7] **Settles B.** Active Learning Literature Survey. Computer Sciences Technical Report 1648. University of Wisconsin – Madison, Wisconsin, 2010.

[8] **Lewis D., Gale W.** A sequential algorithm for training text classifiers. Proceedings of 17th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 1994, p. 3-12.

[9] **Di X., Han T., Yang B. S.** Application of random forest algorithm in machine fault diagnosis, Inaugural World Congress on Engineering Asset Management, Gold Coast, Australia, 2006.

[10] **Breiman L.** Random forests. Machine Learning, Vol. 45, Issue 1, 2001, p. 5-32.

[11] **Quinlan J. R.** C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, 1986.

[12] **Yang B. S., Park C. H., Kim H. J.** An efficient method of vibration diagnostics for rotating machinery using a decision tree. International Journal of Rotating Machinery, Vol. 6, Issue 1, 2000, p. 19-27.

[13] **Breiman L.** Random Forest User Notes. Statistics Department, University of California, Berkeley, ftp://ftp.stat.berkeley.edu/pub/users/breiman/notes_on_random_forests_v2.pdf, 2006.

[14] **Debarr D., Wechsler H.** Spam detection using clustering, random forests, and active learning. 6th Conference on Email and Anti-Spam, Mountain View, California, 2009.