Artif Intell Rev (2007) 28:1–15
DOI 10.1007/s10462-008-9081-6

Using support vector machines and acoustic noise signal for degradation analysis of rotating machinery

Patricia Scanlon · Susan Bergin

Published online: 20 September 2008

Abstract An automated approach to degradation analysis is proposed that uses a rotating machine's acoustic signal to determine Remaining Useful Life (RUL). High resolution spectral features are extracted from the acoustic data collected over the entire lifetime of the machine. A novel approach to the computation of Mutual Information based Feature Subset Selection is applied, to remove redundant and irrelevant features, that does not require class label boundaries of the dataset or spectral locations of developing defect to be known or pre-estimated. Using subsets of the feature space, multi-class linear and Radial Basis Function (RBF) Support Vector Machine (SVM) classifiers are developed and a comparison of their performance is provided. Performance of all classifiers is found to be very high, 85 to 98%, with RBF SVMs outperforming linear SVMs when a smaller number of features are used. As larger numbers of features are used for classification, the problem space becomes more linearly separable and the linear SVMs are shown to have comparable performance. A detailed analysis of the misclassifications is provided and an approach to better understand and interpret costly misclassifications is discussed. While defining class label boundaries using an automated k-means clustering algorithm improves performance with an accuracy of approximately 99%, further analysis shows that in 88% of all misclassifications the actual class of failure had the next highest probability of occurring. Thus, a system that incorporates probability distributions as a measure of confidence for the predicted RUL would provide additional valuable information for scheduling preventative maintenance.

Keywords Support vector machines · Degradation analysis · Acoustic signal processing · Feature selection

This work was supported by IDA Ireland.

P. Scanlon (✉)
Bell Labs Ireland - Alcatel Lucent, Dublin 15, Ireland
e-mail: scanlon@alcatel-lucent.com

S. Bergin
Department of Computer Science, NUI Maynooth, Maynooth, Ireland
e-mail: sbergin@cs.may.ie

1 Introduction

Automated monitoring of machines typically involves the detection and diagnosis of defects. However in industry, there is an increasing demand for machine reliability and optimal management of spare parts inventory to prevent machine downtime. This demand requires such machine monitoring systems to also predict the Remaining Useful Life (RUL) of the machine in order to schedule materials, logistics and maintenance.

Machine monitoring systems reported in the literature are predominantly focused on defects developing in the components of Rolling Element Bearings (REBs). Detecting and diagnosing defects in REBs by manual visual inspection of the vibration or acoustic measurements has been reported in the literature (Tandon and Nakra 1990; Heng and Nor 1998; Shi et al. 2004). However, such manual approaches are inefficient, require expert training and are subjective.

Automated approaches to machine monitoring have been the focus of much research in recent times. These approaches can be broadly divided into detection, diagnosis and degradation analysis. Defect detection is a dual case problem where the classifier determines whether a defect exists or not. Defect diagnosis is a multi-class problem which attempts to classify which type of defect exists (Goumas et al. 2002; Samanta et al. 2003). Degradation analysis is also approached as a multi-class problem where developing defects are classified to several 'wear states' in order to determine the RUL of the machine (Lao and Zein-Sabatto 2001).

Artificial neural networks (ANNs) have been applied in automated detection and diagnosis of machine faults (Samanta et al. 2003) and degradation analysis for determining a bearings RUL (Lao and Zein-Sabatto 2001). In (Goumas et al. 2002) Euclidean, Mahalanobis, and Bayesian distance classifiers, learning vector quantization (LVQ) classifier and the fuzzy gradient classifier are used for classification of various defects in washing machines vibration signals. Recent work in machine fault detection has employed the use of SVMs with RBF kernel (Samanta et al. 2003; Rojas and Nandi 2005). In this work the task of determining the RUL of rotating machinery is approached using SVM classifiers using both linear and RBF kernels.

Some early studies used sound pressure signals to explain the mechanism of vibration and noise generation in bearings (Jayaram and Jarchow 1978; Tandon and Choudhury 1999). Several studies have shown that sound intensity, sound pressure and vibrational data measurements provide enough information to manually differentiate between 'good' and 'bad' bearings using spectral analysis (Tandon and Nakra 1990) and statistical analysis (Heng and Nor 1998). An in-depth review is given in (Tandon and Choudhury 1999) on the application of vibration measurements to both manual and automated machine monitoring systems as well as studies which examine acoustic measurements for manual defect detection. While vibrational signal measurements have been used in automated approaches to machine monitoring, to the best of our knowledge, no automated approach to defect detection, diagnosis or degradation analysis using acoustic measurements has been reported in the literature. Note that while vibrational analysis requires contact with the machine being monitored, acoustic analysis is advantageous as it allows for remote monitoring.

Various signal analysis techniques have been used in condition monitoring on both acoustic and vibrational data. These can be broadly classified into time domain techniques such as root-mean-square, crest factor, and statistical parameters such as mean, variance, skewness and kurtosis (Tandon and Choudhury 1999; Heng and Nor 1998) and frequency domain techniques such as spectral, Short Time Fourier Transform (STFT), Wavelet Transform (WT) and envelope detection (McFadden and Smith 1984; Shi et al. 2004; Tandon and Choudhury 1999).

As spectral feature extraction results in irrelevant, noisy parts of the spectrum being included in the feature vector, information theoretic based feature subset selection (FSS) technique is employed to remove noisy features and select a compact and relevant feature set for classification. While Mutual Information (MI) or relevance has been previously used for FSS (Bell and Wang 2000; Scanlon et al. 2007a), a novel approach to the computation of Mutual Information (MI) for each candidate spectral feature is proposed. Once MI is computed the features are ranked according to MI values and the top n features are retained and used as input to the classifier. This approach does not require a-priori information regarding class label boundaries or the spectral location of potential defects but utilizes only the chronological order of collected data samples to select the relevant features. This feature subset is used as input to multi-class linear and RBF SVM classifiers and a comparison of their performance is provided.

The studies on machine monitoring reported in the literature describe interesting approaches to this problem. However, the focus of the research described in this paper is an automated approach to machine monitoring using acoustic noise measurements to determine the RUL, which has not been addressed before in the literature. This study focuses on three key aspects of this novel system: first the novel approach to the selection of relevant spectral features extracted from the machine's acoustic noise signal and second, the application of SVM classifiers with linear and RBF kernels to the multi-class classification problem of determining the RUL. Another key aspect of the system is the labeling of acoustic data collected over the lifetime of the machine, using automated k-means clustering to define class boundaries.

The following section describes the characteristics of bearing noise and how degradation might affect the noise is described as well as the feature extraction and FSS. Section 3 describes the SVM classifiers used in the experiments and Sect. 4 describes two different approaches to data labeling. Section 5 describes the experimental setup and implementation and Sect. 6 describes and discusses the results of the experiments.

2 Acoustic signal processing

Analysis can be defined as the process of decomposing something complex into simpler, more basic parts. It is useful to note that some signals are easier to interpret (and take less information to define) in the frequency domain than in the time domain and vice versa. In signal processing Fourier Analysis is typically thought of as decomposition of a signal into its composite frequency (sine and cosine) components. Such analysis can be used to isolate individual components of a complex signal, concentrating them for easier detection and/or removal. Oscillating and periodic signals, such as speech, electrocardiograms, surface vibrations or the acoustic noise signal emitted from a rotating machine used in this study, are examples of signals that are easier to interpret in the frequency domain.

The Fourier transform works on an infinite length continuous signal to produce a continuous spectrum, where the spectrum is the set of sine and cosine magnitudes at different frequencies. Since computers cannot work with continuous or infinitely long signals for obvious reasons, typically an approximation to the Fourier transform known as the Discrete Fourier Transform (DFT), is used instead (Proakis and Manolakis 1996). The DFT works on a finite length sampled signal and produces a Fourier spectrum with values at a finite number of discrete frequencies. The DFT is widely employed in signal processing and related fields to analyze the frequencies contained in a sampled signal. The DFT and its inverse can be written as

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi nk/N} \quad \text{for } k = 0, 1, \ldots, N-1 \tag{1}$$

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k)\, e^{j2\pi nk/N} \quad \text{for } n = 0, 1, \ldots, N-1 \tag{2}$$

where $x(n)$ denotes the input signal at time (sample) n, and $X(k)$ denotes the kth spectral sample, which includes both real and imaginary components (Proakis and Manolakis 1996). The Fast Fourier transform (FFT) is an efficient algorithm used to compute the DFT and its inverse, reducing operations for an n-point DFT from $O(N^2)$ arithmetical operations to only $O(N \log(N))$.

The conversion from continuous time to samples (discrete-time) to perform DFT and analysis, generally entails a type of distortion called aliasing. Choice of an appropriate sample-rate based on the Nyquist theory is the key to minimizing that distortion, i.e. to correctly resolve a signal at some frequency, you have to sample it at more than twice that frequency (Proakis and Manolakis 1996). In this study the acoustic data was acquired at a sampling rate of 50,000 samples/s, which results in the DFT components spanning the spectrum from 0 to 25 kHz.

The sampled acoustic data was first split into windowed non-overlapping time frames. A window function is a function that is zero-valued outside of some chosen interval, in this way a finite sequence is extracted for transformation using FFT algorithm. Each time frame results in a set of spectral components which is an estimate of the short-term, time-localized frequency content of the acoustic signal. The duration of the window has a pronounced effect on the nature of the features. A short duration window will result in good time resolution but poor frequency resolution and the converse is also true.

In this study, the steady state nature of the acoustic noise signal emitted from the rotating machine is considered relevant for monitoring and the effects of any brief transient events or random noise are suppressed. Therefore, time resolution is less significant than frequency resolution for predicting RUL. Note also that the Fourier transform of a random waveform is also random. Therefore, spectral averaging can be used to remove the effects of random noise and transient events and create a clearer picture of the signals underlying frequency content. In this study, a sequence of data samples is extracted at regular intervals over the lifetime of the machine e.g. every 20 min. Each sequence is first divided into a number of shorter time-segments, e.g. 20 ms segments. These segments are frequency transformed and the magnitude of the frequency components of the transforms are averaged over approximately a hundred segments to remove the effect of unwanted noise and reduce random variance over short periods of time. The averaged power spectrum for data acquired from a new machine, machine midway through its life and a machine close to failure is illustrated in Fig. 1.

Spectral analysis for machine monitoring, decomposes the acoustic noise signal into frequencies so the influence of individual mechanical components can be ascertained as each type of fault has its own characteristic spectral signature or defect frequencies. In addition to these defect frequencies, their harmonics are also detectable. The harmonic of a wave is a component frequency of the signal that is an integer multiple of the fundamental frequency. The shape and resonance properties of a machine and housing cause harmonics to be boosted or dampened. Harmonics, whose frequencies are close to a resonance frequency of the machine or housing will be enhanced, and harmonics whose frequencies are not close to resonance frequencies become weakened. Due to the considerable broadband noise that occurs in the low frequency range, below 5 kHz, the fundamental defect frequency and low order harmonic

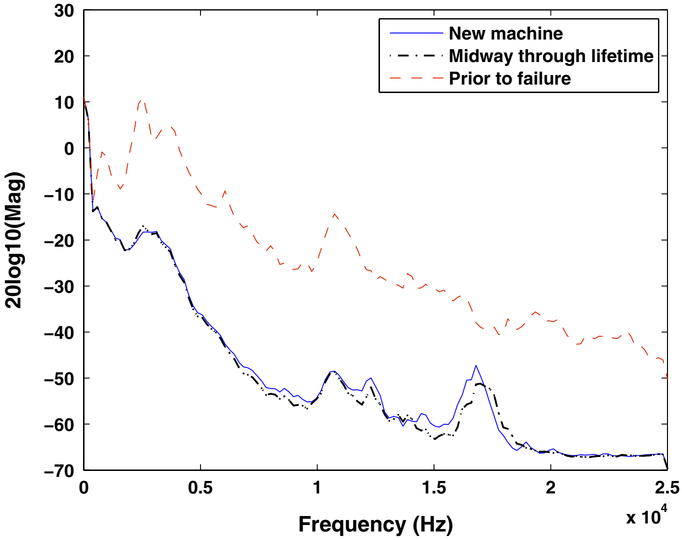

Fig. 1 The averaged power spectrum for data acquired from the new machine, the machine midway through its life and the machine close to failure

resonant frequencies may be ‘swamped’ and therefore not useful for monitoring. However, one or more of the higher order harmonic resonances may have a higher signal-to noise ratio which makes them easier to detect and monitor.

As defects develop, the amplitude at the location of the defect frequency increases. If the frequency resolution is too coarse, this may cause relevant defect frequency or resonances information to be hidden during the early stages of defect development. Increasing frequency resolution may uncover relevant information that would otherwise be hidden. Figure 2 illustrates stacked averaged power spectral feature vectors sampled periodically over the lifetime of the machine using different window lengths (frequency resolutions) i.e. using 32, 256, 1,024 and 2,048 point FFT. Earlier work by the authors, (Scanlon et al. 2007b), showed that using a 1,024 point FFT maximized performance accuracy for predicting the RUL of the rotating machine over the other frequency resolutions and is hence used in this study.

However, using such high resolution also increases dimensionality of the feature vector and a considerable amount of irrelevant noisy features will be included in the feature vector. Therefore, in order to effectively detect and monitor defects and yet eliminate irrelevant noisy features, a novel approach to FSS is proposed in the following section to extract a limited number of relevant features for degradation analysis in order to determine the RUL.

Envelope Analysis is a feature extraction technique widely reported on in the literature for machine monitoring (Ho and Randall 2000; Tandon and Choudhury 1999). This approach requires that the frequency band of interest, the cut-off frequencies, must be pre-determined. However, in normal operating conditions it is generally not known a-priori which defects will develop, whether multiple defects will develop and the severity of manufacturing variances. As a result the theoretical and actual defect frequencies will vary. The proposed FSS technique, described in the following section, has the additional benefit that it does not require a-priori knowledge as to the spectral location of defect frequencies, and therefore makes this approach highly suitable to the task of Rotating Machine Monitoring.

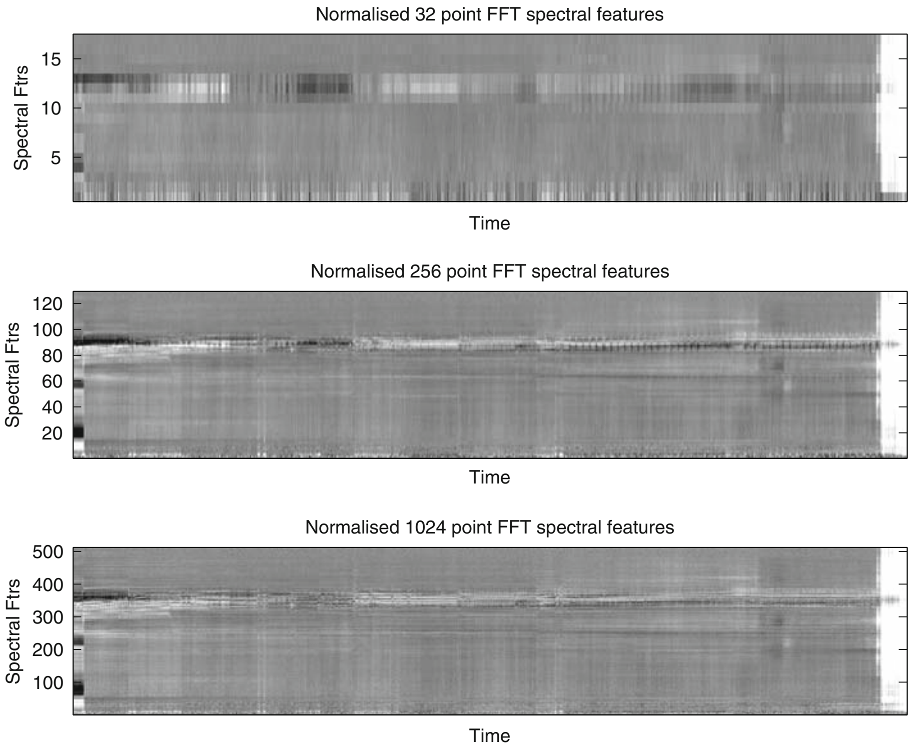

Fig. 2 Stacked averaged power spectral feature vectors sampled periodically over the lifetime of the machine for different frequency resolutions

2.1 Feature subset selection

In this paper, Mutual Information (MI) is used as a basis for selecting a subset of all possible spectral features to optimize the choice of inputs to the classifier in order to predict the RUL of the machine being monitored. The work in this study extends the MI based approach to FSS used in (Scanlon et al. 2007a), as our proposed approach does not utilize any class label boundaries in the computation of MI as described later in this section.

2.2 Background

The entropy of a random variable is a measure of its unpredictability (Cover and Thomas 1991). Specifically, if a variable X can take on one of a set of discrete values $\{x_i\}$ with a probability $Pr(X = x_i)$ then its entropy is given by:

$$H(X) = - \sum_{x \in \{x_i\}} Pr(X = x) \log Pr(X = x) \ , \tag{3}$$

If a second random variable C is observed, knowing its value will in general alter the distribution of possible values for X to a conditional distribution, $p(x|C = c)$. Because knowing the value of C can, on average, only reduce our uncertainty about X, the conditional entropy $H(X|C)$ is always less than or equal to the unconditional entropy $H(X)$. The difference between them is a measure of how much knowing C reduces our uncertainty about X, and is known as the MI between C and X,

$$I(X;C) = H(X) - H(X|C) = H(C) - H(C|X)\ . \tag{4}$$

Further, $0 \leq I(X;C) \leq \min\{H(X), H(C)\}$, and $I(X;C) = 0$, if and only if X and C are independent.

2.3 Implementation

The MI based FSS algorithm (MI-FSS) for feature selection within the candidate pool of spectral features can be expressed as:

$$X_i = \underset{X \in \mathcal{X} \setminus \mathcal{X}_{i-1}}{\operatorname{argmax}} \{I(X;C)\} \quad \text{and} \quad \mathcal{X}_i = \mathcal{X}_{i-1} \cup X_i \tag{5}$$

for $i = 1, 2, \ldots, d$, with $\mathcal{X}_o = \emptyset$, where d is the desired dimensionality of the selected feature vector. Note that this approach represents a simple sorting of all MI values and it results in a nested selected feature set $\mathcal{X}_1 \subset \cdots \subset \mathcal{X}_d \subset \mathcal{X}$. Note also, however, that this greedy strategy does not find the optimal set of d points since there may be information ‘overlap’ between the successively-chosen X points. In the worst case, two spectral components that have identical values would have equal $I(X;C)$ (and would thus be neighbours in the sorted list), but including the second would not add any additional information about C over that provided by the first.

To obtain estimates of the MI values, needed in (5) the *histogram approach* was used to approximate the density functions required in (4). The histogram approach requires choosing the number of bins to be used and their bin widths. The upper and lower bounds are set equal to the maximum and minimum of the samples. Following (Doane 1985), Doane’s rule, $K = \log_2 n + 1 + \log_2(1 + \widehat{k}\sqrt{n/6})$ is used to determine the number of bins to estimate $p(X|C)$ and $p(X)$. In this rule, $\widehat{k}$ is the estimate of the kurtosis of the spectral components (i.e., of random variable X), and n is the total number of training samples. Note that the kurtosis estimates indicate that the spectral components are non-Gaussian.

Given the number of bins, equally spaced intervals are formed b_k, $k = 1, 2, \ldots, K$, between the upper and lower bounds computed for each X. Then $p(x) \approx n_k/n$, iff $x \in b_k$, is approximated where n_k denotes the number of observations $x \in b_k$.

The class labels $c \in \{c_i\}$ are required for the training samples to obtain the n_c and $n_{k,c}$ counts, thus estimating $p(c) = n_c/n$ and approximating $p(x|c) \approx n_{k,c}/n_c$, for all $x \in b_k$, $k = 1, 2, \ldots, K$, and $c \in \{c_i\}$.

In order to obtain estimates of the density functions to compute (4), the class labels are required. Throughout its life, the machine progresses through several stages of physical wear (classes) e.g. increased friction, crack in bearing initiated, crack develops, etc. However, for the task of predicting the RUL of machines, the exact locations in time of the class label boundaries are difficult to ascertain.

Data samples are collected at regular intervals throughout the lifetime of the machine and stored chronologically. This approach to MI-FSS removes any a-priori decisions regarding class label boundaries and takes advantage of the fact that there is a chronological order to the collected data samples. For the purpose of MI-FSS (and not for classification experiments in Sect. 5), the set of classes, $c \in \{c_i\}$, is a set of overlapping short, localised time frames of fixed length e.g. 24 h or 100 samples. The samples that fall within each short, localised time frame are labeled as belonging to that class.

Given that N, the chronologically ordered set of samples acquired over the entire lifetime of the machine, is indexed by n, where $0 \leq n \leq N-1$ and h is the pre-defined number of samples in a localised time frame, then the set of samples belonging to each class c can be defined as

$$c_i = \{n_i \; n_{i+1} \; \ldots \; n_{i+h-1}\} \tag{6}$$

This novel approach computes the local entropy within each class and then extracts an average local entropy across all classes. This averaged local entropy is then subtracted from the global entropy (computed over the entire lifetime) to obtain the MI estimate for each spectral feature. The most relevant features are then selected based on the highest MI criterion, as described in (5). The spectral locations of feature subsets, selected using the MI-FSS approach, are shown in Fig. 3. The chosen feature subset is then used as input to multi-class linear and RBF SVM classifiers, described in the following section.

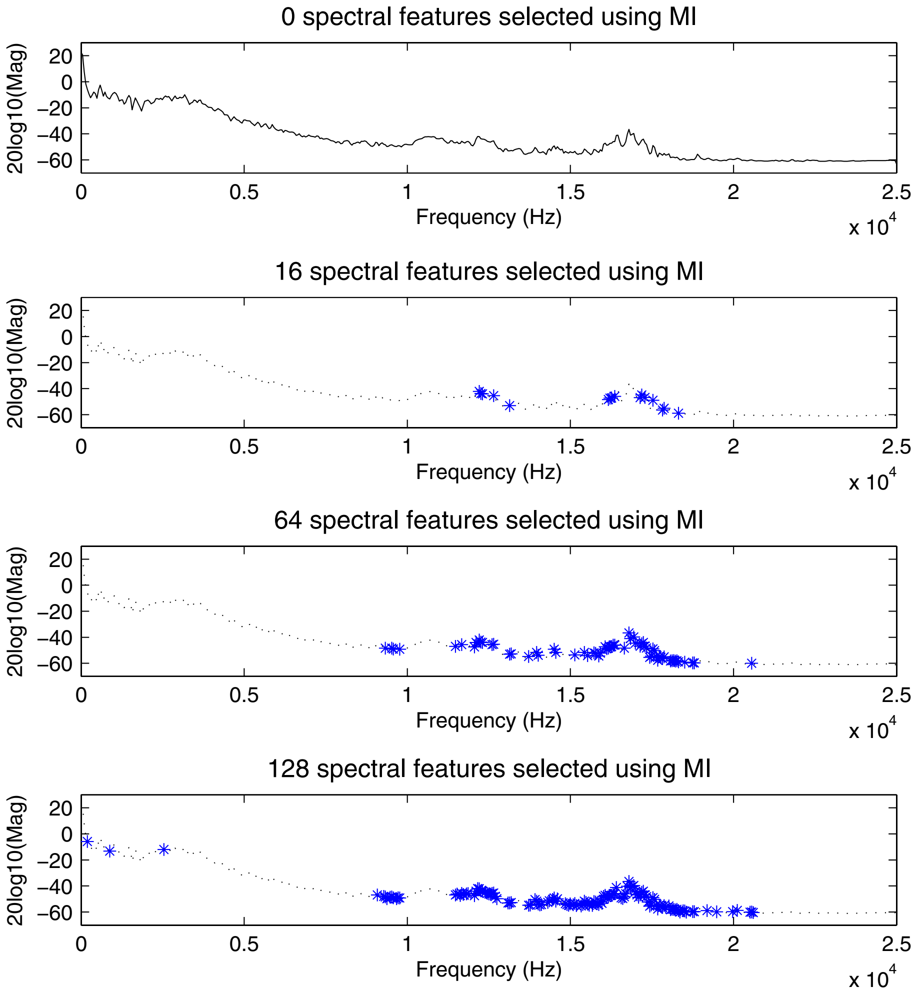

Fig. 3 Selecting spectral components using the data driven Mutual Information Feature Subset Selection approach

3 Classification

Support Vector Machines (SVMs) are a relatively recent set of supervised machine learning algorithms that have been shown to have either equivalent or significantly better generalization performance than other competing methods on a wide range of classification problems (Burges 1998). They can be used to classify linearly separable data using the original input space or non-linearly separable data by mapping to a higher dimensional feature space in which a linear separator can be found.

In a typical binary classification problem composed of a training dataset $\{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_m}, y_m)\}$ where $\mathbf{x_i} \in \Re^d$ and $y_i \in \{\pm 1\}$, SVMs seek a solution to the following Lagrangian optimization function:

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{7}$$

subject to the following constraints

$$C \geq \alpha_i \geq 0 \quad \forall i \quad \text{and} \quad \sum_{i=1}^{m} \alpha_i y_i = 0. \tag{8}$$

C is an optional parameter that controls the trade off between allowing training errors and forcing rigid margins. That is, it represents a soft margin that allows some misclassifications which can be beneficial in noisy datasets. Where a soft margin is not allowed, the constraint is simply $\alpha_i \geq 0$. K represents the kernel function and numerous choices exist, including:

- Linear: $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \cdot \mathbf{x}_j)$
- Polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^d$
- Radial Basis Function: $K(\mathbf{x}_i, \mathbf{x}_j) = exp(-\gamma ||\mathbf{x}_i - \mathbf{x}_j||^2), \ \gamma > 0.$

Once an optimal solution is found, the decision function for a new point $\mathbf{z}$ is given by

$$f(\mathbf{z}) = sign\left(\sum_{i=1}^{m} y_i \alpha_i K(\mathbf{x}_i, \mathbf{z}) + b \right). \tag{9}$$

$\mathbf{z}$ is a training example, b is the bias and non-zero α_i values represent support vectors, the points that lie closest to the hyperplane.

In this study two SVM kernels are implemented—a linear SVM and an RBF SVM. Linear SVMs have very fast training times and do not require any parameter tuning (except C when soft margins are used). Moreover, if the samples can be correctly classified using a linear decision boundary the computational complexity associated with non-linear kernels can be avoided. RBF SVMs are currently the most popular choice of non-linear SVM and thus are an appropriate algorithm for a first experiment in this problem space (Burges 1998; ScholKopf and Smola 2002).

A number of approaches have been proposed to extend SVMs to handle multi-class classification problems, for example, one-against-all, one-against-one and directed acyclic graph SVM (DAGSVM). 'One-against-one' (Knerr et al. 1990) is implemented in this study as it has been shown to have comparable if not better generalization accuracy than alternative techniques and requires considerably less training time (Hsu and Lin 2002; Milgram et al. 2006). The method consists of constructing an SVM for each pair of classes. Thus for a problem with n classes $n(n-1)/2$ SVMs are trained to distinguish between the samples of one class form the samples of another class. For an unknown pattern, each SVM votes for one class and the class with the highest number of votes is chosen.

4 Data labeling

In order to predict the RUL, a classification approach is proposed that determines what state of degradation, or 'wear state', the machine is currently in. Throughout its lifetime the machine transitions through several stages of physical wear: from initial friction caused by lubricant degradation of the bearings; to initial micro-cracks in bearing components; to development of cracks; to bearing seizure and many stages in between. In this study, 10 such 'wear states' or classes are used to predict the RUL e.g. class 1 refers to 100% RUL (new machine) and class 10 to 10% RUL (failure imminent).

As the exact location in time where such degradation events occur is difficult to ascertain, two different approaches to data labeling are employed. The first divides the data (in chronological order) into 10 equal segments for labeling. The second approach uses a automated unsupervised approach to clustering of the data using the k-means algorithm(MacQueen 1967).

K-means clustering is an algorithm to classify or to group samples based on their attributes into K number of groups where K is positive integer. The k-means algorithm is not guaranteed to return a global optimum. The algorithm is very sensitive to the initial set of clusters. In this study, the initial cluster centroids were set to the 10 means of the equal segments as discussed previously. The samples are then partitioned into K new sets by associating each sample in the data set with the closest centroid from the set of K possible centroids. The centroids are then recalculated for the new clusters and the algorithm is repeated by alternate applications of these two stages until convergence, which is obtained when the data samples no longer switch clusters or the centroids are no longer changed.

Note that, using the k-means clustering approach allows cluster assignments that are not contiguous in time. Therefore a smoothing algorithm was also employed where the mode of the clusters assigned to ±15 feature vectors around each feature was used to determine which cluster a feature vector belonged to. The output of the k-means/smoothing clustering algorithm redefines the class label boundaries as shown in Fig. 4.

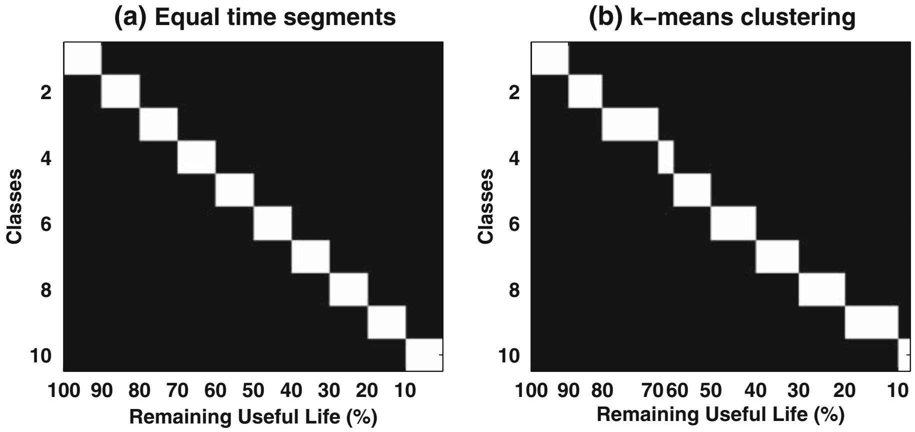

Fig. 4 Automated clustering using k-means to redefine class label boundaries. (**a**) Equal time segments. (**b**) k-means clustering

5 Experimental setup and implementation

Acoustic data was collected from a rotating machine running at approximately 2,200 RPM (37 Hz) in high heat conditions to accelerate failure over a period of approximately 6 months. Note that under normal operating conditions the typical lifetime of this fan is 7 years. The final failure was due to complete bearing seizure. Acoustic data was acquired at a sampling rate of 50,000 samples/s over the lifetime of the machine. Short-term log-spectral features were extracted from the acoustic signal using 1,024 point FFT to provide sufficiently high frequency resolution to uncover the relevant information required for effective automated monitoring that would otherwise be hidden using a coarser resolution. The resultant spectral features span the entire spectrum from 0 to 25 kHz. Information pertinent to machine monitoring occurs at localised spectral locations across the spectrum. The rest of the spectrum contains background noise that is irrelevant to machine monitoring. To remove such noisy components the MI approach to FSS is employed as described in Sect. 2.1. The average MI over the entire 513 extracted spectral features is approximately 0.5 with a standard deviation of 0.25, maximum MI is 1.6. Therefore, only the 200 features with a MI value above the mean were considered for use in classification. To investigate smaller sets of features that could be used to predict RUL with high accuracy, subsets of 25, 50, 100 and 150 features were also examined.

The procedure taken to implement the linear SVMs was as follows. First the attributes are scaled to avoid attributes in greater numeric ranges dominate those in smaller numeric ranges. Initially a fixed penalty parameter (C) of 1.0 is used as if high performance results can be achieved using a default value, then lengthy parameter tuning can be avoided. Generalization accuracy was determined using 10-fold stratified cross validation. In this procedure, data is randomly split into 10 parts, with each part representing the same proportion of each class or wear state. Each part is held out in turn and the learning scheme is trained on the remaining nine parts. The error rate is calculated on the holdout (test) set. The procedure is executed 10 times on different training sets and the results are averaged over all of the testing datasets. Although this approach is more computationally intensive than the commonly used ‘hold out’ method, all examples in the dataset are used for training and testing and thus confidence on the generalisability of the results is increased. In addition the stratification process improves the representativeness of each fold as the process seeks to represent the same proportion of each class in a fold as is in the original full dataset.

The implementation process for the RBF kernel was more involved as two parameters (γ, the kernel parameter and C) are required. First, the attributes are scaled. Then a grid search using 10-fold cross validation (as with the linear SVM) was performed to find the best γ and C parameters.

Unless otherwise stated, results outlined in Sect. 6 are obtained using the first approach to data labeling outlined in Sect. 4 which divides the data (in chronological order) into 10 equal time segments for labeling.

6 Results and discussion

6.1 Support vector machines

The results from the two SVM procedures outlined in the previous section are provided in Table 1. From Table 1 it can be seen that when fewer features ($n = 25, n = 50, n = 100$) are used the RBF SVM considerably outperforms the linear SVM. However, the difference

Table 1 Comparison of linear and RBF kernel Support Vector Machine classifiers using subsets of the spectral, features based on the Mutual Information criteria

No. of features	Linear SVM (%)	RBF Kernel (%)
25	84.86	92.72
50	91.69	96.41
100	94.68	97.90
150	96.06	98.09
200	96.29	98.24

is far less pronounced where more features are used ($n = 150, n = 200$). It appears that the problem becomes more linearly separable as more features are included. Given that a comparision between linear SVM classifiers built with no penalty parameter tuning with RBF SVM classifiers built using considerable parameter tuning is rather harsh, a further evaluation of performance was carried out using alternative default C values (between 2^0 and 2^3, increments of 1.0). It was found that generalization accuracy could be considerably improved to 97.6% when $C = 7.0$. Whilst the difference between the linear SVM and RBF SVM classifiers is statistically significant, that is, the difference is unlikely to be due to chance, it is not meaningful. The performance, in particular of the 200-feature linear SVM ($C = 7.0$) and the 200-feature RBF SVM, is highly comparable. The training time of the linear kernel is significantly less than that of the RBF kernel, for example on the 200-feature set problem, the RBF takes approximately 25 times longer to run on the same machine where no other processes were active. Although, both provide very high performance, this paper promotes the use of a linear SVM using 200 features for several reasons. First, it is far less computationally intensive than the RBF SVM and achieves comparable results. Second, as outlined earlier, it has a considerably faster training time. This is an important consideration from a company perspective where delays in deploying a monitoring component in the field potentially results in increased costs and lost revenue. Third, it is intuitively easier to understand and interpret the problem space and solution. However, additional studies evaluating alternate classifiers in this problem space are warranted. All further results discussed in this paper are based on the 200-feature linear SVM ($C = 7.0$).

6.2 Analysis of misclassifications

Further analysis of the results is valuable to augment preventative maintenance scheduling and machine replacement. In terms of misclassification, predicting something is going to fail earlier than it truly will, results in, at worst, a waste of resources, for example preventive maintenance or machine replacement happens earlier than necessary. The cost incurred is a function of how early this maintenance or replacement takes place, for example, replacing a machine one time-interval before it would have failed is much more cost-efficient than replacing the machine six time-intervals before it might have failed. On the other hand, predicting something is likely to fail later than it actually does could be far more serious, causing, for example, machine down-time and customer dissatisfaction. To this end, an analysis of how 'inaccurate' the misclassifications are is a worthy addition to this study. As illustrated in Table 2, 42% of failures are predicted as happening earlier (1 time-interval earlier) than they truly do and as such are not considerably costly. Furthermore, the 2% predicted to have failed two time-intervals earlier than they truly would is arguably insignificant given how small an error this is. Thus, the real cost of misclassification is the 56% that are predicted as happening later than when they actually do and further analysis is required.

Table 2 Misclassification analysis

Description	Misclassified (#)	Misclassified (%)
Failed 1 time-interval later than predicted	126	52
Failed 2 time-interval later than predicted	8	3
Failed 3+ time-intervals later than predicted	0	0
Failed 1 time-intervals earlier than predicted	102	42
Failed 2 time-intervals earlier than predicted	4	2
Failed 3+ time-intervals earlier than predicted	3	1
Total % of samples misclassified	243	(2.4)

Although SVMs typically only output a target label for each input, an extension to the algorithm is possible to generate probability estimates for each sample. The estimates are based on the distance each test point is from the separating hyperplane, the further the point is from the hyperplane, the higher the probability it belongs in the class (Platt 2000). Analysis of the probabilities reveals that for 95% of the misclassified instances the actual class had the next highest probability to the predicted class. This is very important as it provides a measure of confidence for the predicted RUL and allows preventative maintenance and replacement to be scheduled in a more knowledgeable fashion.

Further analysis of results in Table 2 were carried out to determine exactly where in time misclassifications were made. Most of the misclassifications appeared around the middle and late classes or time-intervals and this suggests that equal time segmentation is not the most suitable technique for class boundaries in these regions. A more sophisticated separation of the class boundaries, for example using a clustering technique, could further improve this misclassification error.

6.3 Further analysis of class label boundaries

The simple approach of setting class boundaries by dividing the data (in chronological order) into 10 equal time segments for labeling yielded very impressive results. However, the authors believed that further improvements could be made by using a more sophisticated labeling approach. Section 4 describes an automated approach using k-means clustering to redefine the class label boundaries.

A linear support vector machine that utilizes the new redefined class label boundaries on the 200 feature set was developed. Using this model, generalisation accuracy improves from 96.29 to 98.11%, effectively halving the previous number of misclassifications. Furthermore, developing a linear SVM utilising an optimal C value determined from a fast grid-search algorithm of values between 2^0 and 2^6 resulted in even further improvements, with generalisation accuracy increasing to 98.96% and the total number of misclassified instances reducing to 105 (out of a total 10, 155 examples).

As discussed earlier, predicting something is going to fail earlier than it truly will, results in, at worst, a waste of resources where the cost incurred is a function of how early this maintenance or replacement takes place. Replacing a machine one time-interval before it would have failed is much more cost-efficient than replacing the machine six time-intervals before it might have failed. Notably with the optimised linear SVM, 41% of the samples failed one class later than predicted, thus the avoidable cost of preventative maintenance for these instances is very low.

Further analysis indicated that 44% of the samples failed one time-interval earlier than expected. However, for each sample the correct class (actual class of failure) had the next highest probability of occurring. The remaining 15% of misclassifications were out by between 2 and 4 intervals. For these instances 66% of the correct classes had the next highest probability distribution value. Thus overall for 88% of all misclassifications the actual class had the next highest probability of occurring. As discussed earlier, a system that incorporated probability distributions as a measure of confidence for the predicted RUL would provide additional valuable information for scheduling preventative maintenance. Finally, it would appear that future work focused on optimising the existing algorithm on these sensitive cases could improve generalisation accuracy even further.

7 Conclusion

Commonly employed machine monitoring techniques, such as measuring changes in speed and current are only useful in indicating when failure is imminent. To successfully determine the RUL of the machine to increase reliability and decrease machine downtime more sophisticated measurements such as vibrations and acoustic noise can be used. Employing the use of acoustic noise measurements as opposed to the commonly used vibrational signal allows for remote, non-contact, monitoring of the machine. Automated machine monitoring using the acoustic noise signal to determine RUL, to the best of our knowledge, has not been attempted before in the literature.

The results of the experiments described in this paper have indicated that there exists sufficient information in the acoustic noise signal of rotating machines in order to effectively determine the RUL. A novel approach to FSS on a high dimensionality spectral feature vector was proposed in order to remove noisy spectral components before classification using MI. This approach does not require information regarding the spectral locations of the defect frequencies or class label boundaries and only uses the chronological order of the collected data samples in the computation of MI to extract the relevant features. In addition, the detailed misclassification analysis provides valuable knowledge and a measure of confidence in the predictions that can be used to further optimize preventative maintenance scheduling and machine replacement. Further improvements in performance are achieved using an automated k-means clustering approach to re-define the class label boundaries for classification, yielding an accuracy of approximately 99%. Close analysis of the misclassifications show that for 88% of all misclassifications the correct class had the next highest probability distribution value. Thus a system that incorporated probability distributions as a measure of confidence for the predicted RUL would provide additional valuable information for scheduling preventative maintenance.

References

Bell D, Wang H (2000) A formalism for relevance and its application in feature subset selection. Mach Learn 41(2):175–195

Burges C (1998) A tutorial on support vector machines for pattern recognition. Knowl Discov Data Min 2(2):121–167

Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York

Doane D (1985) Aesthetic frequency classifications. Am Stat 30:181–183

Goumas S, Zervakis M, Stavrakakis G (2002) Classification of washing machines vibration signals using discrete wavelet analysis for feature extraction. IEEE Trans Instrum Meas 51(3):497–508

Heng R, Nor M (1998) Statistical analysis of sound and vibration signals for monitoring rolling element bearing condition. Appl Acoust 53:211–226

Ho D, Randall RB (2000) Optimisation of bearing diagnostic techniques using simulated and actual bearing fault signals. Mech Syst Signal Process 14(5):763–788

Hsu CW, Lin CJ (2002) A comparison of methods for multiclass support vector machines. IEEE Trans Neural Netw 13(2):415–425

Jayaram V, Jarchow F (1978) Experimental studies on ball bearing noise. Wear 46:321–326

Knerr S, Personnaz L, Dreyfus G (1990) Single-layer learning revisited: a stepwise procedure for building and training a neural network. In: Hérault J, Fogelman F (eds) Neurocomputing, algorithms, architectures and applications, Springer-Verlag, Berlin, pp 41–50

Lao H, Zein-Sabatto S (2001) Analysis of vibration signal's time-frequency patterns for prediction of bearing's remaining useful life. Proc 33rd Southeast Symp Syst Theory 1:25–29

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. Proc 5th Berkeley Symp Math Stat Probab 1:281–297

McFadden P, Smith J (1984) Vibration monitoring of rolling element bearings by the high frequency resonance technique a review. Tribol Int 17(1):1–18

Milgram J, Cheriet M, Sabourin R (2006) 'one against one' or 'one against all': which one is better for handwriting recognition with svms? Tenth International Workshop on Frontiers in Handwriting Recognition

Platt J (2000) Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Smola A, Bartlett P, Schölkopf B, Schuurmans D (eds) Advances in large margin classifiers, MIT Press, Cambridge, MA, USA, pp 61–74

Proakis J, Manolakis D (1996) Digital signal processing—principles, algorithms and applications. Prentice Hall, Englewood Cliffs, NJ, USA

Rojas A, Nandi A (2005) Detection and classification of rolling-element bearing faults using support vector machines. IEEE Workshop on Machine Learning for Signal Processing, pp 153–158

Samanta B, Al-Balushi K, Al-Araimi S (2003) Artificial neural networks and support vector machines with genetic algorithm for bearing fault detection. Eng Appl Artif Intell 16(7):657–665

Scanlon P, Ellis D, Reilly R (2007a) Using broad phonetic group experts for improved speech recognition. IEEE Trans Speech Audio Process 15(3):803–812

Scanlon P, Lyons A, O'Loughlin A (2007b) Acoustic signal processing for degradation analysis of rotating machinery to determine the remaining useful life. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics

ScholKopf B, Smola AJ (2002) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge, MA, USA

Shi D, Wang W, Qu L (2004) Defect detection for bearings using envelope spectra of wavelet transform. J Vib Acoust 126(4):567–573

Tandon N, Choudhury A (1999) A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings. Tribol Int 32:469–480

Tandon N, Nakra B (1990) The application of the sound intensity technique to defect detection in rolling element bearings. Appl Acoust 29(3):207–217