Tribology International Vol. 30, No. 10, pp. 739–752, 1997

Printed in Great Britain
0301–679X/97/$17.00 + 0.00

PII: S0301–679X(97)00056–X

Using neural networks for the diagnosis of localized defects in ball bearings

M. Subrahmanyam* and C. Sujatha†‡

Two neural network based approaches, a multilayered feed forward neural network trained with supervised Error Back Propagation technique and an unsupervised Adaptive Resonance Theory-2 (ART2) based neural network were used for automatic detection/diagnosis of localized defects in ball bearings. Vibration acceleration signals were collected from a normal bearing and two different defective bearings under various load and speed conditions. The signals were processed to obtain various statistical parameters, which are good indicators of bearing condition, and these inputs were used to train the neural network and the output represented the ball bearing states. The trained neural networks were used for the recognition of ball bearing states. The results showed that the trained neural networks were able to distinguish a normal bearing from defective bearings with 100% reliability. Moreover, the networks were able to classify the ball bearings into different states with success rates better than those achieved with the best among the state-of-the-art techniques. © 1998 Elsevier Science Ltd. All rights reserved

Keywords: *defect diagnosis, bearing vibration, ball bearings, neural networks*

Introduction

Machine monitoring and diagnosis involves intermittent or continuous collection and interpretation of data relating to the condition of critical components. Constant monitoring of machinery has been considered to be an essential and integral part of any modern manufacturing facility, because any unexpected failure or breakdown will result in costly consequences. Adequate monitoring greatly reduces the frequency of breakdowns before they actually occur. Therefore, a machine monitoring system can be seen as a decision support tool which is capable of identifying the failure of a machine component or system, and which also predicts its occurrence from a symptom[1].

*Fired Heaters Section, Engineers India Limited, New Delhi 110 001, India
†Machine Dynamics Laboratory, Department of Applied Mechanics, Indian Institute of Technology, Madras 600 036, India
‡Corresponding author
Received 20 June 1995; revised 17 January 1996; accepted 23 July 1997

Bearings are essential components of most machinery and their operating conditions influence directly the operation of the whole machinery. The majority of the problems in rotating machines are caused by faulty bearings[2]. The classical failure mode of rolling element bearings is localized defects, in which a sizable piece of the contact surface is dislodged during operation, as a result of fatigue cracking in the bearing metal under cyclic contact stressing[3]. In industry, it is required not only to diagnose the faults of rolling element bearings in operation, but also to assess the quality of new bearings before use[4,5].

The existing techniques for detecting localized defects in bearings and the rate of success achieved for classifying the condition of bearings in each scheme are shown in Table 1[3]. From the table, it is observed that the highest success rate possible to detect localized defects in the bearing is around 92%. Moreover, most of the bearing condition monitoring methods in vogue need the assistance of an expert in the interpretation of results, and the success rates achieved are less than

Table 1 The time utilization of a typical machining center

Activity	Time (%)
Metal cutting	23
Positioning and tool changing	27
Gauging and loading	18
Set-up	5
Waiting and idle	14
Repair and technical	13

those required by the modern automated industries. Hence, the need arises for the development of a new scheme to outperform all the state-of-the-art techniques.

Neural networks

The present study aims at developing a method of bearing condition estimation using neural networks which give higher success rates in their condition estimation than the existing methods. Two types of neural network models, a multi-layered feed forward neural network trained with Error Back Propagation (EBP) algorithm and an unsupervised Adaptive Resonance Theory-2 (ART2) based single layered competitive neural network have been used. These neural networks have an edge over conventional monitoring methods in that they can classify the condition of machine components even in the absence of explicit input–output relationships. Besides, the networks can classify well even in the case of noisy or incomplete information obtained from the signals being monitored. For the bearing condition estimation problem, both types of neural networks have been used, the exact architecture and the training parameters of the network being problem dependent.

An MLFNN trained with EBP algorithm

A multi-layered neural network has one or more hidden layers along with the input and output layer (Fig 1). Each layer has a certain number of nodes and all the nodes in one layer are connected with all the other nodes in the succeeding layer. Associated with each connection, a numerical value is assigned, which is termed as weight, where the actual associative knowledge between the inputs and outputs is stored[6].

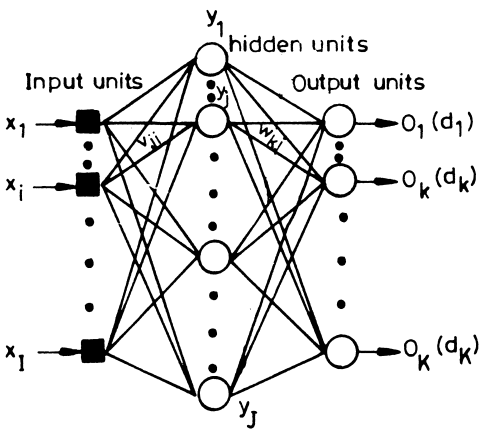

Fig. 1 A two layer feed forward neural network

Input patterns are submitted during the EBP training sequentially. If a pattern is submitted, and its classification or association is determined to be erroneous, the weights are adjusted so that current least square classification error is reduced[7]. Usually, mapping error is cumulative and computed over the full training set[6]. The Total Sum Squared (TSS) error is used as a way of measuring the best fit to the data, and is as follows.

$$\text{TSS Error, } E = \sum_{p=1}^{P} \frac{1}{2} \sum_{k=1}^{K} (d_{kp} - o_{kp})^2 \tag{1}$$

where P is the number of patterns in the training data set, K is the number of output nodes of the network, d_{kp} is the target output for the pth pattern at output node k, and o_{kp} is the actual output for the pth pattern presentation at the kth output node. The squared error is averaged over all output nodes in the output layer, as well as all patterns in the training set[8].

At each node in the hidden and output layers, two functions are performed: (i) weighted summation of all inputs (integration function) and (ii) generation of node outputs (activation function). During the forward pass, output at hidden node, j, i.e.

$$y_j = F(\text{NET}_j) \tag{2}$$

where

$$\text{NET}_j = \sum_{i=1}^{I} v_{ji} x_i \tag{3}$$

where v_{ji} represents the weight on connection between nodes i and j. Outputs of the hidden nodes act as inputs to the output layer, and similarly the output at the output node k (o_k) is computed.

The gradient descent search is performed to reduce the error (E) through the adjustment of weights. The error (E) is back propagated to change the output and hidden layer weights.

Mathematically, change in weight,

$$\Delta w_{kj} \propto (\ -\ ) \frac{\partial E}{\partial w_{kj}} \tag{4}$$

or

$$\Delta w_{kj} = \eta \delta_{ok} y_j \text{ (after simplification)} \tag{5}$$

where η is the constant of proportionality called learning rate parameter and δ_{ok} is the error at node k. So the weight adjustment expression is as follows.

Considering the sigmoid activation function (Fig 2), i.e.

$$F(\text{NET}_k) = \frac{1}{(1 + e^{-\text{NET}k})} \tag{6}$$

After simplification, the following expressions for error terms can be obtained.

$$\delta_{ok} = (d_k - o_k) o_k (1 - o_k) \text{ for output nodes} \tag{7}$$

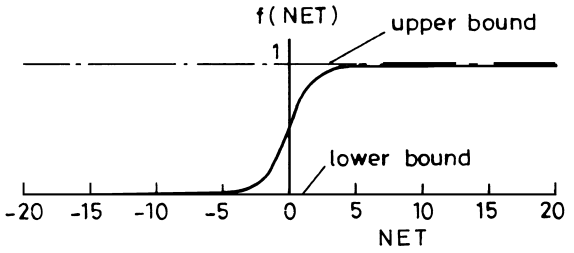

Fig. 2 Sigmoid activation function

$$\delta_{oj} = y_j(1 - y_j)\sum_{k=1}^{K} \delta_{ok} w_{kj} \text{ for hidden nodes} \quad (8)$$

It is to be noted that if the network has more than one hidden layer, the same procedure is extended to compute the weight adjustments of the hidden layer, wherein the next immediate layer acts as an output layer in weight adjustment determination of the hidden layer under consideration. The final weight adjustment expression is as follows.

$$w_{kj}(t) = w_{kj}(t - 1) + \Delta w_{kj}(t) \quad (9)$$

In the above equation, t refers to the current training cycle (at time t) and $t - 1$ refers to the most recent training cycle (at time $t - 1$).

To accelerate the convergence of EBP learning process a momentum term is introduced. The method involves supplementing the current weight adjustments with a fraction of the recent weight adjustments. This is done according to the formula,

$$\Delta w(t) = (-)\eta \nabla E(t) + \alpha \Delta w(t - 1) \quad (10)$$

α is a user-selected positive momentum constant[6,8].

To further accelerate the EBP training, expected values of source nodes are used for updating the weights. The expected value of a node can be approximated as the sum of the output node and a fraction of its error term[9], and the modified EBP training rule is as follows:

$$\Delta w_{kj} = - \eta(y_j + \beta\delta_j)\delta_k \quad (11)$$

where β is a constant, named as accelerating constant. The above expression reverts to the original rule when β equals zero. β can usually be set to 1.0, and superior results (over the conventional EBP training) are consistently obtained. In many cases, however, higher values of β can further accelerate training. As with η, a further increase in β, beyond some problem-specific value, results in oscillations and non convergence.

The values of the input and output variables need to be scaled to a range that is within the bounds of the output node's sigmoid function (0 and 1). The common practice is to use only the relatively linear portion of the sigmoid function, between 0.1 and 0.9, for the selected logistic function, as shown in Fig 2. The scaling is simple. Let the maximum and minimum values of the dependent variable be V_{max} and V_{min}, and let the maximum and minimum scaled target values be T_{max} (0.9) and T_{min} (0.1). Then, for any example, the target value (normalized value) T_{ar} is a function of the value V_{al} (the value to be normalized)[8]:

$$T_{ar} = T_{min} + \frac{(V_{al} - V_{min})}{(V_{max} - V_{min})}(T_{max} - T_{min}) \quad (12)$$

As there is no fixed rule to determine the neural network architecture, different network architectures must be tried to decide which network gives the best results. The various steps adopted in developing this neural network are as follows:

(i) The optimal architectures of the network, for the chosen problem have been arrived at by trial and error.
(ii) The optimal training parameters, i.e. learning rate (η), momentum term (α) and accelerating constant (β) have been determined by trial and error.
(iii) The Total Sum Squared Error (TSS Error) on a validation sample during training has been considered as the main criterion to stop training, keeping in view the classification performance.
(iv) For identifying the condition of a drill, drill average flank wear data were presented as target outputs to the neural network.

An SLCNN based on ART2 algorithm

The supervised BP learning requires external target outputs, the measurement of which demands excessive time and expensive equipment. Moreover, when operating conditions change, a new training set must be compiled and labeled, and the whole training must be repeated. Also, evaluation of target output is itself a tough task in the bearing fault diagnosis problem, as this is not directly measurable. To overcome these bottlenecks of supervised EBP learning, an unsupervised ART2 based neural network has been proposed, where the learning process does not need target outputs and is much faster.

The primary function of an ART2 module is to carry out clustering of input pattern sets such that patterns of the same cluster exhibit a certain degree of similarity. For a clustering procedure, it is necessary to define a similarity measure to be used for evaluating how close the patterns are. The most common one is the Euclidean distance[10] method and is adopted in the present study and defined as given below.

$$\text{Euclidean distance, } d_j(X_i, X_j) = \sqrt{\sum_p (X_{ip}(k) - X_{jp}(k))^2} \quad (13)$$

where X_i and X_j are two column vectors, which represent the input and weight vectors (patterns) in practice.

The ART2 network is shown schematically in Fig 3. It has two layers: the first is the input/comparison layer and the second is the output/recognition layer. These layers are connected together with extensive use of feedback from the output layer to the input layer along with the feed forward connections. Associated with each connection, the ART2 network has feed forward weights (w_{ji}s) from the input layer to the output layer and feed back weights (t_{ij}s), from the output layer to the input layer. Between the input and output layers there is also a reset circuit which is actually responsible for comparing the evaluated Euclidean distance

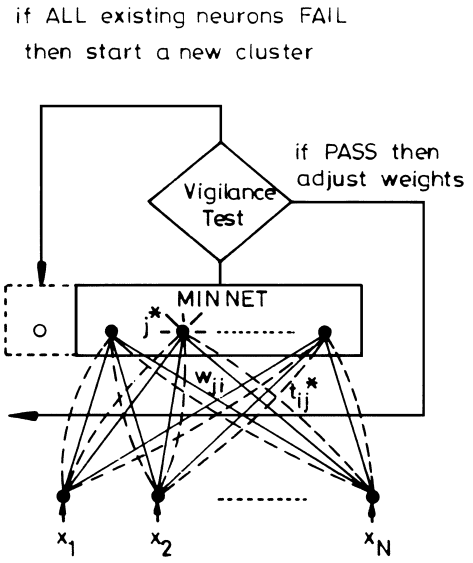

Fig. 3 The configuration of ART2 network

(making use of the current inputs and the most recent weights) to a vigilance threshold that determines whether the pattern under consideration pertains to one of the already generated clusters or a new class (cluster) must be created[11].

The number of nodes in the input layer depends on the number of input features in each input pattern set and the number of output nodes may be unknown *a priori*, and the ART2 mechanism allows adaptive expansion of the output layer until an adequate size (required number of classes) is reached. The ART2 algorithm is designed for clustering continuous valued, e.g. real valued patterns. The salient steps involved in clustering are as follows.

(1) Initialization: Weights $w_{ji}(t)$ and $t_{ij}(t)$ are initialized, and a value is set for ρ, which is problem dependent. In the present case, $w_{ji}(t) = t_{ij}(t)$.
(2) Pattern presentation: A new input vector X_i is presented. [For the first input pattern (X_0), the new output neuron (J_0) is created with weights $W_0 = X_0$.]
(3) Compute matching: Given a new training pattern, a MINNET (minimum net) is adopted to select the winner, which yields the minimum distance, Euclidean distance $d_j(X_i, W_j)$, where

$$d_j^2(X_i, W_j) = \sum_p (X_{ip}(t) - W_{jp}(t))^2 \quad (14)$$

(4) Vigilance test: A neuron j^* passes the vigilance test if and only if $d_j(X_i, W_j) < \rho$, where the vigilance threshold value (ρ) determines the radius of the cluster.
(5) Test failed state: If a neuron fails the vigilance test, a new neuron k is created with weight vector $W_k = X_i$.
(6) Test passed state: When the winner passes the vigilance test, the weight of the winner j^* is adjusted by

$$W_j^*(t+1) = \frac{X_i + W_j^*(t)|\text{cluster}_j^*(t)|}{|\text{cluster}_j^*(t)| + 1} \quad (15)$$

where $|\text{cluster}_j^*(t)|$ denotes the number of nodes in the cluster j at time t.
(vii) Repeat: This procedure is continued till all training patterns have been presented to the network.

The vigilance parameter controls the resolution of the classification process. A low choice of ρ produces a high resolution of classification process, creating larger class types and vice versa. In training, the network ρ is varied by trial and error till the required number of output nodes (clusters/classes/exemplars) is adaptively generated, so as to attribute meaningful interpretation to each cluster generated.

If all input patterns are clustered based on the ART2 learning algorithm, some input patterns may be actually closer to the centroids (updated weights/coordinates) of the other clusters. Also, the ART2 is sensitive to the order of presentation of the input patterns and yields a different clustering on the same input patterns when they are presented in the reverse order, even though the vigilance threshold remains the same. To overcome these effects, reclustering (making use of final weights obtained through initial clustering as initial weights) is done with reverse order of presentation of the input patterns.

In the classification or testing phase, the process is similar to the training phase with the only exemption that the network makes use of the finally arrived at weights to achieve the required classification of all test patterns.

Experimental studies

Objective

The present study focuses on finding out whether a bearing under consideration is good or bad. If it is a bad bearing, the aim is to classify the fault, i.e. ball defect or outer race defect. In a manufacturing plant, the answers to these questions are of immense use, as in producing a component of very high accuracy, one would like to avoid faulty bearings. Besides, faulty bearings often give rise to other problems in rotating machinery.

Selection of parameters

Vibration information has been widely adopted for malfunction detection of bearings in industry. The rolling element bearing supports a load by means of rolling elements and therefore has an unavoidable tendency to produce undesirable vibration and sound. Many studies have therefore been conducted to find a way of preventing or reducing the vibration and sound from rolling element bearings[12,13]. However, the previous studies are, for the most part, concerned with normal rolling element bearings except for a few studies carried out to detect localized defects[14,15]. The vibration and sound from rolling element bearings seem to constitute an important subject of inquiry from the stand point of early detection of defective rolling element bearings. A thorough analysis of bearings and their effects on vibrations can be found in the paper by Igarashi et al.[16]. Malfunction alarms for rolling element

bearings can be based upon the detection of localized defects (which constitute a common mode of failure) by processing the bearing vibration acceleration signals[3].

Liu and Iyer[17] used a feed forward neural network for the recognition of different states of roller bearings by feeding various parameters obtained by processing vibration acceleration signals as inputs to the neural network which decides the bearing condition. Unal[18] made use of jerk fields (time rate of change of acceleration data) as discriminating features in the training of Artificial Neural Networks (ANNs) for the fault diagnosis of ball bearings in a paper mill. Chiou *et al*.[19] made use of vibration signatures in the ultrasonic frequency range (beyond 100 kHz) to train a neural network in order to identify the condition of needle bearings.

Parameters monitored

Vibration acceleration signals have been processed using a signal analyzer and the following parameters are obtained, as suggested in the literature and supported by the observations made in the present study, for estimating the condition of bearings.

(1) Peak value of amplitude in PSR.
(2) Average of top five peak values of amplitude in PSR.
(3) Peak value of autocorrelation function in PSR.
(4) Peak value of amplitude in HFR.
(5) Average of top five peak values of amplitude in HFR.
(6) Peak value of autocorrelation in HFR.
(7) Standard deviation (0–10 kHz).
(8) Kurtosis (0–10 kHz).

These parameters have been fed as inputs to an Artificial Neural Network (ANN), which judges the condition of the bearing as good or bad, and if it is found to be bad, it also pinpoints the fault.

Bearing test rig and instrumentation

The schematic of the bearing test rig on which experiments were carried out is shown in Fig 4. The rig consists of a short shaft, supported between two bearings at its ends. The shaft is coupled to a variable speed dc motor through a coupling arrangement which ensures that the shaft does not experience any of the vibrations from the motor. Also, the coupling accommodates any misalignment present in the assembly. The dc motor is connected to a speed control unit, an auto transformer, to achieve variable speeds. With the present motor, speeds up to 3000 rpm are possible.

The support bearing adjacent to the coupling is an SKF2310 double row self-aligning ball bearing. The other support bearing (the test bearing) is an SKF6307 single row deep groove ball bearing, on which tests are conducted. The specification of the test bearing is given in Table 2.

The loading of the shaft is done through an SKF6411 heavy duty deep groove ball bearing. This load bearing is mounted in between the test bearing and the support bearing. The load bearing is kept near the test bearing, so that nearly three-quarters of the load applied on it would be transferred to the test bearing. The load bearing is equipped with an outer casing, having an eye bolt at the bottom, which is connected to a hydraulic jack via an extensometer used for measuring the applied load, using a rope and pulley arrangement. The hydraulic jack is connected to a hydraulic pump, which is operated manually.

A piezo-electric accelerometer (Bruel and Kjaer, 4332) is stud-mounted on the housing of the test bearing. The accelerometer is connected to a charge amplifier (Bruel and Kjaer, 2626), the output of which is fed to a signal analyzer (SD 380, Scientific Atlanta) to analyze the signals. The signals are monitored on a digital storage oscilloscope (Kikusui Electronics) and also simultaneously recorded on a magnetic tape using an eight channel FM tape recorder (Racal, England). The recorded signals are replayed and processed in the signal analyzer to extract different features from the vibration acceleration signal.

Experimental procedure

Three SKF6307 single row deep groove ball bearings are used in the present study. The details of this bearing are as given in Table 2. One is a brand new bearing (assumed to be free from defects) and defects were created in the other two bearings using EDM, in order to keep size and depth of the dent under control. A dent was created in one of the balls of one bearing and another created in the inside groove of the outer race of the other bearing. The size of the defect is about 1.0 mm in diameter and 0.5 mm in depth, and is the same for both bearings. Before installing, each bearing is properly lubricated with grease and mounted on the shaft. After allowing initial running of the bearing for some time, the acceleration signal from the pick-up mounted on the test bearing is fed to a charge amplifier and the conditioned signal is tape recorded at a tape speed of 3 3/4 inch/s to cater to a frequency range of 0–10 kHz. This test is done for various load and speed combinations, the loads being 2.5–10 kN in steps of 2.5 kN and the speeds being 80, 100, 120, 150, 200 and 250 rpm.

Signal analysis

Significant frequency regions

Bentley[13] pointed out that vibrations produced by machines equipped with rolling element bearings always have components in the following frequency regions:

(1) Rotor Vibration Region: This includes the range of one-quarter to three times the shaft rotating speed, and is the direct result of rotor related malfunctions like imbalance, misalignment, rotor instability, etc. which must be corrected to eliminate bearing overloading and subsequent failure.
(2) Prime Spike Region (PSR): This includes the frequency range which covers bearing characteristic defect frequencies generated by rolling elements traversing either an inner or outer race

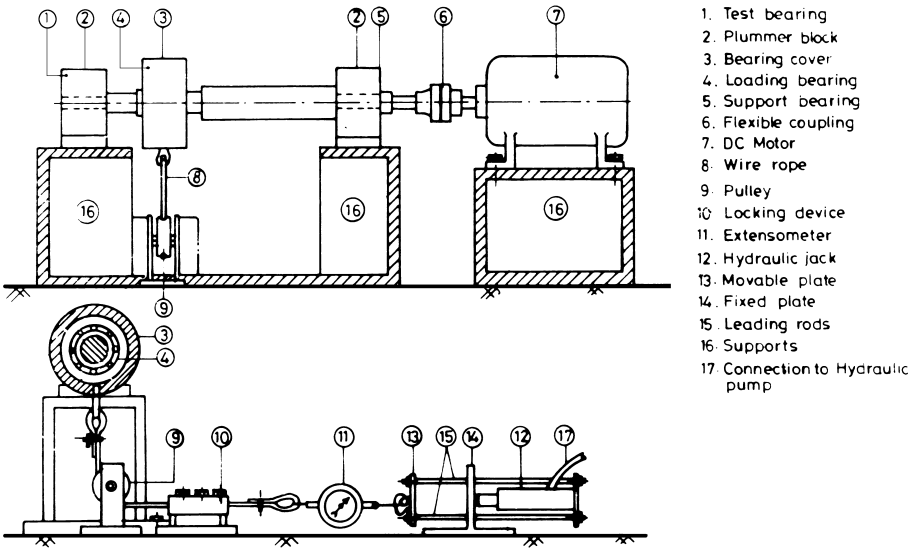

Fig. 4 Schematic of bearing test rig and loading arrangement

Table 2 Specification of test bearing

Test bearing	
Type	SKF6307
Number of balls	8
Ball diameter	13.2 mm
Pitch diameter	57.5 mm
Contact angle	0°

flaw. Generally this region includes one to seven times the element passage rate (the rate at which the rolling elements pass a point on either the inner or outer race of the bearing). Field studies indicate that approximately 90% of rolling element bearing failures is related to either an inner or outer race flaw. The other 10% is related to either a rolling element flaw or cage flaw. For the problem under study, the prime spike region is given as the outer race defect frequency to seven times inner race defect frequency. This has been calculated to cover the frequency range 3–142 Hz and is taken as 0–200 Hz for the sake of analysis.

(3) High Frequency Region (HFR): The third vibration frequency region, to be monitored for machines with rolling element bearings, is the high frequency region. When a flaw develops in a rolling element bearing, the vibration signals generated are in the form of short, sharp impulses. Since an accelerometer is a lightly damped device, it will respond to this type of input by ringing at its resonance frequency. Utilizing the accelerometer mounted resonance and measuring its amplitude, in units of acceleration, it is possible to monitor rolling element bearing condition in HFR. However, due to noise susceptibility problems and the possibility of 'self-peening' of the bearing flaws, which may result in decreasing readings as the bearing damage progresses, high frequency measurements for bearing failure detection should only be used as a supplement to measurements made in rotor vibration region and PSR. The predominant vibration frequencies in the high frequency region are the resonant frequencies of the races modulated by the characteristic defect frequencies and their harmonics. For the present low speed bearing, the high frequency range has been taken as 200 Hz to 10 kHz.

Processing of vibration signals

Typical time domain and frequency domain plots are taken from the three bearings at different speeds and loads. Figs 5–7 show the time domain plots of vibration acceleration of a new bearing, bearing with a ball defect and bearing with an outer race defect, respectively. Similarly, the spectra for these three bearings are shown in Figs 8–10, respectively.

The recorded signals are analyzed in the spectral mode of the analyzer and from the frequency response, in both PSR and HFR, parameters [(1)–(6)] mentioned below have been extracted. The rotor vibration region has not been considered separately as steps have been

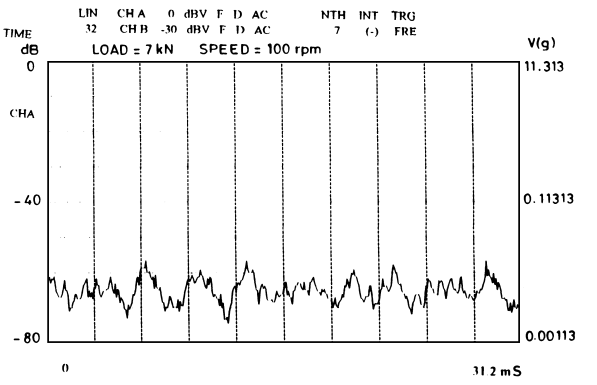

Fig. 5 Time domain plot of vibration acceleration (new bearing)

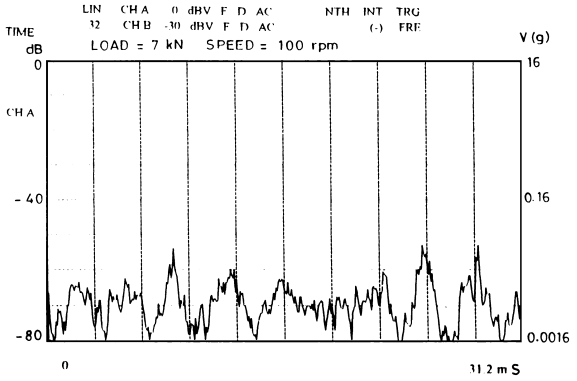

Fig. 6 Time domain plot of vibration acceleration (bearing with a ball defect)

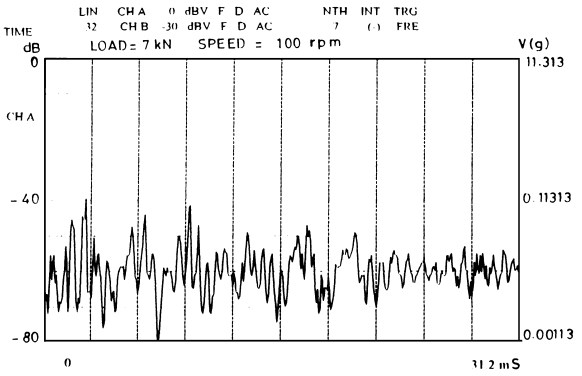

Fig. 7 Time domain plot of vibration acceleration (bearing with an outer race defect)

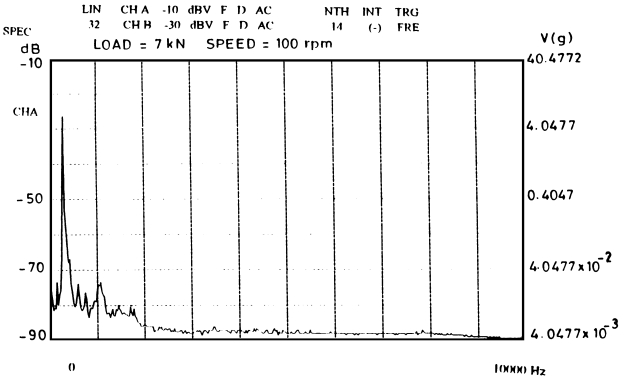

Fig. 8 Frequency domain plot of vibration acceleration (new bearing)

taken to eliminate rotor related malfunctions. In addition, the signals are analyzed in the statistical mode of the analyzer and the statistical parameters, i.e. standard deviation and kurtosis [(7) and (8)] are obtained.

Thus, the various bearing condition descriptors used are:

(1) Peak value of amplitude in PSR: This is a good indicator of bearing condition since the peaks in the PSR correspond to the characteristic defect frequencies[17,20].

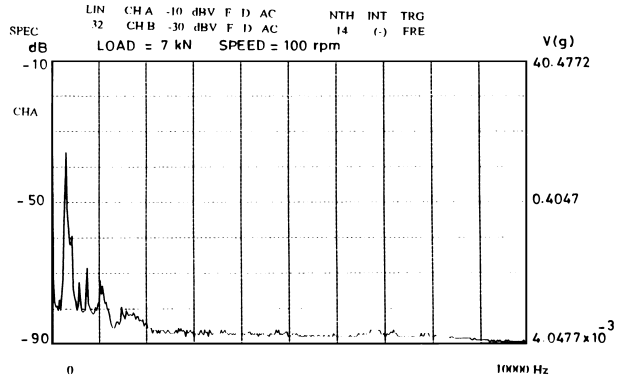

Fig. 9 Frequency domain plot of vibration acceleration (bearing with a ball defect)

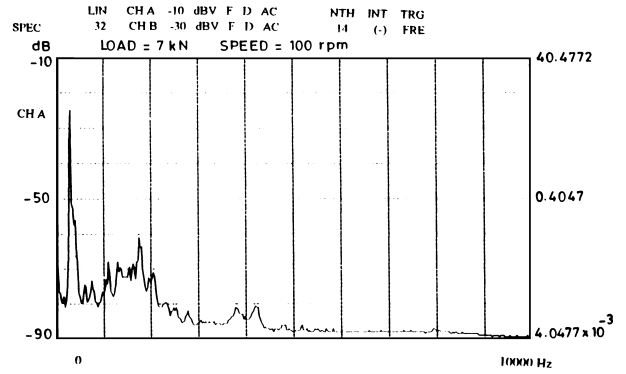

Fig. 10 Frequency domain plot of vibration acceleration (bearing with an outer race defect)

(2) Average of top five peak values of amplitude in PSR: This provides a better indication than the peak value since it gives an indication of the averaged spectral heights corresponding to various defect frequencies[17].

(3) Peak value of autocorrelation function in PSR: In the prime spike region, the bearing vibration signal may be treated as a narrow band (0–200 Hz) random signal. Hence, the autocorrelation function has a sharp spike at $\tau = 0$, that does not die off very rapidly with $\pm \tau$, as shown in Fig 11. This

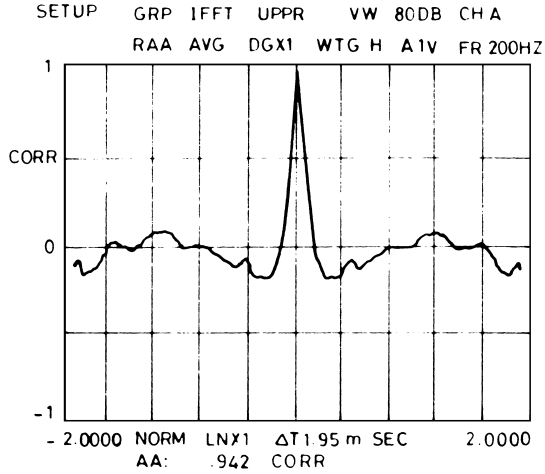

Fig. 11 Autocorrelation in prime spike region

peak value is nothing but the mean square value of the signal[3,17].

(4) Peak value of amplitude in HFR: With degradation in the bearing condition, vibrations in the high frequency region are seen to go up because of the increase in amplitudes of resonances of the races and the amplitudes of the modulating defect frequencies. Therefore, many researchers have used the peak value of frequency response amplitude in HFR to detect faulty bearings[17,20]. This value corresponds to the predominant resonance frequency.

(5) Average of top five peak values of amplitude in HFR: Instead of relying on a single peak (as above), which is indicative of one predominant resonance frequency, by considering more peaks, resonances of other components of the bearing are also brought into the picture along with the modulating defect frequencies. Hence, the average of top five peaks in the HFR is probably a better indicator of bearing condition than the predominant peak value[17,20].

(6) Peak value of autocorrelation in HFR: The peak value of autocorrelation function in the HFR (Fig 12) is nothing but the mean square value of the bearing vibration signal in the 0–10 kHz range and is highly indicative of the condition of the bearing[17].

(7) Standard deviation (0–10 kHz): This is a measure of the effective energy or power content of the vibration signal and clearly indicates deterioration in the bearing condition.

(8) Kurtosis (0–10 kHz): Kurtosis is related to the shape of the probability density distribution and is indicative of flatness or spikiness of the signal being considered. Kurtosis is a unitless quantity, which emphasizes the tails of the probability density histogram and is 3 for a normal Gaussian noise. Thus, for a healthy bearing the kurtosis is around 3 and for a defective bearing it is always high due to the spiky nature of the signal. Kurtosis is highly sensitive to bearing damage and is independent of shaft and bearing dimensions[21].

The above mentioned features are chosen because of their successful application in previous works[13,17,22,23] related to bearing damage detection. The use of these parameters is also justified from the observations made in the present investigation.

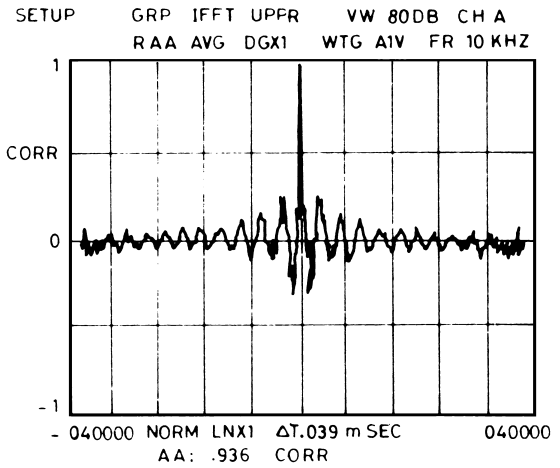

Fig. 12 Autocorrelation in high frequency region

Table 3 shows the variation in these eight input parameters for the new and defective bearings for various speeds and for a load of 5 kN.

Problems associated with low speed bearing condition monitoring

Defects in rotating rolling element bearings generally produce impacts when the mating parts of the bearing come in contact with one another. These impacts occur regularly and the frequencies of occurrence correspond directly to the ball or roller pass frequencies. These characteristic defect frequencies as they are known, can usually be easily identified when monitoring the vibrations of rolling element bearings operating at high speeds. Various techniques have been proposed and evaluated by Alfredson and Mathew[24,25]. It was shown that spectrum analysis of bearing cap acceleration was usually sufficient to detect and diagnose damage of high speed bearings. At low speeds ($\leq$ 250 rpm), however, these defect frequencies are sometimes difficult to detect, usually due to the presence of dominant background noise and the lack of sensitivity of acceleration measurements at low frequencies[26].

Besides, due to inherent performance limitations of the FFT approach, i.e. limited frequency resolution and masking of weak signal spectral responses by stronger spectral responses, the spectral analysis method of monitoring does not offer very reliable diagnosis. Also, low speed rolling element bearings are not considered for monitoring in most industrial situations because of the long sampling and data analysis times required. When considering vibration data in the low frequency bands, long records of signals are required in order to achieve repeatable and trustworthy results. Up to 10 min of vibration signal may be required if a considerable amount of noise is present in the signal[27].

Despite these problems associated with low speed rolling element bearings, neural networks seem to play a very promising role for fault detection and isolation, by making use of multiple features extracted from a relatively short vibration signal (about 10 s), though the signal is dominated by background noise. Hence, low speed bearings have purposely been chosen for this study, to prove that, in spite of all the above mentioned difficulties in monitoring vibration signals of these bearings, neural networks are able to predict their condition with very high success rates by making use of multiple parameters.

Results and discussion

Results of EBP based neural network

The neural network has to be trained *a priori* to learn the complex input/output association for bearing condition recognition. Here, the training process has been divided into two cases. In the first case, effort has been concentrated on distinguishing a defective ball bearing from a normal one. In the second case,

Table 3 A typical input data set

Speed (rpm)	1×10^{-4}	2×10^{-4}	3×10^{-4}	4×10^{-4}	5	6	7	8×10^{-4}
New bearing								
80	2.15	2.00	2.25	2.13	0.92	0.92	1.40	81.40
120	2.22	3.22	4.67	3.68	0.92	0.93	1.96	83.60
160	2.92	2.59	5.15	3.74	0.93	0.91	2.32	87.60
200	3.22	2.57	4.61	4.41	0.94	0.92	3.32	88.60
Bearing with a ball defect								
80	2.61	1.64	3.97	3.27	0.96	0.91	4.01	90.10
120	1.23	0.89	7.59	5.57	0.93	0.82	4.34	90.70
160	2.95	2.01	10.04	7.66	0.96	0.87	4.40	91.80
200	2.00	1.68	12.60	9.92	0.95	0.80	5.21	94.70
Bearing with an outer race defect								
80	7.42	3.66	6.05	5.48	0.95	0.94	2.21	84.70
200	9.53	4.62	7.19	5.88	0.93	0.95	3.82	86.30
160	9.84	4.48	8.72	7.59	0.93	0.93	5.80	99.10
200	10.02	5.41	9.77	8.45	0.93	0.94	7.38	102.10

1: Peak value of vibration acceleration frequency response in PSR (g)
2: Average of top five peaks in PSR (g)
3: Peak value in HFR (g)
4: Average of top five peaks in HFR (g)
5: Peak value of autocorrelation in PSR (non-dimensional)
6: Peak value of autocorrelation in HFR (non-dimensional)
7: Kurtosis (non-dimensional)
8: Standard deviation (g)

the training process aims at identifying various ball bearing states (normal, ball defective and outer race defective). Training involves determination of the network architecture (number of nodes in each layer) and the network training parameters: Learning rate (η), momentum parameter (α) and accelerating constant (β), which are problem dependent.

Case 1

In the first case, the neural network is trained using 94 sets of normalized input/output data. A typical set of normalized input data, normalized between 0.1 and 0.9, corresponding to Table 3 is shown in Table 4. Each data set consists of the eight input parameters and the target output of the network, which is normalized to 0.1 for normal bearing data, 0.6 for data taken from a bearing with a ball defect and 0.9 for data corresponding to a bearing with an outer race defect. Initially, weights are assigned at random (between 0 and 0.5) and are continuously updated as per EBP training algorithm. To obtain optimal network architecture and the training parameters, training has been stopped after completion of 2000 training iterations, to save computational time, as the training convergence behavior remains the same after this.

(1) Learning rate (η): The effectiveness of convergence of training depends significantly on the

Table 4 Typical normalized input data set

Speed (rpm)	1	2	3	4	5	6	7	8
New bearing								
80	0.18	0.29	0.10	0.10	0.10	0.74	0.10	0.10
120	0.19	0.51	0.28	0.25	0.10	0.79	0.17	0.18
160	0.25	0.40	0.32	0.26	0.30	0.68	0.22	0.33
200	0.28	0.39	0.28	0.33	0.50	0.74	0.35	0.37
Bearing with a ball defect								
80	0.22	0.23	0.23	0.21	0.90	0.68	0.44	0.43
120	0.10	0.10	0.51	0.45	0.30	0.20	0.49	0.45
160	0.25	0.29	0.70	0.66	0.90	0.47	0.50	0.50
200	0.17	0.23	0.90	0.90	0.70	0.10	0.60	0.61
Bearing with an outer race defect								
80	0.66	0.59	0.39	0.44	0.70	0.84	0.20	0.22
120	0.85	0.76	0.48	0.48	0.30	0.90	0.42	0.28
160	0.88	0.73	0.60	0.66	0.30	0.79	0.68	0.84
200	0.90	0.90	0.68	0.74	0.30	0.84	0.90	0.90

value of η, the optimal value of which is problem dependent. As the numbers of input and output nodes are fixed, eight and one, respectively, in this case, for eight hidden nodes, chosen arbitrarily initially, the learning rate (η) is varied as shown in Fig 13(a) and when $\eta = 0.375$, the convergence has been found to be faster and better without leading to oscillations.

(2) Hidden nodes: The size of the hidden nodes is one of the most important considerations during training. The function of these nodes is to capture the complex association between inputs and outputs, instead of mapping them directly. The number of hidden nodes is varied as shown in Fig 13(b). For the chosen $\eta = 0.375$, 16 hidden nodes are found to be sufficient to achieve the required classification.

(3) Momentum parameter (α): The purpose of α is to accelerate the training process, which involves supplementing the current weight adjustments with a small fraction of the most recent weight adjustments. The value of α is varied as shown in Fig 13(c) and $\alpha = 0.2$ is found to be optimal in conjunction with $\eta = 0.375$ and the architecture $8 - 16 - 1$ (no. of input nodes $-$ no. of hidden nodes $-$ no. of output nodes).

(4) Accelerating constant (β): To further accelerate the training process, an accelerating term has been introduced and the effect of this term is shown in Fig 13(d). For the selected optimal architecture ($8 - 16 - 1$) and the training parameters ($\eta = 0.375$, $\alpha = 0.2$), at $\beta = 1.0$ much faster convergence is achieved and hence $\beta = 1.0$ has been taken as the optimal value for the final learning.

For the finally obtained optimal architecture ($8 - 16 - 1$) and the training parameters ($\eta = 0.375$, $\alpha = 0.2$, $\beta = 1.0$), the training has been continued and the final learning curve is shown in Fig 14. After completion of training, the weights and thresholds along with the network architecture are stored to judge the network's classification ability to recognize the condition of the bearing.

The ability of the fully trained network to distinguish a defective bearing from a normal bearing is tested using 30 fresh test patterns (which have been obtained from interpolation of the original data). The ball bearing states are categorized into two classes, based on the network predicted output values: Normal bearing (0.100–0.200) and defective bearing (0.201–0.900). It has been noticed that the network clearly distinguished a defective bearing from a normal bearing with cent per cent accuracy, as seen from the first row of Table 5.

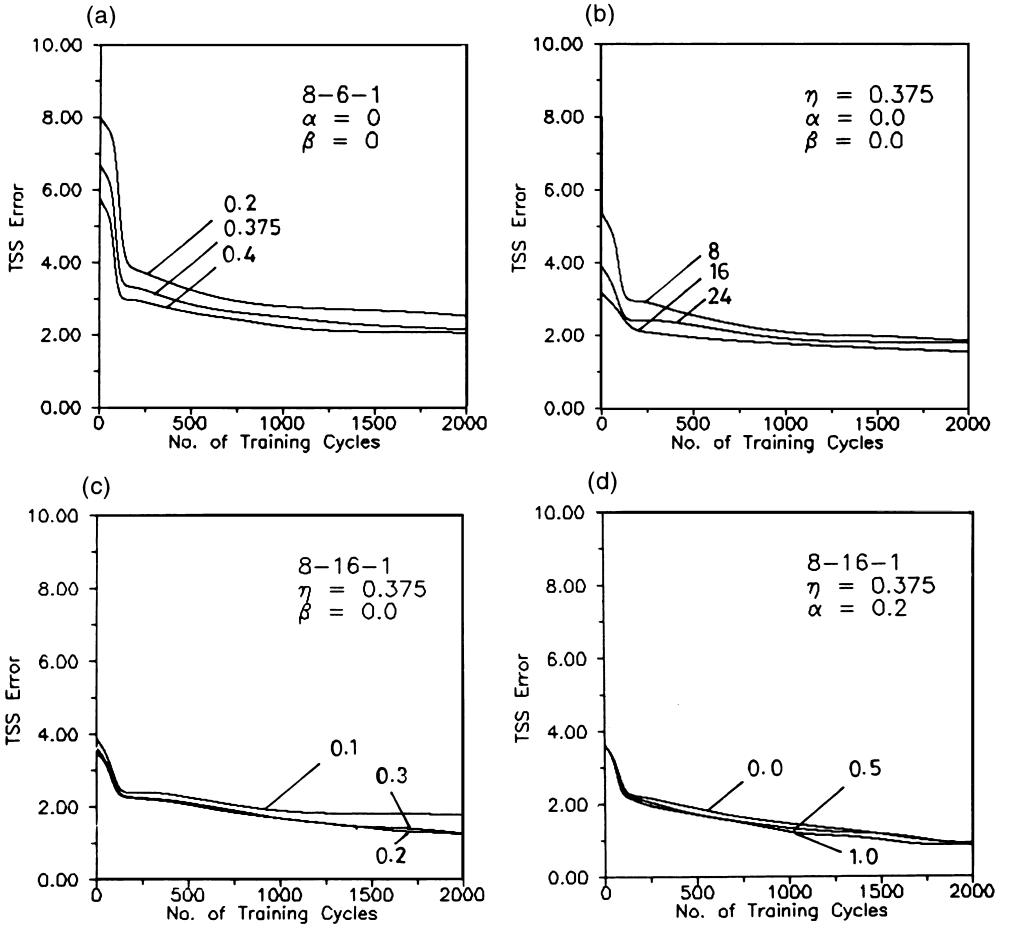

Fig. 13 Effect of training parameters on convergence. (a) Effect of learning rate (η). (b) Effect of hidden nodes. (c) Effect of momentum (α). (d) Effect of accl. const. (β)

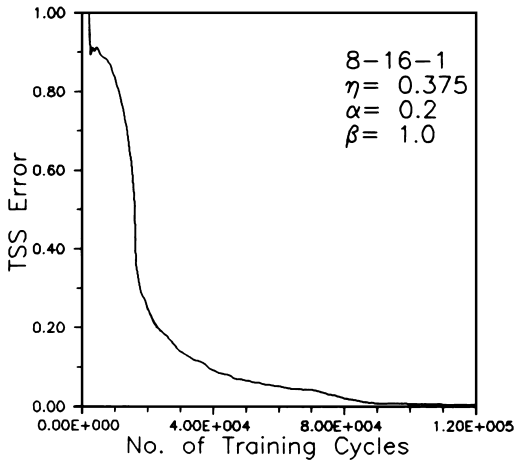

Fig. 14 Learning curve (case 1)

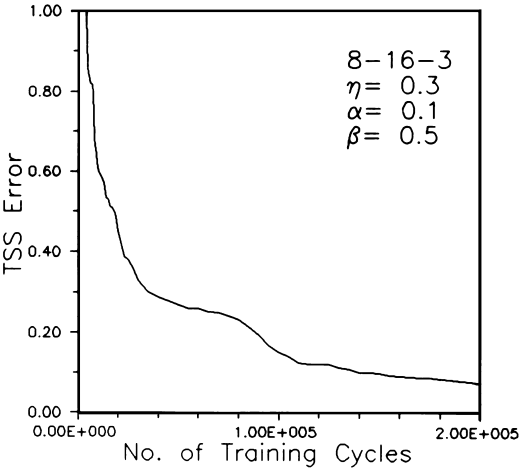

Fig. 15 Learning curve (case 2)

Table 5 Classification performance of the neural network in deciding the component condition

Problem case	Architecture (EBP)/vigilance threshold (ART)	No. of test patterns	No. of correct classifications	Does the network estimate the component condition correctly?
1. Bearing (EBP)	8 − 16 − 1 (η = 0.375,α = 0.2, β = 1.0)	30	30	Yes
2. Bearing (ART2)	0.8	30	30	Yes

Case 2

In the second case, the network is also trained with the same 94 sets of data, but the number of output nodes is taken as three and the target outputs are assigned as follows.

New bearing (NB) [0.90, 0.01, 0.01]

Bearing with a ball defect (BD) [0.01, 0.90, 0.01]

Bearing with an outer race defect (OD) [0.01, 0.01, 0.90].

The training process is repeated as discussed earlier to ascertain optimal architecture and other training parameters, and the learning curves are typical of those given in Fig 13. For the finally chosen optimal architecture (8 − 16 − 3) and the training parameters (η = 0.3, α = 0.1, β = 0.5), the learning process has been continued, as discussed earlier and the final learning curve is shown in Fig 15. The classifying ability of the fully trained network to recognize various ball bearing states, has been tested on 12 fresh test patterns and the results are shown in Table 6. From the table it can be seen that the network correctly estimated the bearing states on all test patterns.

ART2 results

The same normalized input data samples, as used in EBP learning, without target outputs, are presented to the ART2 type neural network for clustering the input samples for condition monitoring of the bearing. The clustering process (training) involves the selection of vigilance threshold (ρ) and optimal number of reclusterings (with the same inputs), and also considers the effect of order of presentation of the inputs.

Training

The normalized data were presented to the network (94 data sets, each set consisting of eight parameters) and the training process has been done as described below.

(1) Vigilance threshold (ρ): Vigilance threshold controls the resolution of the classification process. A low choice of ρ produces a high resolution of the classification process, creating larger class types (clusters). A high ρ produces a smaller classification, creating fewer class types. The effect of ρ is shown in Table 7. At ρ = 0.7, it has been observed that the ART2 network clustered all 94 input patterns into five groups.

(2) Reclustering: Once all input patterns are clustered based on the ART2 learning algorithm, some input patterns may be actually closer to the centroids (weights) of other clusters. Hence, reclustering has been done, using current centroids obtained through initial clustering, as initial reference weights for reclustering. This has been repeated until no change of clustering during one entire sweep occurs, i.e. there is no jumping of patterns from one cluster to other clusters. The effect of reclustering on the number of clusters

Table 6 ANN classification performance (EBP)

Test pattern	Actual class	EBP classification		
		NB	BD	OD
1	NB	NB	—	—
2	NB	NB	—	—
3	NB	NB	—	—
4	NB	NB	—	—
5	BD	—	BD	—
6	BD	—	BD	—
7	BD	—	BD	—
8	BD	—	BD	—
9	OD	—	—	OD
10	OD	—	—	OD
11	OD	—	—	OD
12	OD	—	—	OD

Table 7 Effect of vigilance threshold (ρ) on number of clusters

ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	1.0	1.1
No. of clusters	89	62	33	23	15	11	5	3	2	1

Table 8 Effect of reclustering on number of clusters

ρ	No. of reclusterings					
	0	1	2	3	4	5
0.6	11	11	13	14	14	14
0.7	5	5	6	6	6	6
0.8	3	3	6	6	6	6

Table 9 Effect of order of presentation of input patterns on number of clusters

ρ	Normal order with five reclusterings	Reverse order with five reclusterings
0.6	14	10
0.7	6	3
0.8	6	2

is shown in Table 8 (though the number of clusters is the same with three or four reclusterings, the number of patterns in each cluster is different). It has been noticed that after five reclusterings, no shifting of patterns from one cluster to another is found and hence five is taken as the optimal number of reclusterings.

(3) Order of presentation of input patterns: ART2 is sensitive to the order of presentation of inputs and yields a different clustering for the same input patterns. To overcome this, the input data have been presented in the reverse order and reclustering has been allowed five times. The effect of this is shown in Table 9.

After completion of training, the weights (updated centroidal coordinates of each cluster) along with the optimal vigilance threshold (ρ) and network architecture are stored for testing purposes.

Testing

The classification performance of the trained network has been tested on 12 fresh test patterns (same as those used in EBP training) and the results are summarized below.

(1) At $\rho = 0.8$, with five reclusterings, ART2 created two clusters on reverse order presentation of

Table 10 ANN classification performance (ART2)

Test Pattern	Actual Class	ART2 Classification		
		NB	BD	OD
1	NB	NB	—	—
2	NB	NB	—	—
3	NB	NB	—	—
4	NB	NB	—	—
5	BD	—	BD	—
6	BD	—	BD	—
7	BD	—	BD	—
8	BD	—	BD	—
9	OD	NB	—	—
10	OD	—	—	OD
11	OD	NB	—	—
12	OD	—	—	OD

Table 11 Neural network training time comparison

Problem	Optimal architecture	EBP training time (CPU time in min)	ART2 training time (CPU time in min)
1. Bearing condition estimation (case 1)	8 − 16 − 1 (η = 0.375, α = 0.2, β = 1.0)	600	8
2. Bearing condition estimation (case 2)	8 − 16 − 3 (η = 0.3, α = 0.1, β = 0.5)	840	8

inputs, clearly separating defective bearing data from normal bearing data. Testing has been carried out on 30 fresh test patterns and the results are shown in the second row of Table 5. From the table, it can be seen that ART2 network's classification is exceptionally good.

(2) At ρ = 0.7, with five reclusterings, ART2 created three clusters with reverse order presentation of inputs. Testing has been done on 12 fresh test patterns, and the ART2 classification is shown in Table 10. The results show that ART2 is not so effective in isolating bearing defects as compared to EBP network, but the utility of the model has been overwhelmingly satisfying in detecting whether the bearing is defective or not.

Comparison of training time

Table 11 shows a comparison of training time for both the network models for the chosen problem for both the cases. The training time mentioned is the CPU time on a PC386 machine at 40 MHz with a coprocessor at 33 MHz. Referring to the table, it can be noticed that the ART2 network is about 100 times faster than the EBP neural network in the training mode (but the classification performance achieved with the latter is much better than that achieved with the former). Testing can be implemented on both the network models in a fraction of a second.

Conclusions

Two neural network approaches, based on the supervised Error Back Propagation (EBP) learning algorithm and the unsupervised Adaptive Resonance Theory-2 (ART2) based training paradigm, have been developed for bearing condition recognition. The results of these studies show the potential suitability of these approaches for the chosen application, for use in industry.

(1) The performance of the error back propagation neural network in recognizing ball bearing states has been found to be exceptionally good. Using the proposed neural network, any defective ball bearing can be distinguished from a normal one with cent per cent reliability. Moreover, the network has been capable of estimating the three different ball bearing states (having different localized defects) for diagnostic purposes with a success rate of over 95%.

(2) The results of bearing condition monitoring using ART2 neural network lead to the following conclusions.

The ART2 neural network has been found to be extremely fast, about 100 times faster than EBP learning.

A faulty bearing can be distinguished from a normal bearing with 100% reliability.

The recognition of different ball bearing states for diagnostic purposes is not so effective as compared to the classification achieved through EBP network, but the performance of the ART2 neural network is overwhelmingly satisfying in distinguishing a normal bearing from defective bearings.

References

1. Wang, S.S., Artificial intelligence and expert systems for diagnostics. *Proc. 1st Int. Conf. Machinery and Diagnostics and Exhibition*, Las Vegas, Nevada, 1989, pp. 516–512.
2. Widner, R.L. and Littmann, W.E. (eds.) *Bearing Damage Analysis*, National Bureau of Standard Publication, April 1976.
3. James Li, C. and Wu, S.M., On-line detection of localised defects in bearings by pattern recognition analysis. *ASME J. Engng. Ind.*, 1989, **111**, 331–336.
4. Shi, X.Z., Xu, Z.Q. and Xu, M., A study on the automatic recognition of vibration signal for ball bearing faults—the FFT-AR feature extraction and classification methods. *Proc. IEEE Int. Workshop on Applied Time Series Analysis*, World Scientific, 1988, pp. 149–154.
5. Wang, X.F., Shi, X.Z. and Xu, M., The fault diagnosis and quality evaluation of ball bearing by vibration signal processing. *Proc. 1st Int. Machinery Monitoring and Diagnostics Conference*, 1989, pp. 318–321.
6. Zurada, J.M., *Introduction to Artificial Neural Systems*, Jaico Publishing House, Bombay, 1995.
7. Rumelhart, D. and McClelland, J., *Parallel Distributed Processing*, Vol. 1. M.I.T. Press, Cambridge, MA, 1986.
8. Smith, M., *Neural Networks for Statistical Modelling*, Van Nostrand Reinhold, New York, 1993.
9. Samad, T., Neural network with expected source values. *Neural Networks*, 1991, **4**, 615–618.
10. Kung, S.Y., *Digital Neural Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
11. Carpenter, G.A. and Grossberg, S., *Pattern Recognition by Self-organising Neural Network*, M.I.T. Press, Cambridge, MA, 1991.
12. Dyer, D. and Stewart, R.M., Detection of rolling element bearing damage by statistical vibration analysis. *ASME J. Mechanical Design*, 1978, **100**, 229–235.
13. Bentley, D., *Predictive Maintenance through the Monitoring and Diagnostics of Rolling Element Bearings*, Bentley Nevada Corporation, Minden, Nevada, 1989.
14. Braun, S. and Datner, B., Analysis of roller/ball bearing vibrations. *ASME J. Mechanical Design*, 1979, **101**, 118–125.
15. Gustafssan, O.G. and Tallun, T., Detection of damage in assembled rolling bearings. *ASLE Transactions*, 1979, **5**, 197–209.
16. Igarashi, T. and Kato, J., Studies on the vibration and sound of defective rolling bearings. *Bulletin JSME*, 1985, **28**, 492–499.
17. Liu, T.I. and Iyer, N.R., On-line recognition of roller bearing states. *Japan/USA Symp. on Flexible Automation*, 1992, **1**, 257–262.
18. Unal, A., Intelligent diagnostics of ball bearings. Feature Article, Nov./Dec., *Shock and Vibration Digest*, 1994, 9–12.
19. Chiou, Y.S., Tavakoli, M.S. and Liang, S., *Bearing Fault Detection Based on Multiple Signal Features using Neural Network Analysis. Proc. 10th Int, Modal Analysis Conf.*, Vol. 1, San Diego, CA, 1992, pp. 60–64.
20. McFadden, P.D. and Smith, J.D., Vibration monitoring of rolling element bearings by high frequency resonance technique—A review. *Tribol. Int.*, 1984, **17**, 3–10.
21. Technical Bulletin on Kurtosis Meter, Condition Monitoring Ltd., UK.
22. Prashad, H., Ghosh, M. and Biswas, S., Diagnostic monitoring of rolling element bearing by high frequency resonance technique. *ASLE Transactions*, 1985, **28**, 439–448.
23. Bentley, D., Rolling Element Bearing Activity Monitor, Bentley Nevada Corporation, Minden, Nevada, 1989.
24. Alfredson, R.J. and Mathew, J., Time domain methods for monitoring the condition of rolling element bearings. *Transactions of the Institution of Engineers, Australia*, 1985, **2**, 102–107.
25. Alfredson, R.J. and Mathew, J., Frequency domain methods for monitoring the condition of rolling element bearings. *Transactions of the Institution of Engineers, Australia*, 1985, **2**, 108–112.
26. Mathew, J., Szezepanik, A., Kuhnell, B.T. and Stecki, J.S., Incipient fault detection in low speed bearings using demodulated resonance analysis technique. *Int. Tribol. Conf.*, Melbourne, 1987, pp. 366–369.
27. Mechefske, C.K. and Mathew, J., Fault detection and diagnosis in low speed rolling element bearings, Part-1: The use of parametric spectra. *Mechanical Systems and Signal Processing*, 1992, **6**, 297–307.