

Research Article

A New Feature Selection Algorithm Based on the Mean Impact Variance

Weidong Cheng,¹ Tianyang Wang,¹ Weigang Wen,¹ Jianyong Li,¹ and Robert X. Gao²

¹ School of Mechanical Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China

² Department of Mechanical Engineering, University of Connecticut, Storrs, CT 06269, USA

Correspondence should be addressed to Weidong Cheng; wdcheng@bjtu.edu.cn

Received 19 January 2014; Revised 1 May 2014; Accepted 9 June 2014; Published 26 June 2014

Academic Editor: Weihua Li

Copyright © 2014 Weidong Cheng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The selection of fewer or more representative features from multidimensional features is important when the artificial neural network (ANN) algorithm is used as a classifier. In this paper, a new feature selection method called the mean impact variance (MIVAR) method is proposed to determine the feature that is more suitable for classification. Moreover, this method is constructed on the basis of the training process of the ANN algorithm. To verify the effectiveness of the proposed method, the MIVAR value is used to rank the multidimensional features of the bearing fault diagnosis. In detail, (1) 70-dimensional all waveform features are extracted from a rolling bearing vibration signal with four different operating states, (2) the corresponding MIVAR values of all 70-dimensional features are calculated to rank all features, (3) 14 groups of 10-dimensional features are separately generated according to the ranking results and the principal component analysis (PCA) algorithm and a back propagation (BP) network is constructed, and (4) the validity of the ranking result is proven by training this BP network with these seven groups of 10-dimensional features and by comparing the corresponding recognition rates. The results prove that the features with larger MIVAR value can lead to higher recognition rates.

1. Introduction

Feature extraction is key factor in pattern recognition because only sufficient and effective features can describe a given sample comprehensively and then differentiate between classes [1–4]. In general, there always exist tens or hundreds of variables to describe an object. However, the use of too many features in pattern recognition is not suitable for the following reasons: first, the number of feature dimensions should be far fewer than the number of training sets; second, too many features increase training and utilization times, which then cause the entire recognition algorithm to be time-consuming; last, prediction performance is negatively affected by inappropriate features. Because of these problems, the algorithms for multivariate feature selection and feature ranking have become the focus of much research in several areas [5].

Feature selection is a necessary preprocessing step between feature extraction and pattern recognition. Its main

purpose is to choose more sensitive features from the original multidimensional features as the subset that should maintain the same ability of recognition. To achieve this goal, several algorithms based on the principal component analysis (PCA), artificial neural network (ANN), genetic algorithm (GA), support vector machine (SVM), and pattern recognition theory-based algorithm are proposed. The PCA algorithm is the most common linear dimensionality reduction algorithm that can map multidimensional features into a space of lower dimension. Reference [6] employs this algorithm in face recognition and [7] makes use of it in machine defect classification. However, PCA can only lower the dimension by generating new features that are not suitable if the physical meaning of the features must be given [8]. As intelligent algorithms, the ANN and GA algorithms can be used in feature selection. Among them, an ANN-based feature-selection method called the UTA algorithm (named after the author [9]) is used to predict the American business cycle. The GA algorithm is used to select features for SVM [10].

However, the GA algorithm is too complicated and cannot quantitatively determine the feature that is more suitable for classification [11]. Reference [12] proposed a recursive SVM feature selection for mass-spectrometry and microarray data. Another pattern recognition theory-based algorithm was proposed in [13]. Its main principle is to maximize the quotient obtained by dividing the mean distance between the samples of different classes by the mean distance between the samples of the same classes. This method is widely used in parameter evaluation [14] because of its efficiency and clear mathematic meaning.

In this study, an interesting method called the mean impact variance (MIVAR) method is constructed to determine the feature that is more sensitive to classification. This method is obtained after the BP network training step by changing the magnitude of all the features separately. The feature with the larger MIVAR value is considered the better choice when the BP network is used as the classifier. To verify the effectiveness of this method, we use it to rank multidimensional time-domain features and select more representative features for a bearing fault diagnosis.

The rest of the paper is organized as follows: Section 2 specifies the algorithm of the MIVAR-based feature selection; Section 3 describes the databases and the all-waveform feature extraction method, which is used to generate multidimensional features; Section 4 uses the MIVAR method to rank the aforementioned multidimensional features in the order of their sensitivity and the BP network to testify the validity of the rank result. Finally, the conclusion is presented in Section 5.

2. MIVAR-Based Feature Selection Algorithm

MIVAR is a new method that can be used to select more representative features from multidimensional features. To specify the algorithm in detail, n is used to represent the number of classes, L is used to represent the total sample number of the in the training set, and l is used to represent the number of each class ($L/l = n$). The dimension of the multidimensional feature is m . i , j , and k represent the feature sequence number, the class, and the sample in one class, respectively. The specific algorithm is described as follows.

Step 1. First, m -dimensional features are extracted from the training sets of n different classes. A BP network is then constructed and trained with the training sets. The input size of the network is m , which is equal to the dimension of the multidimensional features. The output size is equal to n , which represents the type number.

Step 2. The k th sample is chosen from the training sets of the j th class, and the results are obtained by feeding the trained BP network with the corresponding m -dimensional feature. Then, the value of the i th dimension varied by $\pm 30\%$ (the other $m - 1$ dimensions are maintained at the same values) to form the following two new features:

$$P_{i,k,j,UP} = P_{i,k,j} \cdot (1 + 30\%), \quad (1)$$

$$P_{i,k,j,DOWN} = P_{i,k,j} \cdot (1 - 30\%), \quad (2)$$

where i is the dimension sequence number, k is the sequence number of the sample in the j th class, UP means that the new feature is generated by increasing the value of the i th dimension by 30%, and DOWN means that the new feature is generated by decreasing the value of the i th feature by 30%. $P_{i,k,j}$ is the original feature and $P_{i,k,j,UP}$ and $P_{i,k,j,DOWN}$ are the two new generated features. Except for the i th feature, the other $n - i$ pairs of new features should be calculated also following (1) and (2).

Step 3. The network is simulated with these n pairs of new features, and n pairs of outputs, $O_{i,k,j,UP}$ and $O_{i,k,j,DOWN}$, where i varies from one to n , are obtained.

Step 4. The absolute value of the difference between the j th bits of $O_{i,k,j,UP}$ and $O_{i,k,j,DOWN}$ is calculated. Here, we use $IV_{i,k,j}$ to denote the difference, which represents how much the i th feature affects the correct recognition of the k th sample of the j th class. $O_{i,k,j,UP}$ and $O_{i,k,j,DOWN}$ are both $j \times 1$ matrices, and the j th bit can determine whether the k th sample belongs to the j th class. We call the j th bit the judging bit of the j th class. Consider

$$IV_{i,k,NO} = |O_{i,k,j,UP} - O_{i,k,j,DOWN}|. \quad (3)$$

Step 5. The process is repeated from Step 2 to Step 4 for the other $l - 1$ samples of the j th class, and another set of $l - 1$ differences of the i th feature is obtained for the j th class. In addition to the $IV_{i,k,NO}$ obtained in Step 4, we have a total of l differences of the i th feature. By calculating the mean value of these m differences, we obtain MIV, which represents how much the i th feature influences the correct recognition of a sample of the i th as follows:

$$MIV_{i,NO} = \frac{1}{l} \sum_{k=1}^l IV_{i,k,NO}. \quad (4)$$

Step 6. The process is repeated from Step 2 to Step 5 using the samples that belong to $n - 1$ classes. This way, we can obtain all the MIVs of every feature for the samples of four different states: $MIV_{i,1}, \dots, MIV_{i,j}, \dots, MIV_{i,n}$.

Step 7. The variance of the four MIVs of each feature is calculated for the four different states, and a method called MIVAR, which represents the fluctuation in the MIVs, is obtained. Consider

$$MIVAR_i = \frac{\sum_{j=1}^n (MIV_{i,j} - \overline{MIV}_i)^2}{n}. \quad (5)$$

MIVAR is a proposed method that can determine the feature that is more suitable for classification. Thus, we should select a feature with a larger MIVAR as the one for final classification.

3. Database Description and Features Generation

In this paper, the effectiveness of the MIVAR-based feature selection algorithm is proven by selecting more representative

TABLE 1: Top three MIV sequence numbers for different classes.

Ordinal number	Top NO sequence numbers	Top IR sequence numbers	Top RE sequence numbers	Top OR sequence numbers
1	33	1	1	1
2	32	2	3	2
3	9	10	6	3

features for a bearing fault diagnosis using the data from the Bearing Data Center of Case Western Reserve University [15]. In detail, there are four data classes: normal (NO), inner race fault (IR), rolling element fault (RE), and outer race fault (OR). We choose 300 samples as the network training sets, and the sample number of each state is 75. In similar way, we choose another 100 samples as the testing sets that consist of 25 samples for NO, 25 samples for IR, 25 samples for RE, and 25 samples for OR. To acquire multidimensional features, a new feature-extraction method called all waveform feature extraction is proposed.

Step 1. The raw signal is rounded to the nearest hundredth, and the original signal data are divided into N groups (ensuring that the data from the same group are equal to each other). Then, each group number is counted and denoted with x_i , where i ranges from one to N .

Step 2. P_i represents the proportion of the i th group data number to the original signal total number, and it is obtained as follows:

$$P_i = \frac{x_i}{N}. \quad (6)$$

Step 3. The P_i curve is the probability density curve of the original signal. The four curves in Figure 1 represent the probability density curves of the vibration signal of NO, an IR, a RE, and an OR, respectively.

Step 4. New features are extracted on the basis of the probability density curve. The corresponding y -axis represents the percentage of each number in the different groups; thus, its upper bound is 100%. We choose 1/1,000 as the unit, equally divide the entire y -axis into 1,000 parts, and draw 1,000 lines parallel to the x -axis from every point along the y -axis. These 1,000 secants can be divided into two types, which are illustrated in Figure 2: the first type of secant intersects the curve more than two times (indicated by the lower two solid lines), and its corresponding features are equal to the distance between the intersections on the far right and the far left. The upper dotted line represents the second type of secant that has one or no intersections with the curve, and we let the feature obtained by this secant type be equal to zero. We let d_i denote the feature generated by the i th secant line $i = (1, 2, \dots, 1,000)$, which is the all waveform feature.

While extracting the all waveform features of the training and testing sets, we find that the maximum value of P_i s for all features is always less than 7% in all the training sets and testing sets. Therefore, we decrease the dimension of the all waveform feature from 1,000 to 70. Figure 3 shows how to

obtain 70-dimensional features using a sample of NO: the lower two lines are the first secant and the 30th secant that belong to the first type. The bold solid lines in the middle represent the all waveform features, number 1 feature and number 30 feature, extracted by these two secant lines. The upper line that has no interaction with the curve belongs to the second secant type. It can generate number 70 feature whose value is equal to zero. Using the method described above, we extracted the 70-dimensional feature vector as the feature that represents the bearing vibrational signal of the training sets and testing sets.

4. Effectiveness Proof of MIVAR-Based Feature Selection Algorithm

In this section, the MIVAR-based feature selection algorithm is proven by ranking the aforementioned 70-dimensional all waveform features.

First, a network with a structure of $70 \times 35 \times 4$ is constructed and trained with all waveform features of the samples in the training set. To calculate the MIV of every dimension for different working conditions, we choose four samples that separately belong to NO, OR, IR, and RE and calculate the corresponding MIVs of every dimension using the algorithm proposed in Section 2, Step 2 to Step 5. Figures 4(a) to 4(d) show the MIVs of all 70-dimensional features for the four different states, where the x -coordinate represents the sequence number of the feature and the y -coordinate represents their MIVs.

In Figure 4, we separately mark three features with the largest MIVs among the 70 features of each state. Considering Figure 4(a) as an example, we mark the features numbers 33, 32, and 9 beside their columns with the form “ $j(i)$,” which means that the MIV of the i th feature places j th among the 70-dimensional features for NO. In other words, if we want to recognize a sample of NO with the network, these three features have the greatest effect on the recognition results. From Figure 4, we find that the sequence numbers of the top three features of different states are different, as listed in Table 1.

In Table 1, we can see that number 1 feature places first for IR, OR, and RE, thereby having the greatest effect on sample recognition, and number 33 feature places first for NO. It seems that features numbers 1, 2, and 3 are the best three features for classification because their MIVs are relatively larger than the MIVs of the other features in the three states. Correspondingly, feature number 33 is less suitable for classification because it performs well only for state NO. However, we claim that if a feature affects most of the states at the same level, as is the case with numbers 1, 2, and 3, it

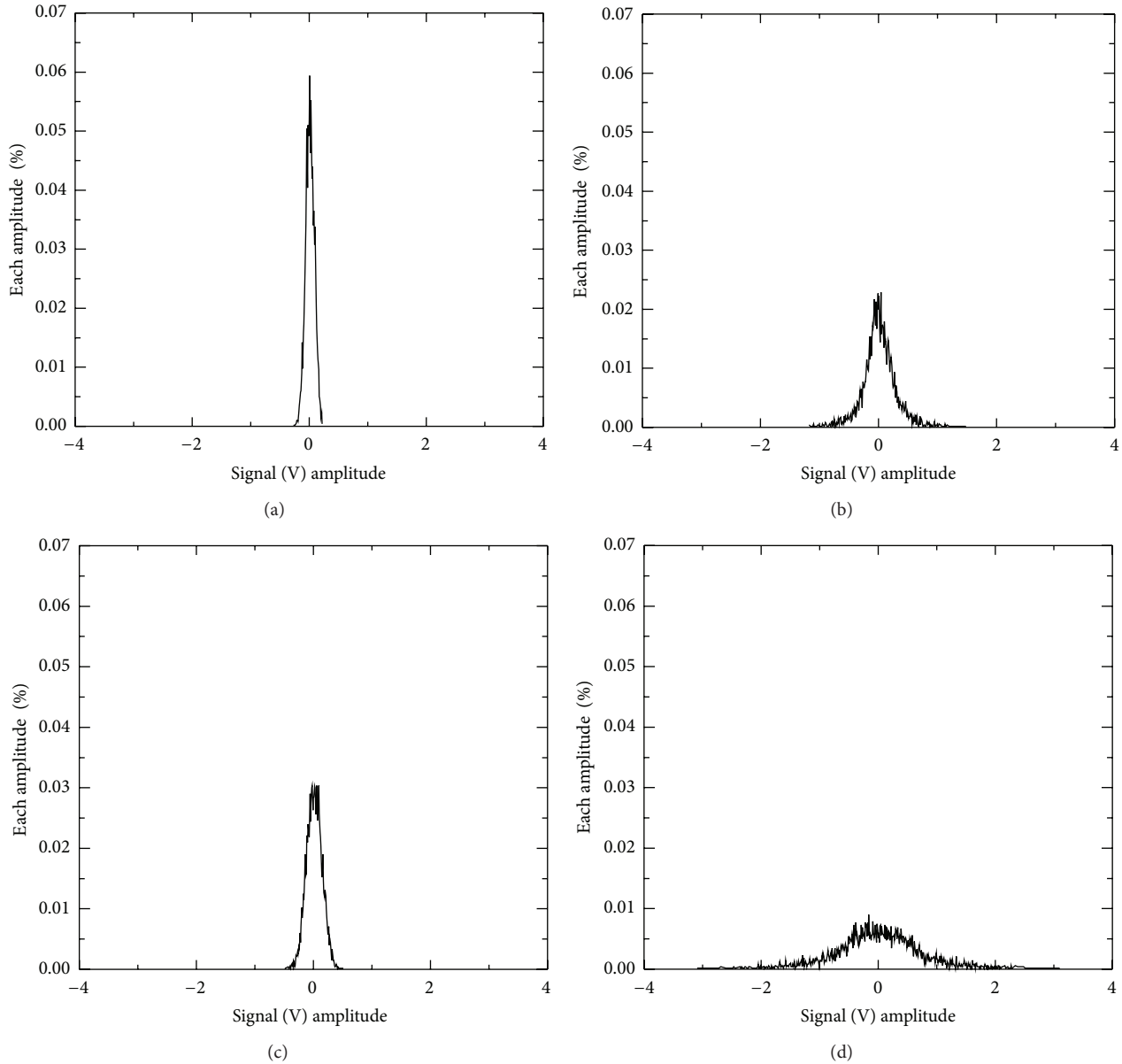


FIGURE 1: Probability density curves for four different running states: (a) NO, (b) IR, (c) RE, and (d) OR.

is probably not the most suitable feature for classification. At minimum, such a feature cannot efficiently classify the MIV types at the same level. On the contrary, feature number 33 might be more suitable for classification, even if its MIVs place first for NO only, because its MIVs for different classes are at different levels, which might make it a better feature for classification. Thus, the MIV cannot determine the feature that is more suitable for classification.

Second, the corresponding MIVARs of every feature are calculated by (5). Figure 5 shows the MIVARs of every feature in a histogram with the top ten features denoted in the form of “ $j(i)$,” which means that the MIVAR of the i th feature places j th among the 70-dimensional feature. The x -coordinate represents the feature sequence number, and the y -coordinate represents the MIVAR value.

In Figure 5, we can readily find the top ten features with the largest MIVARs among the 70-dimensional features. According to the MIVAR method, these ten features have the greatest effect on classification. We can see that the sequence numbers of these ten features are not the same as those in Table 2. As Figure 5 suggests, number 33 feature, whose MIVAR value places first among the 70 features, is the most efficient feature. However, it only performs well in just one state in Table 2. Only half of the features listed in Table 2 are marked in Figure 5; they are features numbers 1, 3, 9, 32, and 33. Among them, numbers 9, 32, and 33 perform well in only one state. According to the MIVAR method, numbers 33, 1, 32, 9, 3, 38, 4, 28, 35, and 19 are selected as the most efficient features for classification. The specific ranks of all 70-dimensional features are listed in Table 2.

TABLE 2: MIVAR ranks of 70-dimensional features.

Ranking number	1	2	3	4	5	6	7	8	9	10
Sequence number	33	1	32	9	3	38	4	28	35	19
Ranking number	11	12	13	14	15	16	17	18	19	20
Sequence number	40	10	2	11	23	7	47	12	6	31
Ranking number	21	22	23	24	25	26	27	28	29	30
Sequence number	17	8	22	29	25	34	13	46	16	49
Ranking number	31	32	33	34	35	36	37	38	39	40
Sequence number	44	39	14	20	15	43	30	41	27	42
Ranking number	41	42	43	44	45	46	47	48	49	50
Sequence number	24	50	58	18	21	5	57	37	45	52
Ranking number	51	52	53	54	55	56	57	58	59	60
Sequence number	26	54	55	48	53	36	51	59	61	62
Ranking number	61	62	63	64	65	66	67	68	69	70
Sequence number	56	60	63	64	66	67	65	68	70	69

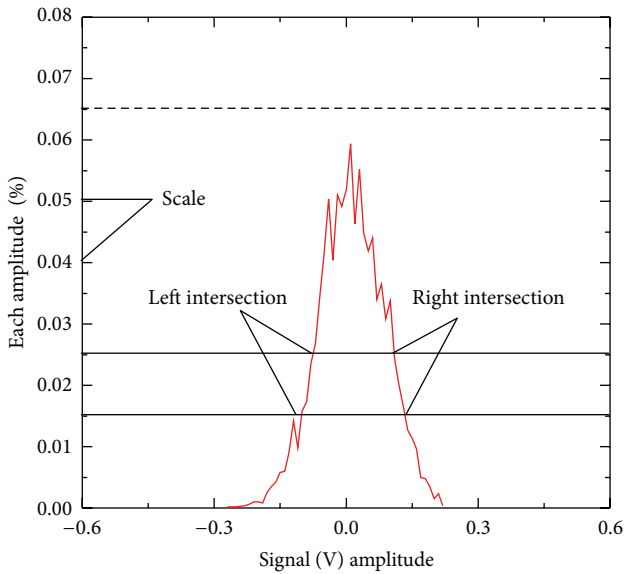


FIGURE 2: Schematic diagram for conceptual explanation.

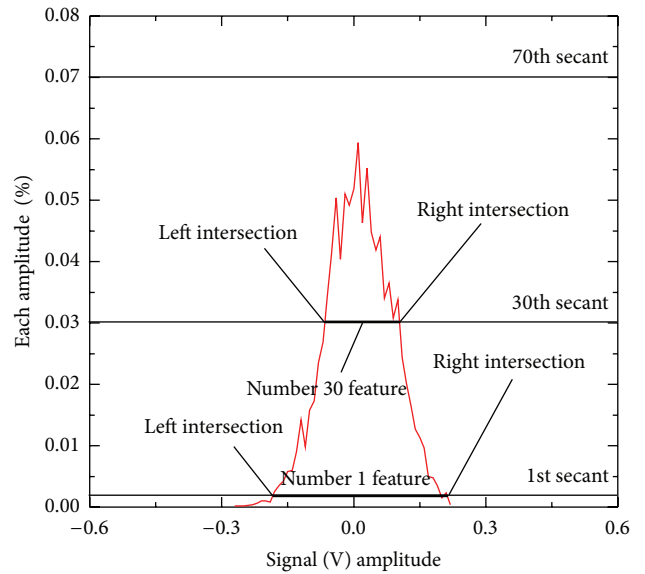


FIGURE 3: All waveform features with 70 dimensions.

Third, several comparisons are presented to prove the validity of the ranking results by constructing 14 groups of features as follows:

- (1) Features 33, 1, 32, 9, 3, 38, 4, 28, 35, and 19, whose sequence numbers are the top ten;
- (2) Features 40, 10, 2, 11, 23, 7, 47, 12, 6, and 31, whose sequence numbers are the second top ten;
- (3) Features 17, 8, 22, 29, 25, 34, 13, 46, 16, and 49, whose sequence numbers are the third top ten;
- (4) Features 44, 39, 14, 20, 15, 43, 30, 41, 27, and 42, whose sequence numbers are the fourth top ten;
- (5) Features 24, 50, 58, 18, 21, 5, 57, 37, 45, and 52, whose sequence numbers are the fifth top ten;
- (6) Features 26, 54, 55, 48, 53, 36, 51, 59, 61, and 62, whose sequence numbers are the sixth top ten;
- (7) Features 56, 60, 63, 64, 66, 67, 65, 68, 70, and 69, whose sequence numbers are the bottom ten;
- (8) new constructed 10-dimensional features with the top ten scores based on the PCA algorithm;
- (9) new constructed 10-dimensional features with the second ten scores based on the PCA algorithm;
- (10) new constructed 10-dimensional features with the third ten scores based on the PCA algorithm;
- (11) new constructed 10-dimensional features with the fourth ten scores based on the PCA algorithm;
- (12) new constructed 10-dimensional features with the fifth ten scores based on the PCA algorithm;
- (13) new constructed 10-dimensional features with the sixth ten scores based on the PCA algorithm;

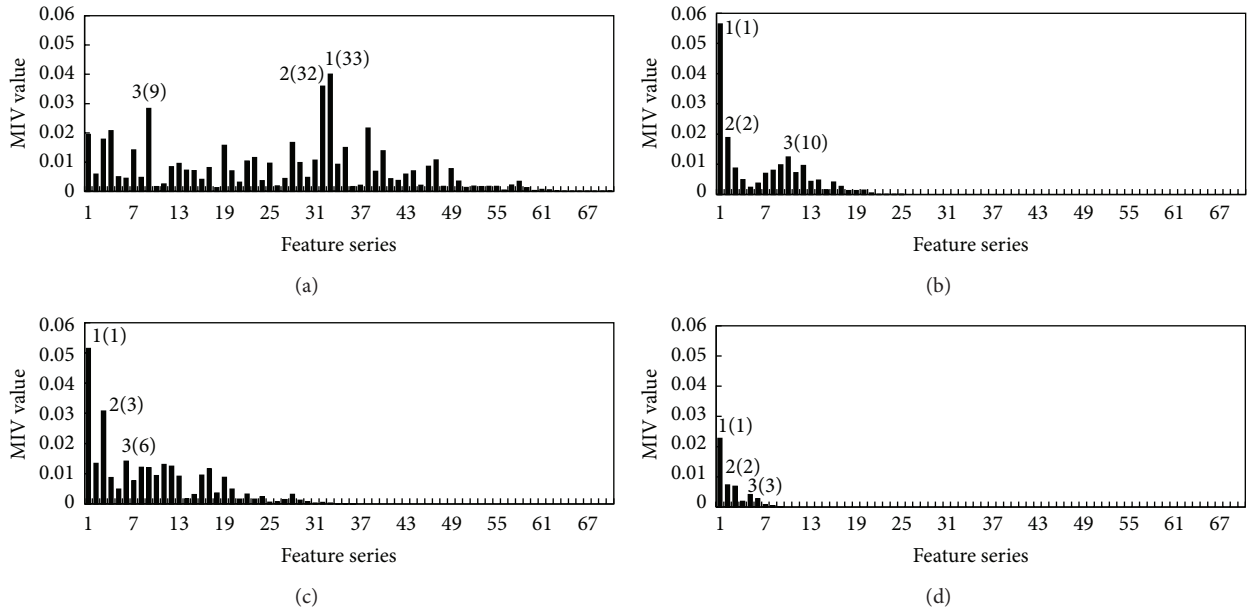


FIGURE 4: Feature MIVs in different states: (a) NO, (b) IR, (c) OR, and (d) RE.

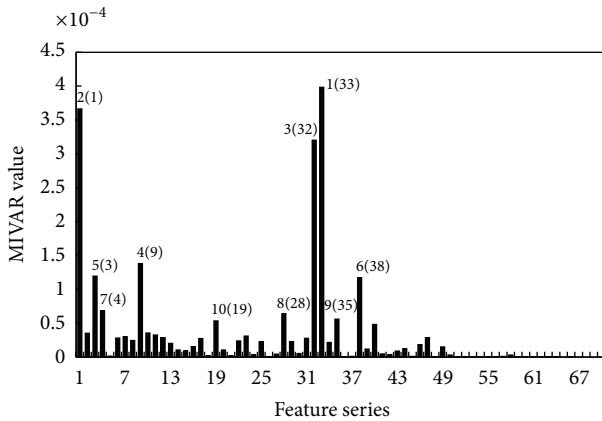


FIGURE 5: MIVAR histogram of every feature.

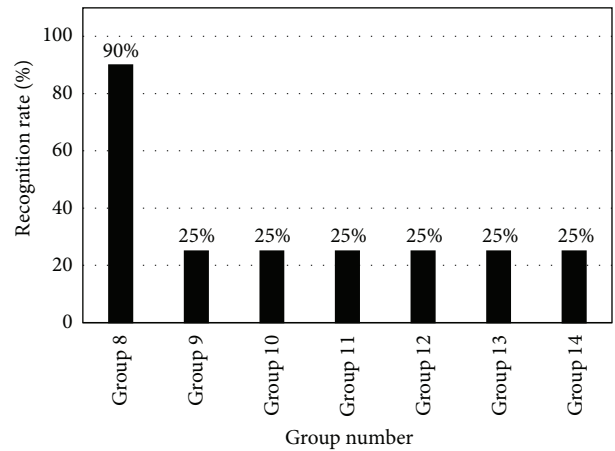


FIGURE 7: Recognition rate of PCA-based features.

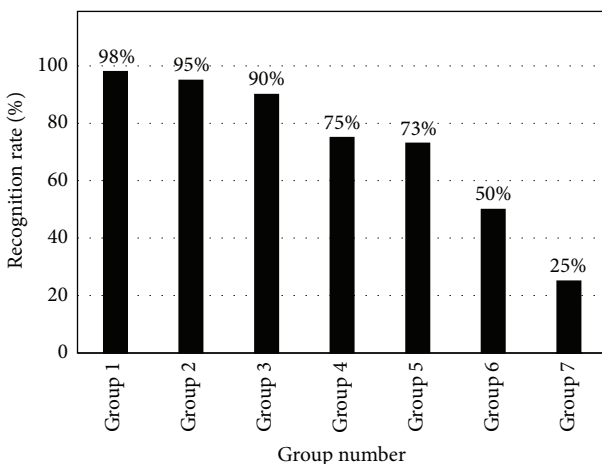


FIGURE 6: Recognition rate of MIVAR-based features.

(14) new constructed 10-dimensional features with the bottom ten scores based on the PCA algorithm.

In detail, 14 new training sets and testing sets are generated to train and test a newly constructed network with the structure $10 \times 6 \times 4$. The corresponding recognition rates listed in Table 3 can be then used to prove whether the MIVAR-based ranking result is appropriate. To ensure the fairness of the comparison, the initial weights and training times during the training processes of the different groups should be the same.

According to the comparison results listed in Table 3, we can see that the features whose MIVAR ranking sequences are the top ten and the second top ten can lead to recognition rates of 98% and 95%, respectively. Moreover, the recognition rate decreases from 90% to 25% when the features in Groups 3 to 7 are used to represent the vibration signal. It is

TABLE 3: Recognition rate of different groups.

Network input	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7
Recognition rate (%)	98	95	90	75	73	50	25
Network input	Group 8	Group 9	Group 10	Group 11	Group 12	Group 13	Group 14
Recognition rate (%)	90	25	25	25	25	25	25

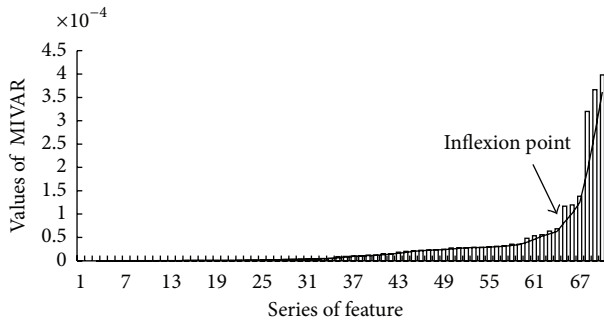


FIGURE 8: Ranking results of all 70-dimensional all waveform features.

proven that the MIVAR-based feature selection algorithm can be used to select more representative features from the multidimensional features. As for the features generated by the PCA algorithm, we use the features in Group 8 to Group 14 train of the same network. It is shown that the new constructed 10-dimensional features with the top ten scores can lead to the recognition rate of 90% and all the other 6 groups of 10-dimensional features can only lead to 25% recognition rate. Figures 6 and 7 show their histograms. By comparing the recognition rates in Figures 6 and 7, we find that the recognition rate of Group 8 is not as good as the ones of Group 1 and 2 which can partly explain the advantage of MIVAR based feature selection algorithm. It should be mentioned that the principal component contribution rates summation of the top 3 vectors is more than 95%. So, the features in Group 9 to Group 14 are useless for the final classification.

Last, we display the 70-dimensional all waveform features in the order of the corresponding MIVAR value in Figure 8, where the MIVAR of all the features are displayed by hollow histograms, and the corresponding trend line is presented simultaneously. We can see that the MIVAR value of the 65th histogram indicated by the arrow is obviously larger than for the 64th and the changing rate of the trend line after the 65th feature is much larger than before it. This way, we consider 65th feature as the inflexion point and recommend the features (number 33, 1, 32, 9, 3, and 38) whose MIVAR values are the top six largest most representative features when the BP network is used as the classifier.

5. Conclusion

In this paper, a MIVAR method was proposed to determine the feature that is more suitable for ANN-based classification. The MIVAR values of all the features were calculated by

changing the input vectors and then measuring the differences of the output vectors after the training process of the BP network. It was proven that using the features with higher MIVAR values can lead to higher recognition rates.

As an example, 70-dimensional all waveform features of a rolling bearing vibration signal were ranked based on the MIVAR method. The features with the largest ten MIVAR values can lead to a recognition rate of 98%, and the corresponding recognition rate of the second, third, fourth, fifth, sixth, and seventh largest ten MIVAR values are 95%, 90%, 75%, 73%, 50%, and 25%, respectively. This decreased recognition rate proved the effectiveness of the MIVAR method. To compare the effectiveness of the MIVAR method to the traditional algorithm, the PCA algorithm is then used to generate 7 groups of 10-dimensional features (Group 8 to Group 14). And the 10-dimensional features with the top ten scores can lead to a recognition rate of 90%, which is not as good as that for Groups 1 and 2.

In addition, it should be pointed out that the discussion is limited to the use of time-domain features to describe a steady vibration signal. Moreover, the MIVAR algorithm can be extended also to the selection of frequency-domain features.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (51275030) and the “Fundamental Research Funds for the Central Universities M11JB00210.”

References

- [1] A. Jain and D. Zongker, “Feature selection: evaluation, application, and small sample performance,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997.
- [2] L. Yu and H. Liu, “Efficient feature selection via analysis of relevance and redundancy,” *The Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.
- [3] H. Brunzell and J. Eriksson, “Feature reduction for classification of multidimensional data,” *Pattern Recognition*, vol. 33, no. 10, pp. 1741–1748, 2000.
- [4] S. B. Serpico, M. D’Inca, F. Melgani, and G. Moser, “Comparison of feature reduction techniques for classification of hyperspectral remote sensing data,” in *Image and Signal Processing for Remote Sensing VIII*, 347, vol. 4885 of *Proceedings of SPIE*,

International Society for Optics and Photonics, Crete, Greece, September 2002.

- [5] G. Isabelle and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
- [6] H. K. Ekenel and B. Sankur, "Feature selection in the independent component subspace for face recognition," *Pattern Recognition Letters*, vol. 25, no. 12, pp. 1377–1388, 2004.
- [7] A. Malhi and R. X. Gao, "PCA-based feature selection scheme for machine defect classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 53, no. 6, pp. 1517–1525, 2004.
- [8] Y. Lu, I. Cohen, X. S. Zhou, and Q. Tian, "Feature selection using principal feature analysis," in *Proceedings of the 15th International Conference on Multimedia*, New York, NY, USA, 2007.
- [9] J. Utans, J. Moody, S. Rehfuss, and H. Siegelmann, "Input variable selection for neural networks: application to predicting the U.S. business cycle," in *Proceedings of the IEEE/IAFE Computational Intelligence for Financial Engineering (CIFER '95)*, pp. 118–122, New York, NY, USA, April 1995.
- [10] C. Huang and C. Wang, "A GA-based feature selection and parameters optimization for support vector machines," *Expert Systems with Applications*, vol. 31, no. 2, pp. 231–240, 2006.
- [11] B. Samanta, K. R. Al-Balushi, and S. A. Al-Araimi, "Artificial neural networks and genetic algorithm for bearing fault detection," *Soft Computing*, vol. 10, no. 3, pp. 264–271, 2006.
- [12] X. Zhang, X. Lu, Q. Shi et al., "Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data," *BMC Bioinformatics*, vol. 7, article 197, 2006.
- [13] B. S. Yang, T. Han, and J. L. An, "ART-KOHONEN neural network for fault diagnosis of rotating machinery," *Mechanical Systems and Signal Processing*, vol. 18, no. 3, pp. 645–657, 2004.
- [14] Y. Lei, Z. He, Y. Zi, and Q. Hu, "Fault diagnosis of rotating machinery based on multiple ANFIS combination with GAs," *Mechanical Systems and Signal Processing*, vol. 21, no. 5, pp. 2280–2294, 2007.
- [15] <http://csegroups.case.edu/bearingdatacenter/home>.



Hindawi

Submit your manuscripts at
<http://www.hindawi.com>

