Feature Extraction Using Wavelet Analysis with Application to Machine Fault Diagnosis

by

Reza Tafreshi

B.Sc., K. N. Toosi University of Technology, Iran, 1991

M.Sc., K. N. Toosi University of Technology, Iran, 1995

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

Mechanical Engineering

THE UNIVERSITY OF BRITISH COLUMBIA

April 2005

ABSTRACT

Two different approaches have been used to diagnose faults in machinery such as internal combustion engines. In the first approach, a mathematical model of the specific engine or component under investigation is developed and a search for causes of change in engine performance is conducted based on the observations made in the system output. In the second approach, the specific engine or component is considered a black box. Then, by observing some sensory data, such as cylinder pressure, cylinder block vibrations, exhaust gas temperatures, and acoustic emissions, and analyzing them, fault(s) can be traced and detected. In this research the latter approach is employed in which vibration data is used for the detection of malfunctions in reciprocating internal combustion engines.

The objective of this thesis is to develop effective data-driven methodologies for fault detection and diagnosis. The main application is the detection and characterization of combustion related faults in reciprocating engines; faults such as knock, improper ignition timing, loose intake and exhaust valves, and improper valve clearances.

To perform fault diagnosis in internal combustion engines, cylinder head vibration data are used for characterizing the underlying mechanical and combustion processes. Fault diagnosis includes two main stages: feature extraction and classification. In the feature extraction stage, we have utilized wavelets for the analysis of acceleration data acquired at the cylinder head to capture meaningful features that include necessary information about the state of the engine. Wavelets have shown to provide suitable signal processing means for analysis of transient data and noise reduction. Wavelet packets, as a generalization of wavelets, offer even a more powerful data analysis structure to extract features that are capable of identifying combustion malfunctions. Various concepts of wavelets, wavelet packets, related algorithms and assessment techniques have been reviewed, analyzed and discussed.

As a result of this research, a novel methodology for fault diagnosis has been developed. This has been achieved through critically investigating available methodologies employed in fault diagnosis and classification, and by understanding their shortcomings. The developed method not only avoids the demerits of the previous techniques, but also demonstrates superior performance.

To compare the performance of the proposed approach with major existing methods, various sets of real-world machine data acquired by mounting accelerometer sensors on the cylinder head, as well as a set of synthetic data, have been extensively tested.

CONTENTS

Chapter 1- Introduction

Chapter 2- Pattern Recognition and Wavelets

Chapter 3- Experimental Setup, Data Collection and Preparation

Chapter 4- Mutual Information and Informative Wavelet

Chapter 5- Dictionary Projection Pursuit

Chapter 8- Conclusions

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I owe an enormous debt of gratitude to my supervisor, Dr. F. Sassani, for his help, support, encouragement, and patience in the course of this study. I really enjoyed every minute of working with him. I express my sincere thanks to my co-supervisor Dr. G. Dumont for his continued support and indispensable comments and discussions. I would especially like to thank my very active supervisory committee member, Dr. H. Ahmadi, for his great support in both academic and research areas, and for many hours of engaging discussions. My sincere appreciation for their invaluable comments also goes to university examiners Dr. J. R. Marti and Dr. M. Gadala, supervisory committee member Dr. C. W. de Silva, and external examiner Dr. M. J. Zuo.

I wish to thank Dr. Howard Malm of REM Technology Inc., Port Moody, BC, Canada, for his continued assistance for providing both data and technical support, Professor R. L. Evans and Mr. Conor Reynolds of the Mechanical Engineering Department at the University of British Columbia for generously providing access to their engine test facility and technical support. I would also like to express my appreciation to Professor N. Saito of the Department of Mathematics at the University of California, Davis, Dr. B. Liu of the School of Mechanical and Production Engineering, Nanyang Technological University, and Dr. G. Rutledge of the Mechanical Engineering Department at the University of Victoria for their invaluable comments.

I spent a very pleasant time with my dear colleagues in the Process Automation and Robotics Laboratory: Dr. Reza Ghodsi, Mr. Mohammadreza Mallakzadeh, Mr. Tao Fan, Ms. Shabnam Khatami, Ms. Nazly Pirmoradi, Mr. Mohammad Amin Karami, and Mr. Daniel Wong. I would like to thank them for their moral support.

Finally, I wish to express my deepest gratitude to all members of my big family for their unwavering support and encouragement, especially my late father, my never-ending source of inspiration and encouragement, my mother, who inspired me with her encouragement and moral support, my brothers and sisters, who were there for me, especially my eldest brother who was always supportive. Last but not least, to my wife without whose understanding and support I could not finish this, a special thank-you.

To the memory of my beloved father

To my dearly loved mother

To my dear wife and adorable daughter

CHAPTER 1

Introduction

1.1. Foreword

As machinery becomes more complex and costly to build and operate, preventive maintenance measures also become increasingly important. Implementing machinery performance analysis has been found to offer several benefits including financial, operational, and even environmental advantages. When a machine breaks down it can lead to significant problems including downtime cost, loss of functionality and productivity, catastrophic failure which might be beyond repair, and loss of life.

There is currently a great need for equipment and software to automatically predict, detect, and diagnose faults in machines and components. Important examples of such systems include rotors and turbines, compressors and engines, bearings and gearboxes, and cutting and drilling tools.

Large reciprocating engines, such as diesel and spark ignited engines, are essential elements of most industries. They can be seen in a wide range of applications, such as in power plants, chemical plants, petroleum industries, and in the operation of compressors and pumps, pipelines, ships, and trains.

Condition monitoring for fault detection and prevention is now an integral part of the operation of many industrial processes and machinery [1]. It plays an important role in maintaining quality standards, increased productivity and cost reduction. Unlike traditional reactive maintenance practices, condition monitoring and predictive maintenance attempt to avoid unnecessary shutdowns and thus reduce machine downtime and increase productivity.

This thesis focuses on the fault diagnosis of reciprocating engines, though once the procedure is developed, the method can be extended for similar pattern recognition applications.

In this chapter, the performance fundamentals of different internal combustion engines along with some possible engine faults are introduced. The significance of using vibration data in machine diagnosis is discussed. Such data are collected by means of accelerometer sensors mounted on the cylinder head. A literature survey of fault diagnosis is also followed.

1.2. Performance of Internal Combustion Engines

Internal combustion engines can be classified by the method of ignition into two categories: spark-ignition (SI) and compression-ignition (CI) engines. SI engines usually use gasoline, natural or liquid gas as fuel, while CI engines typically consume diesel fuel or natural gas.

In conventional spark-ignition engines appropriate amounts of fuel and air is supplied to the engine cylinder. The mixture of fuel, air and residual gases are compressed, and then combustion is initiated toward the end of the compression stroke by the electric discharge of a spark plug. Inflammation develops and propagates by a turbulent wave of flame throughout the entire air-fuel mixture up to the combustion chamber (cylinder) walls.

In compression-ignition (*diesel*) engine combustion process, fuel-injection system sprays the fuel into the combustion chamber just before the desired start of combustion which is toward the end of the compression stroke. Sprayed fuel which has been atomized through the high velocity injection is vaporized and mixed with high-temperature high-pressure cylinder air. In a few crank angle degrees, pressure and temperature reach the fuel's ignition point. The portion of fuel which has properly been mixed by air ignites spontaneously. As a result, the pressure and temperature increase and cause the unburned portion to continue the combustion at an accelerated rate.

Usually the performance of every engine is measured by some parameters such as power, torque, and specific fuel consumption [2]. Any significant deviation from nominal conditions results in a change from the optimal condition in the above characteristic parameters, which is generally referred to as a *fault*. This thesis focuses on the detection of some of these faults that occur in the engine cylinder. Examples of such faults are introduced in the next section.

1.3. Engine Faults

Requirement for efficient power generation, compliance with exhaust control standards for lower emissions, and finally prevention of engine wear caused by occurrence of combustion fault are some of the main motives for the research in this area. There are several faults associated with combustion; among them, intake and exhaust systems related faults, knock, and ignition timing faults are our focus.

Faults in intake and exhaust systems include small and large valve gaps and gas leakage in both intake and exhaust valves. Opening and closing of intake and exhaust valves with appropriate timing are issues that also engage engineers. Research shows that correct exhaust valve timing reduces hydrocarbon emissions [3], and leads to better engine performance.

In SI engines, knock phenomenon as a state of combustion event, refers to the *autoignition* of end-gas (the unburned fuel-air mixture ahead of flame) inside the cylinder. In fact, autoignition refers to any combustion activity which is not initiated by an external ignition source. Knock, a "sharp metallic noise", is caused by high compression and combustion of end-gas between the piston and burned gas behind the flame. Autoignition of the end-gas is accompanied by shock waves inside the combustion chamber that causes high frequency pressure resonance, which is transmitted to the cylinder head assembly. The frequency band of knock condition is determined by the specifications of a given engine such as cylinder diameter and engine parameters [2]. Knock is a common abnormal combustion phenomenon in SI engines.

Even though CI engines operate based on autoignition (as there is no spark plug, and ignition is based on compression as discussed in previous section), knock can still occur in these engines. The non-uniformity of in-cylinder pressure generates some combustion waves, known as *detonation* waves, that contribute to sonic velocities of the hot gases in the cylinder. These fast moving combustion waves create knock.

Two important parameters that promote the occurrence of knock, are high load and over-advanced (too early) ignition. Detection of deviations in ignition timing is another important issue in engine fault diagnosis since engine Efficiency is highly influenced by ignition timing. In fact, optimal power generation and lower pollutant emissions are achieved when ignition timing is set at the onset of knock occurrence.

1.4. Use of Acceleration Data

Non-intrusive measurements such as the use of cylinder head acceleration data in engine monitoring applications offer clear advantages over other measurement types such as intrusive cylinder pressure measurement. The use of acceleration data – obtained from vibration sensors located on cylinder head or engine block – has become popular in a wide range of fault diagnosis applications, including detection of knock [4,5,6,7], detection of valve clearance and gas leakage in both intake and exhaust valves [8], as well as detection of drift in ignition timings [9]. However, the accuracy of the diagnostic results is highly machine dependent and is subject to large errors when the speed of engine is variable or the number of cylinders is high. Other factors may also affect the accuracy of diagnostic results including size and type of the engine, manufacturer, and more significantly, operating conditions such as sizable load variation.

There are inherent complexities in machine diagnosis that are accentuated using non-intrusive methods. Vibration data are often contaminated with considerable noise generated from overall engine vibration and from sources external to a given machine fault. Noise masks the true fault signatures and can complicate information extraction. The extent of noise content often differs for different engines and different application environments.

While the detection of combustion related faults has been a focus of study by several groups in the past, there are still some challenges that are discussed next.

1.4.1. Challenges of Engine Diagnosis Using Acceleration Signals

The use of vibration data for engine diagnosis is accompanied with complexities outlined below.

(1) Path dynamics. Acceleration signals that are acquired at the location of the sensors are characteristically different from the signals that were originally created within the combustion process. While traveling from "source" to "sensor", the signal is in fact affected by the dynamics and the mechanical properties of the path. This phenomenon is referred to as *path dynamics*. Different source-to-sensor paths will impact different "path dynamics" upon the signal being transmitted. For example, accelerations measured in horizontal and vertical directions caused by the same combustion event can exhibit different characteristics.

It is worth noting that path dynamics is a non-linear phenomenon whose modeling is an intricate task, specially that the modeling must be done for every engine type.

(2) Machine Event Variations. Machine events, despite their cyclic and repetitive nature, exhibit variability from one cycle to the next [2]. Such variability is triggered by the underlying variability of source causes and processes governing different machine events. Variability is best described by the statistical nature of the variables defining these events that inhibits extraction of signatures for fault detection. It is observed under both steady state nominal and faulty conditions.

Variability is caused by several factors; they may be categorized as internal or external for a given machine event. For example, variation in combustion quality may be caused by variation in fuel mixing process or compression due to valve operation. Variability as measured by standard deviation of vibration signals, may be different for different engines, machine events, and locations. Vibration signals at the cylinder head exhibit a higher degree of variability as compared with cylinder pressure signals; they carry the effect of the variability of both source event and external factors such as valve opening or closing events as well as non-linearity of machine response to combustion event.

Fig.1.1 illustrates spectral variability as measured by power spectral density of vibration signals in 16 consecutive combustion cycles of a 12-cylinder industrial engine. Power spectral density is a global measure of the engine operation and does not reflect localized time variations of signal behavior. As such, it is a suitable measure to examine variability of engine operation from one cycle to the next. Variability of vibration data is more observable in time domain. Using standard deviation as a measure of variability, the ratio of std/mean remains nearly unchanged in all frequencies. Other global metrics of acceleration data such as signal energy indicate similar statistical variability from one cycle to the other.

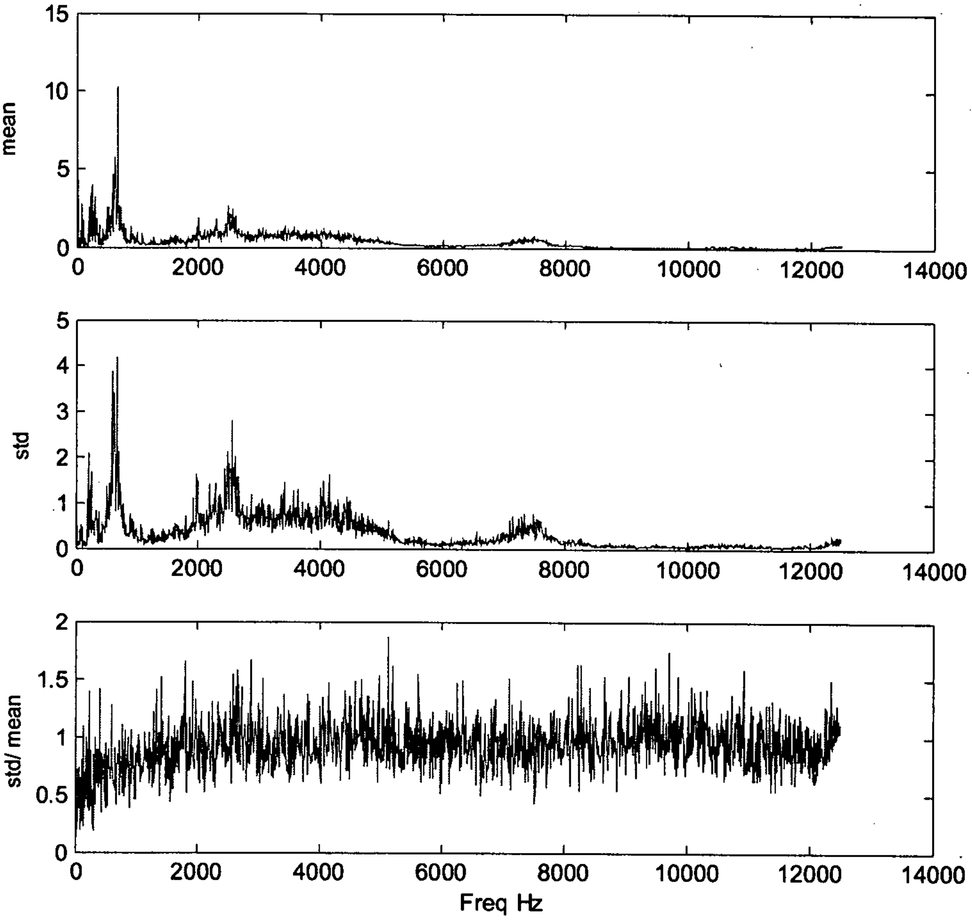


Fig. 1.1. Mean and standard deviation of FFT of acceleration data, and spectral variability measured by std/mean.

Variability causes dilution and disguising of fault signatures and introduces inaccuracies that are more pronounced when the intensity of faults is low. Under mild faults, fault signature falls within the variability range of the measured signals and remains undetected. For high intensity faults, such as hard knock, fault signature is dominant and variability is of lesser significance. In acceleration data, different frequency bands may exhibit different variability.

(3) Noise. Vibration signals are often contaminated with noise that inhibits the extraction of true signal signatures for diagnosis. As mentioned before, Noise in machine data is generated from the overall engine vibration caused by totality of machine events. In diagnosis of combustion-related faults, noise is also caused by events that are external to combustion event such as valve opening and closing, and cross-talk effects of consecutive combustion in adjacent cylinders. Vibrations of engine components oscillating at their harmonic frequencies are also transmitted to cylinder head position and are compounded with acceleration signals. Noise reduction is considered as an important part of machine data analysis; it can play an effective role in reducing inaccuracies of diagnosis results.

1.4.2. Possible Solution

An approach used widely in dealing with non-linearity as well as variability effects in fault detection and classification problems is to search for changes in the statistical distribution of the respective data derived under different operating conditions. In this approach, training data are used to extract necessary statistics under different fault conditions. For example in knock detection, spectral energy at a set of designated frequency bands are considered as feature variables and used as inputs to a fault classifier such as a neural network. While statistical modeling and neural network classifiers have been used in numerous fault detection problems, there are still limitations to their use in real-world applications. Access to sufficient number of training data under a given condition and known intensity level is not always possible. Ample data is necessary in order to capture the statistics of a given condition with sufficient accuracy in order to obtain acceptable discriminatory classification results.

Furthermore, changes in operating conditions of an engine require the use of a new set of training data under the new conditions. In the absence of the necessary knowledge about such changes, accuracy of diagnosis cannot be ensured. In addition, while spectral and time signal energies are considered as valid indicators of knock intensity – as several research groups have used for knock quantification, such as in [10] – spectral energy is a global measure which does not carry necessary information about the time behavior of the acceleration signal under a given knock condition. Changes in the knock conditions are sought in both global and local features of the acceleration signals.

Due to the fact that vibration characteristics of faults are complex and normally buried within wide band engine background noise as well as high frequency structural resonance, they cannot be easily identified through simple signal processing. The signals obtained from an engine are usually contaminated with noise or cross-talk effects. The main purpose of signal processing is to manipulate the information contained within the signal to enhance the view of a desired feature.

In recent years, there have been many attempts to diagnose different faults relating to knock [11,12,13], loose or cracked roller bearing [14,15], ignition timing [9], valve clearance and operation under loose valve condition [16], cracked teeth in gear train [17,18,19], cylinder and ring wear, and injection system problems occurring in engines [20]. In terms of diagnosis, the problem becomes more complex when these begin to develop concurrently. A more detailed literature review is given in the next section, in which wavelets have been employed as an efficient analytical tool used in fault diagnosis. Wavelets as a classes of functions have information localization ability in both time and frequency and will be presented in chapter 2.

1.5. Review of Previous Works

Coifman and Wickerhauser [21] established a mathematical foundation for a method that permits efficient compression of a variety of signals such as sound and images. Their method selects a set of functional forms known as basis that is best adapted to the global properties of signal. Such bases are appropriately selected from a dictionary of orthogonal bases such as local trigonometric functions or wavelet packets family. They used Shannon entropy measure as a cost function to match a basis to a given signal or family of signals.

In wavelet packet transformation (which will be defined in the next chapter), the selection of best decomposition tree for signal representation is usually done through entropy cost function as introduced in [21]. Even though entropy is considered appropriate for signal compression, it may be unsuitable for signal classification [22,23]. Saito and Coifman [22] used the concept of relative entropy as a cost function in classification applications, and termed their algorithm the *local discriminant basis* (LDB). Similar to the process in [21], LDB selects an orthonormal basis from a dictionary, which most discriminates different classes in a given set of data belonging to several classes. We will expand more on their algorithm in chapter 6. They later modified their algorithm by employing relative entropy of the empirical probability density estimate of each class in a wavelet packet domain [24].

Englehart *et al.* [25] applied LDB for myoelectric signals (MES) for clinical diagnosis in biomedical engineering. As in LDB, they used the time-frequency energy maps of each class as input to a symmetric relative entropy measure, in conjunction with principal component analysis (PCA) for dimensionality reduction. In another MES application, they proposed time-frequency methods such as short time Fourier transform (STFT), wavelet transform (WT) and wavelet packet transform (WPT) for feature extraction, along with PCA, and found WPT superior for classification purposes [3,23].

Mallat and Zhang [26] introduced an algorithm called *matching pursuit* which searches for a set of wavelets to represent an individual signal, then efficiently decomposes the signal into a linear expansion of waveforms that belong to a library of functions. Matching pursuit is referred to as a greedy algorithm in which a signal is decomposed into a sequence of components generated iteratively with a projection direction that has the highest match with the residual at each stage. This algorithm is closely related to projection pursuit strategy developed in [27]. These methods have proved their utility in signal presentation and compression problems.

Using matching pursuit algorithm, in 1999, Liu and Ling [20] proposed a measure to identify a small set of wavelets that carry meaningful information about machinery faults and tried to identify the wavelets that are sensitive to fault occurrence. The application of this set of wavelets, which is called *informative wavelet*, has been expanded in [16] for combustion fault diagnosis.

Samimi and Rizzoni [4] used time-frequency analysis of pressure signals in order to detect knocks in an internal combustion engine. Since the pressure measurement is not considerably affected by signals or noises from other mechanical sources, contrary to vibration measurements, it is a very reliable measurement for knock detection, but using pressure transducers in existing industrial engines is neither easy nor economical.

Yang *et al.* [28] used a dynamic model to simulate instantaneous angular speed to obtain cylinder pressure information to diagnose combustion-related faults. This could be a useful approach for single cylinder engines, but a slight pressure changes in one cylinder of a multi-cylinder engine is expected to have a very little effect on the angular velocity of the engine. For this reason their method can hardly be applied on engines with several cylinders and large inertia.

Along the same line, Zavarehi and Schricker [29] employed the actual crankshaft angular velocity information of a six-cylinder diesel engine to find kinetic energy of each cylinder. They used such energy to detect possible power loss in a cylinder. They divided the crank angle degree of one engine cycle by the number of cylinders, and used the velocity fluctuations of crankshaft in each segment to find the kinetic energy of each cylinder. This method may be appropriate for detecting complete power loss of one cylinder, but is impractical for detecting small changes in the power of one cylinder, say in a large 12-cylinder industrial engine.

Wang and McFadden [17] applied the wavelet transform to the analysis of the vibration signals of a helicopter gearbox in order to represent gear condition and detect faults. They found that the Gaussian-enveloped oscillating wavelet is well-suited for detection of gear faults. In another attempt Dellomo [24] also used accelerometer data to detect gearbox fault in helicopters. He employed some elementary signal analysis to identify the frequency band of faults and then applied Fourier analysis for fault monitoring. The particular gearbox signal he examined is a very simple example. Wang [31] viewed time-frequency-scale distribution as a three-dimensional image and tried to obtain the detailed features of a signal. He applied this method to the signals from gears and a steel mill roll, but it's difficult to interpret and quantify the produced three-dimensional images.

Shiroishi *et al.* [32] investigated defect detection methods in frequency domain for rolling element bearings through accelerometer and acoustic emission (AE) sensor and found that the latter is better at detecting the defect types. Their main goal was to increase the signal-to-noise ratio. Ma *et al.* [33] presented a method to design a filter with combination wavelets which is formed by frequency shift and single wavelet superposition. They designed the Gaussian combination wavelet filter which has been applied by Luo *et al.* [34] for condition monitoring. Their measure for diagnosis was monitoring the system natural frequencies via vibration signals. They extended the design in [33] to a very narrow frequency band wavelet filter to have more accuracy in recognizing the system natural frequency, then used it in obtaining power spectrum of different conditions in a turning machine to recognize the fault in bearings. Their method is applicable to only high intensity faults which change the natural frequency of the system; there is no guarantee to ensure that these changes are because of faults in bearings and not from other sources.

Friedman [27] proposed an approach referred to as the projection pursuit method (PP). This is a method for exploratory data analysis to find low dimensional projections of high dimensional multivariate data which optimize a projection criterion via numerical computation. Rutledge and Mclean [35] extended this method to choose a subset of basis functions from the wavelet packet dictionary which are orthogonal to one another. They called this method the dictionary projection pursuit (DPP), which will be presented in chapter 5.

1.6. Thesis Layout

Numerous research efforts have been made towards the diagnosis of different engine faults. In this chapter, some of these were reviewed; more work carried out in classification and engine fault detection will be outlined in subsequent chapters.

Chapter 2 outlines the process of pattern recognition, reviews Fourier and Short Time Fourier transforms and sets the context for introduction, definition and application of wavelets and wavelet packets in machine diagnosis. This is followed by an explanation of the concept of entropy and discriminant measures, and their role in signal processing.

In chapter 3, different normalization and preprocessing methods will be presented. The data sets that have been used throughout the thesis along with some preliminary data analysis will be introduced.

In chapter 4, using mutual information and entropy defined in wavelet domain, informative wavelet algorithm [20] will be explained and applied to real-world machine data for classification and diagnosis of a set of designated faults in diesel engines. Several prototype wavelets and data under different operating conditions are employed to examine the effectiveness of the algorithm for the classification of two categories of faults, namely, excess valve clearance and knock condition.

Chapter 5 covers another method used in fault diagnosis known as the dictionary projection pursuit and the underlying algorithm in which its relevance to the research conducted in thesis is investigated in detail.

In chapter 6 the best-basis algorithm and its extension, the local discriminant bases algorithm, are investigated and their role in the research carried out in the present work is outlined. Then, some representative test results are presented and thoroughly analyzed.

In chapter 7 some novel methods for fault diagnosis and classification are introduced, and in chapter 8 conclusions, future work, along with contributions made in this thesis are presented.

1.7. Closing Remarks

In many of the papers reviewed in this chapter, the respective authors have investigated relatively harsh faults, such as gear faults, which are primarily time-invariant and could practically be detected by Fourier analysis techniques as well. Nevertheless, they have shed light on new trends and approaches and some have developed innovative ideas in the field of fault diagnosis. In this research, we develop a methodology for detecting faults with lower levels of intensity in more transient and time-variant environments. In the process, we highlight the shortcomings of several available methods and deal with development and application of wavelet-based techniques for engine fault detection and diagnosis.

CHAPTER 2

Pattern Recognition and Wavelets

2.1. Introduction

Wavelets are classes of functions with properties suitable for the analysis of a wide spectrum of signals found in engineering and scientific applications. Wavelets have been utilized successfully in system modeling and control, image and signal processing, data compression, communication, signal identification, pattern recognition, and feature extraction. An important property of wavelet analysis is the ability to localize feature of a signal analysis both in time and in frequency. As such, wavelet transform is viewed as a generalization of Fourier transform in the sense that it provides both spatial and frequency localization of a given signal. Often wavelets are considered for the analysis of signals to characterize discontinuities, breakpoints, nonstationary and transients behavior [36]. For feature extraction, they have found applications in industrial and biomedical diagnostics where features for change detection are often sought in wavelet coefficients or in their characteristic parameters such as correlation structure and statistics of the coefficients at different scales of signal decomposition.

In this chapter, basic concepts of pattern recognition and classification are presented; then the fundamentals of wavelets signal processing is reviewed. The significance of applying wavelets in machine diagnosis is discussed followed by introducing the concept of Shannon entropy [45]. While entropy is successfully utilized in data compression and signal representation applications, one needs a different criterion in classification. Relative entropy is presented as a natural extension of entropy theory to be employed in classification applications.

2.2. Pattern Recognition

Fault detection and quantification problems may be analyzed within the scope of pattern recognition problems whose goal is to classify objects or patterns into a number of categories or types [12]. Pattern recognition is an integral part of most machine intelligence systems built for decision-making. The major problem associated with pattern recognition is the so-called *curse of dimensionality*, in which the high dimension of data means the existence of high amount of unnecessary information in the original data, and excessive computational time in the data analysis phase.

Both traditional methods such as Fourier analysis and new methods such as time-frequency and wavelet analyses generate large numbers of data features. For pattern recognition and classification applications, it is highly desirable to determine as few features as possible while containing as much information about the faults as possible. There is a pressing need to reduce the dimensionality of these data by extracting a limited number of features which best preserve the useful information. While researchers have applied certain methods such as linear discriminant analysis [12] and neural networks [47], it is clear that much remains to be done in this arena. There are many reasons why feature reduction is essential, including reduction of computational cost, reduced requirements in terms of training time and data, noise reduction, increased robustness, and more rapid training of classifiers. There will also be high mutual correlation among the selected features which could increase the complexity without any gain. Furthermore, high dimensionality causes the information to be diluted.

A pattern recognition system is trained with a finite number of training samples. The trained system must be well generalized to data which were not contained in the training set, but without significant increase in the system complexity.

On the other hand, often, signals acquired for use in a fault detection process carry unnecessary information and cannot be directly used in a given classification problem. Such signals must be suitably preprocessed for order reduction before feature extraction. Preprocessing attempts to associate each class of signals with a certain pattern (signature) that can be used as a feature for classification. In addition, in classification problems, not only we look for features that contain non-superfluous information but also we seek information that can separate classes from each other as distinctly as possible. This type of information is referred to as “discriminant”. Usually, it is the superfluous information that turns the classification into a difficult task. The main objective in feature extraction and classification problems is to find a coordinate system for projecting the signal along its axes, that yields high discriminatory information residing on a few axes, with insignificant information along most axes.

2.2.1. Definition and Process

A linear projection from R^n to R^m is a linear map B represented as an n x m matrix:

$$Z = B^T X, \quad X \in R^{n \times l}, \quad Z \in R^{m \times l} \tag{2-1}$$

which transforms n-dimensional data set X (consists of l data in each column) into an m-dimensional space; Z is the m-dimensional transformed data set. Suppose b_i is the n-dimensional column vectors of matrix B=[b_1 b_2 ...b_m]. If b_is are orthogonal to each other, the projection is called orthogonal, and if they have unit magnitudes, the projection is called orthonormal. If m = 1, then B is a one-dimensional projection, and Z is a scalar sometimes referred to as the *projection score.*

Fig. 2.1 shows main stages of classification in which X is input signal, Y, corresponding class label (e.g. *faulty* or *healthy* conditions), and F, feature space, which is the discriminant subspace of reduced dimension ($m \ll n$). The maps $f : X \to F$ and $g : F \to Y$ are called *feature extractor* and *classifier*, respectively. It is computationally more efficient to analyze the data in a discriminant subspace of lower dimension.

Classification is a complex task because: 1) the signal space dimension is usually very high which makes the classification computationally expensive, 2) as mentioned before, signal space usually includes some unnecessary information, 3) signal space is usually contaminated with noise. Classification goal is to determine which class a given data X belongs to by constructing a feature space F that provides the highest discriminant information among all classes.

This thesis deals with the analysis of cylinder-head acceleration data for engine fault detection and diagnosis. Vibration signals of a machine always carry information about its dynamic behavior, which can be used to identify faults in machine operation.

Vibration signals in internal combustion engines are characterized as being transient, time variant and extremely noisy. Wavelets are considered to be highly suitable for the analysis of transient signals for feature extraction used in fault detection problems.

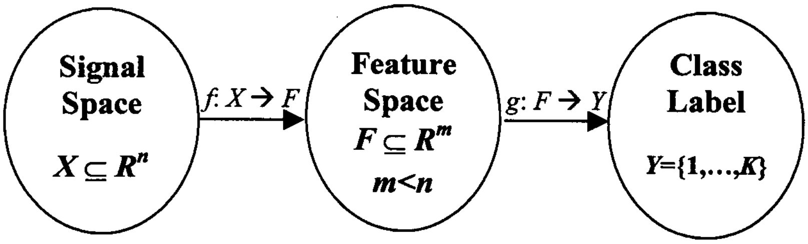


Fig. 2.1. Main stages in classification.

The basic background concepts behind wavelet analysis along with formal definition of wavelets are introduced in the next two sections.

2.3. Fourier and Short Time Fourier Transforms

Conventional techniques such as Fourier analysis are practically valuable for many signals in which the signal's frequency content is of great interest. Fourier analysis decomposes a signal into a sum of constituent sinusoids of different frequencies:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) e^{j\omega t} d\omega .$$

It gives another view of the signal with details that cannot be seen in the signal itself. If f has finite energy; i.e., $f \in L^2(\mathbb{R})$ (L^2 is the space of square integrable functions), then the amplitude $F(\omega)$ of each sinusoidal wave $e^{j\omega t}$ is the Fourier transform of f obtained by:

$$F(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-j\omega t} dt.$$

Sinusoidal waves $\{e^{j\omega t}\}_{\omega \in \mathbb{R}}$ are eigenfunctions (sometimes referred as eigenvectors) and Fourier coefficients $F(\omega)$ are their associated eigenvalues [39,40].

Fourier analysis transforms a time-based signal to a frequency-based one, but in the new (frequency) domain, there is no time information; it is impossible to notify *when* a certain event occurred. If a signal is stationary – as is common in many applications – this drawback may not be so important. However, many signals including engine signals have non-stationary or transitory characteristics; they exist in short periods of time, but consist of local information. Fourier analysis is not capable of detecting these transitory and abrupt changes.

To resolve this deficiency, in 1946 the physicist Dennis Gabor [40] motivated by quantum mechanics, modified the Fourier transform to analyze only a small section of the signal at a time. Gabor's adaptation, called the *Short-Time Fourier Transform* (STFT), maps a signal into a two-dimensional function of time and frequency (Fig. 2.2). It provides some information about both when and at what frequencies a signal event occurs, but its precision is limited by the size of the time window used. Its other weakness is that once one chooses a particular window size, that remains the same for all frequencies. The time-frequency window of any STFT is rigid; in many applications we need a more flexible approach where we can vary the window size to examine an event more accurately either in time or frequency.

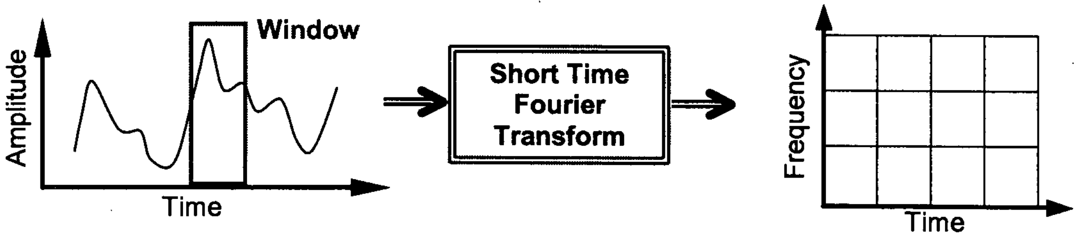


Fig. 2.2. Short Time Fourier Transform.

2.4. Wavelets

Wavelets are classes of wave-like functions that are often irregular, non-symmetric, and with no analytical/mathematical expression. They have finite number of oscillations and an effective length of finite duration. Wavelets are used as basis functions for signal decomposition and signal processing. They allow function expansion in an orthogonal, non-orthogonal or redundant structures. Wavelets are considered as unconditional bases with properties that allow efficient information extraction and coding [41,47].

Wavelets in signal processing can be considered as windowing functions extracting signal information at variable-sized localized regions. It allows the use of windows with long time intervals where we want more precise low frequency information, and short time regions where we want to extract high frequency information. Wavelet transform projects a given signal onto a two-dimensional array of coefficients parameterized by scale (or frequency) and translation (time), while Fourier transform maps a one-dimensional function into a sequence of single parameter coefficients. Two-dimensional signal representation allows localized extraction of signal information both in time and frequency. In standard *discrete wavelet transform* (DWT) this representation is achieved using basis function ψ dilated with a scale parameter j, and translated by k:

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k) \qquad\qquad j,k \in \mathrm{Z}$$

where Z is the set of all integers and the factor $2^{j/2}$ maintains a unity norm independent of scale j. Any finite energy signal f in $L^2(R)$ can be decomposed using wavelet orthogonal basis $\{\psi_{j,k}\}$ [41]:

$$f(t) = \sum_{j,k} a_{j,k} 2^{j/2}\psi(2^j t - k)$$

or

$$f(t) = \sum_{j,k} a_{j,k}\psi_{j,k}(t)$$

where the two-dimensional set of coefficients $a_{j,k}$ is referred to as wavelet coefficients of $f(t)$ and can be determined using inner products defined as

$$a_{j,k} = \left\langle \psi_{j,k}(t), f(t) \right\rangle = \int f(t)\,\psi^*(t)dt\,.$$

Decomposition of a signal can be carried out using a filter bank structure by breaking the signal into a set of low and high frequency components as illustrated in Fig. 2.3, where f=f_v^0 is the original signal, h^j and g^j are low and high pass filters of stage j, $f_v^{\,j}$ and $f_w^{\,j}$, j = $1,\ldots, J$, are called the approximation and detail at resolution level j, respectively. J is the number of decomposition level considered for signal analysis. In standard wavelet transform, low pass and high pass filters h^j and g^j remain unchanged for all stages. At an arbitrary stage j, original signal can be reconstructed from the sum of the all details up to stage j plus approximation at that stage i.e., $f = f_v^j + \sum_{i=1}^{j} f_w^i$ [36].

Filter bank structure can assume orthogonal or biorthogonal structure [42]. They are derived from a prototype wavelet function characterized by a set of parameters including regularity, symmetry and wavelet order [42]. Selection of a particular prototype wavelet for a given application is determined by the requirements of that application and is an area of wide interest in function approximation and signal processing.

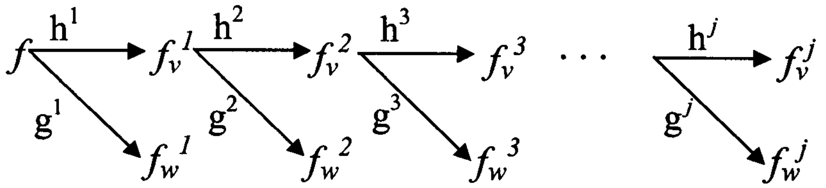
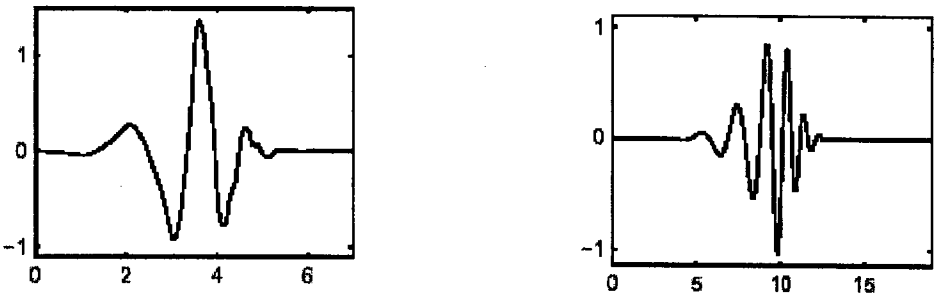


Fig. 2.3. Filter bank analysis and multi-resolution signal decomposition.

There are several families of wavelets that have proven to be useful in different applications. One of the most well-known is the Daubechies family of wavelets, which is shown by DbN, where N is the order; the greater the N, the more oscillating and smooth the wavelet. Two examples of Daubechies family, Db4 and Db10, are shown in Fig. 2.4.

A wavelet transform of a given signal can be interpreted as a decomposition of the signal into a set of components described at different frequency channels. In standard wavelet decomposition, low frequency channels have narrow bandwidth, and high frequencies have wide bandwidth [42]. Whilst this kind of signal decomposition is appropriate for many purposes there are applications that need a more flexible frequency partitioning which is the theme of next section.



Db4

Db10

Fig. 2.4. Two examples of Daubechies family of wavelets.

2.5. Wavelet Packets

Wavelet packet decomposition is an alternative and a more suitable structure for signal decomposition in a narrow frequency band data analysis. Wavelet packet can be considered as an extension of the standard discrete wavelet transform in which the outputs of high pass filters are further decomposed into high and low frequency signal components. Decomposition can be continued up to a level in which the last stage consists of single sample only. A binary tree with a root as the original signal can describe wavelet packet signal decomposition. Each node is associated with a basis function spanning signal component at that node.

While a Fourier basis provides a poor representation of functions that are highly localized in time, standard discrete-time wavelet transform is also not well suited to represent functions whose Fourier transforms have a narrow high frequency bandwidth. To solve this problem, the wavelet packet, as a generalization of standard discrete wavelet decomposition, offers a richer range of possibilities for signal analysis. Standard wavelet technique decomposes the frequency axis in dyadic intervals where size of bandwidth increases in an exponential manner [43]. Wavelet packet, introduced by

Coifman, Meyer and Wickerhauser [44], on the other hand, generalizes dyadic construction by decomposing the frequency axis in separated intervals of varying sizes.

In effect, wavelet packet is a *redundant* signal decomposition. The term "redundant" refers to the fact that there are more than one set of basis functions which can span a particular space. Non-orthogonality of wavelet basis functions at the parent/child nodes leads to a redundant signal representation. Redundancy in wavelet packet provides a wider collection of basis functions for selecting the most suitable projection directions, and can have a great influence on fault diagnosis results. Using wavelet packet we can select bases from a library of basis functions that best match the signal components at different resolution both in time and frequency.

2.6. Machine Diagnosis and Wavelets

Vibration Signals, acquired from engine operation, generally correspond to machine events that are cyclic and are often associated with a burst of high energy. They are highly transient and last only for a short period of time. Acceleration (vibration) signals acquired by sensors located at the cylinder head are of decaying oscillatory nature and can reveal information about various engine events. Such information about the condition of machine operation resides in the overall time-frequency behavior of the signal. Transient nature of machine signals and search for a particular time-frequency behavior for diagnostic purposes render wavelets as highly suitable for the analysis of such signals. Some of the reasons for the use of wavelets in machine diagnosis applications are as follows:

- **Wavelets as Time-Frequency Analysis Tools.** Wavelets are mainly time-frequency analysis tools. They are highly suitable candidates for machine data analysis as information about a given machine operation lie both in time and frequency behavior of the signal.
- **Wavelets and Localized Signal Analysis.** As stated earlier, machine data utilized for diagnosis are highly transient where information about a given machine condition reside in local behavior of the signal; i.e., changes occurring in part or the entire segment of the signal. Wavelets are highly suitable to capture localized changes and behavior.
- **Wavelet Coefficients as Feature Variables.** Signal expansion by wavelets often leads to a few wavelet coefficients of large magnitude and large number of coefficients of small magnitude. This leads to signal approximation with limited number of large amplitude coefficients used as feature variables. Considerable reduction of dimensionality is achieved in this manner.
- **Wavelets as Unconditional Bases.** Signal information lies in coefficient values obtained from wavelet signal decomposition. Wavelets are unconditional bases [41] which imply a very robust basis in which the coefficients drop off fast independent of the sign of the coefficients. Therefore, in an orthogonal signal decomposition, absolute values of the coefficients carry the necessary information about the signal. This allows to use absolute values of wavelet coefficients for feature extraction.
- **Noise Reduction using Wavelets.** Wavelets are used for noise reduction in which wavelet coefficients of small amplitude (below a given threshold) are set to zero. Often such coefficients belong to noise content of the signal at the highest

frequency band. De-noising is different from the commonly used high frequency filtering, as it can be carried-out at all frequencies. On the other hand, we utilize de-noising scheme for reduction of machine background noise corresponding to overall acceleration observed in machine data. Often it is composed of white noise. Careful selection of threshold level in de-noising can successfully reduce machine background noise.

In this work, we employ wavelets for the analysis of vibration signals of a single-cylinder engine. Wavelets are used for engine diagnosis, such as identifying knock, loose valve, and different engine operating conditions produced by various ignition timings.

First, we present the concepts of entropy and discriminant measure, which have been used as cost functions in many applications of signal representation and classification. Then, we introduce the neural network algorithm that has been used throughout the thesis as the classifier.

2.7. The Concept of Entropy

There are many useful cost functions; one of the most well-known is Shannon entropy [45]. Consider a probability distribution $\mathbf{s} = \{s_1, ..., s_n\}$ (where s_is are non-negative) associated with a random variable X, i.e., $P(X = x_i) = s_i$, for i=1,...,n. Then, entropy of the sequence is defined by the expression

$$h(X) \equiv H(\mathbf{s}) = -\sum_{i=1}^{n} s_i \log s_i \qquad (2\text{-}2)$$

where

$$\sum_i s_i = 1 \tag{2-3}$$

assuming that

$$\log 0 = -\infty\,,\ \log(s/0) = +\infty \text{ for } s > 0\text{, and } 0.(\pm\infty) = 0\,. \tag{2-4}$$

Entropy was first encountered in thermodynamics by Clausius in 1867 as a part of the second law of thermodynamics. Since the dynamic theory was unable to articulate all the collisions of the molecules causing the thermal energy, in 1872, Boltzman employed entropy to express the uncertainty of the state of the molecules of a perfect gas. In 1940s, Shanon introduced the concept of entropy in *information theory* [45] as an extension of Boltzman's idea to measure the uncertainty of information. The application of entropy was then expanded to many fields such as information theory, task planning and organization, system communication, image and signal processing, intelligent control and machine intelligence, and stochastic control [49].

By definitions (2-2) and (2-3), entropy is applied on a sequence which forms a probability density function (pdf) of a random variable. Entropy sometimes referred to as a measure of uncertainty or randomness. It can also be considered as a measure of complexity of a system [50].

Shannon entropy has several properties [51]; here three important properties are introduced.

Property 1: Entropy is always non-negative, i.e., $H(\mathbf{s})\geq 0$. Equality holds when sequence **s** follows a single-value distribution, i.e.:

$$H(\mathbf{s}) = 0 \qquad \text{if } s_k = 1 \text{ and } s_i = 0 \text{ for all } i \text{ except } i \neq k.$$

Zero entropy in *property 1* implies that the process is deterministic [51], and there is no uncertainty involved. Since **s** is a pdf then every element of **s** is less than one;

therefore, entropy is always non-negative (considering the negative sign in Eq. (2-2) and the fact that logarithm of a value less than 1 is negative).

Property 2: Maximum entropy occurs when all probabilities are equal; in other words:

$$H(\mathbf{s}) \leq H(\tfrac{1}{n}, \tfrac{1}{n}, \ldots, \tfrac{1}{n})$$

where the maximum value is log n:

$$H(\tfrac{1}{n}, \tfrac{1}{n}, \ldots, \tfrac{1}{n}) = -\sum_{i=1}^{n} \tfrac{1}{n} \log(\tfrac{1}{n}) = -n.\tfrac{1}{n}(-\log n) = \log n. \qquad (2\text{-}5)$$

From statistics and probability point of view, *property 2* suggests that if the probability of every alternative of an n-state system is identical, the entropy of the system equals to $\log n$. In special case, when size of sequence **s** is n = 2, we have:

$$H(\mathbf{s}) = H(s, 1-s) = -s \log s - (1-s)\log(1-s).$$

If **s** has an equal probability, that is, s = 1/2, and assume that logarithm base is 2, the maximum entropy will be:

$$H(\tfrac{1}{2}, \tfrac{1}{2}) = \log_2 2 = 1.$$

Before introducing the third property of entropy, let us define the *additive* concept.

Definition: Let $\mathbf{s}^{(1)} = \{s_i^{(1)}\}_{i=1}^{n} = (s_1^{(1)}, \ldots, s_n^{(1)})$ and $\mathbf{s}^{(2)} = \{s_i^{(2)}\}_{i=1}^{m}$ be two probability distributions related to independent random variables $X^{(1)}$ and $X^{(2)}$ with the joint probability distribution

$$P(X^{(1)} = x_i^{(1)}, X^{(2)} = x_i^{(2)}) = s_i^{(1)}\, s_j^{(2)}, \qquad i = 1, \ldots, n;\ j = 1, \ldots, m.$$

Then operator H (consider H as a general operator) is said to be *additive* if

$$H(\mathbf{s}^{(1)}, \mathbf{s}^{(2)}) = H(s_1^{(1)}s_1^{(2)}, \ldots, s_1^{(1)}s_m^{(2)}; \ldots; s_n^{(1)}s_1^{(2)}, \ldots, s_n^{(1)}s_m^{(2)}) = H(\mathbf{s}^{(1)}) + H(\mathbf{s}^{(2)}) \qquad (2\text{-}6)$$

Property 3: Entropy is additive because:

$$
\begin{aligned}
H(\mathbf{s}^{(1)},\mathbf{s}^{(2)}) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} s_i^{(1)}s_j^{(2)} \log(s_i^{(1)}s_j^{(2)}) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} s_i^{(1)}s_j^{(2)} \log s_i^{(1)} - \sum_{i=1}^{n}\sum_{j=1}^{m} s_i^{(1)}s_j^{(2)} \log s_j^{(2)} \\
&= -\sum_{j=1}^{m} s_j^{(2)} \sum_{i=1}^{n} s_i^{(1)} \log s_i^{(1)} - \sum_{i=1}^{n} s_i^{(1)} \sum_{j=1}^{m} s_j^{(2)} \log s_j^{(2)} \\
&= H(\mathbf{s}^{(1)}) + H(\mathbf{s}^{(2)})
\end{aligned}
$$

Eq. (2-6) implies that entropy of joint distribution of $X^{(1)}$ and $X^{(2)}$ equals the sum of entropies of $X^{(1)}$ and $X^{(2)}$.

It can be shown that for all distributions with the same variance, entropy, defined by (2-2), is maximum if sequence **s** follows a normal distribution .

Entropy has been widely used as a valuable criterion in data and image compression and signal representation. However, in classification applications, one needs a principle to measure the distance between two or more sequences. The following section pays attention to this important issue.

2.8. Discriminant Measures

The principal objective in a classification problem is to develop measures that are capable of discriminating different classes as much as possible. Accuracy of the classification results is highly influenced by the extent of class separation in feature space generated by the chosen discriminating measure. Discriminant measure, in general, is designed to evaluate the statistical distance among different classes. The choice of discriminant measure depends on the application on hand. Different authors have employed different discriminant measures in various applications [20,22,50]. The approach utilized in this work is based on using *relative entropy* as a measure for discriminating different classes that is now defined.

In a two-class case, suppose $\mathbf{s}^{(l)} = \{s_i^{(l)}\}_{i=1}^n$ for $l = 1,2$ be two non-negative sequences satisfying:

$$\sum_i s_i^{(1)} = \sum_i s_i^{(2)} = 1 \tag{2-7}$$

Symmetric relative entropy for two-class is then defined as:

$$D(\mathbf{s}^{(1)}, \mathbf{s}^{(2)}) = \sum_{i=1}^{n} (s_i^{(1)} \log \tfrac{s_i^{(1)}}{s_i^{(2)}} + s_i^{(2)} \log \tfrac{s_i^{(2)}}{s_i^{(1)}}) \tag{2-8}$$

assuming that

$$\log 0 = -\infty,\ \log(s_i / 0) = +\infty \text{ for } s_i > 0, \text{ and } 0.(\pm\infty) = 0. \tag{2-9}$$

If we just use the first term in the right hand side of (2-8) then the *relative entropy* is defined as:

$$D(\mathbf{s}^{(1)}, \mathbf{s}^{(2)}) = \sum_{i=1}^{n} s_i^{(1)} \log \tfrac{s_i^{(1)}}{s_i^{(2)}}. \tag{2-10}$$

Lemma: Equation (2-10) is always non-negative and will be zero if distributions $\mathbf{s}^{(1)}$ and $\mathbf{s}^{(2)}$ are the same.

Proof [11]: Recall the elementary inequality for real numbers

$$\log x \leq x - 1, \tag{2-11}$$

with equality if and only if $x = 1$. Then, considering conditions (2-7) we have

$$\sum_i s_i^{(1)} \log \tfrac{s_i^{(2)}}{s_i^{(1)}} \leq \sum_i s_i^{(1)} (\tfrac{s_i^{(2)}}{s_i^{(1)}} - 1) = \sum_i s_i^{(1)} - \sum_i s_i^{(2)} = 0. \tag{2-12}$$

Therefore, $\sum_i s_i^{(1)} \log \tfrac{s_i^{(1)}}{s_i^{(2)}} \geq 0$. Equality holds if and only if $s_i^{(2)} = s_i^{(1)}$, for all i.

Note: Conditions (2-7) is not necessary for a symmetric relative entropy to be non-negative, because the right-hand side of Eq. (2-12) in symmetric case results in cancellation of the corresponding terms:

$$(\sum_i s_i^{(1)} - \sum_i s_i^{(2)}) + (\sum_i s_i^{(2)} - \sum_i s_i^{(1)}) = 0\,. \tag{2-13}$$

As can be seen from the above lemma, if two random variables have the same distributions, discriminant measure D will be zero. In classification applications, one is interested in those features that can separate the distribution associated with each class; therefore, one should look for those features that maximize D. The more separate are these distributions, the higher the discriminant measure D is.

Now, the challenge is to find appropriate features. By projecting a set of data in different classes (vibration signals in our application) onto a set of basis (coordinates) and using the associated coefficients in different classes as $\mathbf{s}^{(l)}$, the corresponding discriminant measure D can be found. In a classification problem, the objective is to locate those bases (out of a dictionary of bases) that maximize the value of D. In this manner, the distribution of each class can be transformed to a sharp distribution with least overlap with other distributions.

The discriminant function D measures how differently distributions of two classes are. Since $\mathbf{s}^{(l)}$ are non-negative sequences assimilating pdf functions, we can successfully employ the normalized energy of the coefficients in each class as $\mathbf{s}^{(l)}$ and $\mathbf{s}^{(2)}$ in the classification problem.

In general, if there are L classes, one can use this simple approach:

$$D(\{\mathbf{s}^{(l)}\}_{l=1}^{L}) \equiv \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} D(\mathbf{s}^{(i)}, \mathbf{s}^{(j)})\,. \tag{2-14}$$

Equation (2-14) is not always the best choice when the number of classes is more than three, since a large value for D may be obtained because of the summation of many small terms, which is not a favorable outcome. In fact, the desirable situation for the classification is a large D due to a few significant terms with most others having negligible values. To overcome this shortcoming, Watanabe and Kaminuma [52] split training data into class i and non-class i and then form L sets of two-class problems and construct a classifier for each of them.

Relative entropy and its symmetric version have been used in chapters 5 to 7.

2.9. Neural Network Algorithms

Backpropagation algorithms, that use a multilayer perceptron network [71], have been extensively used in classification applications. Backpropagation is an extension of least mean square (LMS), which in turn is an approximation of the steepest descent algorithm [38] that is a simple, but a slow minimization method.

There are two different approaches to improve the convergence speed of backpropagation learning rule [53]. The first approach employs heuristic methods such as varying the learning rate to determine the length of steps in the steepest decent algorithm. The other approach focuses on numerical optimization techniques since training feedforward neural networks to minimize square error is basically an optimization problem. As a result of latter approach, which has been the focus of research for many years, several variations of the main backpropagation algorithm have been developed. For example, conjugate gradient algorithm and Newton's method provide faster convergence. In these methods and many other backpropagation algorithms, in which goal is the optimization of a performance index, errors of network with respect to target values are found from the last layer of the network to the first. The word "backpropagation" refers to this process. The improved backpropagation training algorithms have various computation and storage requirements.

Even though there are some heuristics and general guidelines for the selection of a specific learning method, such as convergence speed or storage needs, usually, a training algorithm is chosen empirically for each application. We examined 12 different backpropagation training algorithms in which Levenberg-Marquardt algorithm was found to be the best algorithm with respect to both high classification accuracy and computational efficiency. Classification errors for each of these 12 algorithms for the Ricardo Hydra engine data (introduced in the next chapter) are shown in Table 7.1 in chapter 7.

In the next two chapters the attention will be on two methods for fault diagnosis based on wavelet analysis: mutual information and dictionary projection pursuit.

2.10. Closing Remarks

In this chapter, wavelets – as the class of tools we would like to examine in this research – were introduced, along with a formal definition of the concept of entropy and its extension, the relative entropy. The importance of feature extraction in pattern recognition applications were emphasized and stated that data must be processed for dimension reduction to degrade the effect of information dilution (due to both redundant and unrelated information contained in the original data) and to decrease computational time. Reviews of literature in the signal processing and fault detection area as well as the various approaches researchers have developed were presented.

CHAPTER 3

Experimental Setup, Data Collection and Preparation

3.1. Introduction

In Chapters 1 and 2, wavelets and wavelet packet transform including the basic theory and potential capabilities in pattern recognition applications in general, and in machine diagnosis in particular, were presented. In the coming chapters a more detailed and in depth analysis of a number of wavelet-based methodologies used in pattern recognition will be introduced. In this chapter, various sets of data that will be used for the evaluations of these methodologies are discussed.

Experimental data used in this thesis are as follows:

- Acceleration data acquired at the cylinder head position of a single cylinder research engine, namely Ricardo Hydra,

- Acceleration data acquired at the cylinder head position of a mid-size single cylinder diesel engine, and
- A set of synthetic data.

It should be mentioned that main experiments were conducted and reported on actual engine data. Synthetic data were used for test and validation and as a secondary source.

To use the real engine data for the analysis, it was necessary to carry out data preparation and preprocessing as described below.

3.2. Preprocessing and Normalization

Data preprocessing is an integral part of statistical pattern recognition. Several preprocessing schemes have been proposed in the literature [53], which were examined in this thesis. Effectiveness of a given scheme in a particular application is highly case-dependent and is influenced by the requirements of a given application. Usually an appropriate preprocessing method should be applied on both signal space and feature space (please refer to Fig. 2.1). The role of preprocessing is to remove bias and disproportionate differences in data so that a meaningful comparative analysis of the results can be made. Normalization of data is a logical technique that may be done through the following three options:

1- Normalization based on mean-centering and norm of every data:

$$data = data - \mathrm{mean}(data) \tag{3-1}$$

$$data = data / \mathrm{norm}(data) \tag{3-2}$$

This normalization has been used in [20] and [22].

2- Individual class normalization in which data are mean-centered and each class is normalized to its Euclidean norm, i.e., to the sum squared of the elements of all data in the given class:

$$data = data - \mathrm{mean}(data)$$

$$data = data / \sum_{i=1}^{N_l} \mathrm{norm}(data_i^{(l)}) \qquad (3\text{-}3)$$

where N_l is the total number of data in the class l. This normalization has been used in [24].

3- Normalization of data in each class to confine the range of data to -1 and +1. This is done by the following transformation:

$$d_N = s.d + b$$

where

$$s = 2/(max - min);\ b = -(Max{+}Min)/(Max - Min).$$

d_N can also be written as :

$$d_N = \frac{(d - Max) + (d - Min)}{Max - Min} \qquad (3\text{-}4)$$

in which d and d_N are actual data and normalized data respectively, "*Max*" and "*Min*" are the maximum and minimum values of all of data.

In our study, option 3 for the normalization of data in both signal and feature domains has been used. Under this scheme, the relative position of data in different classes with respect to each other remains unchanged. Analysis and further discussion on the application of this scheme and the effect of preprocessing are presented in chapter 7.

In the following sections the experimental setups of the sets of data, which will be used throughout the thesis, are presented.

3.3. Ricardo Hydra Engine Experimental Setup

Ricardo Hydra, a single cylinder spark ignition research engine, was used for conducting a set of experiments and test runs. The engine operates on both gasoline and natural gas fuel modes, but only natural gas mode with a compression ratio of 9.26:1 was used. The engine speed and throttling were set at 1500 RPM and 100% (wide open), respectively. Different machine operations with three relative air/fuel ratios namely stoichiometric ($\lambda = 1$), fuel lean ($\lambda = 1.5$), and fuel-rich ($\lambda = 0.8$) mixtures, each with normal, advance, and retard spark timing were obtained. One pressure sensor with 12.5 KHz and two accelerometers with 25 KHz and 12.5 KHz sampling rates were employed to measure cylinder pressure and simultaneous vibrations at two positions on the cylinder head in vertical (V) and horizontal (H) directions. A rotational encoder was used to monitor engine speed and to determine the starting point of each cycle. Data belonging to $\lambda = 1$ condition were utilized in this work. To realize the effect of combustion on cylinder head vibration and pressure signals, data were also collected in the absence of combustion (no ignition) by externally driving the engine. This is referred to as the *motoring mode*.

3.3.1. Data Preparation

The objective of the respective experiment was to collect acceleration data at the cylinder head position with three different ignition timings of: -23 (normal), -33 (advance), and -10 (retard) degrees under stoichiometric conditions assimilating healthy and faulty conditions. The numbers denote ignition timings measured as angles before top dead center. A data acquisition system from REM Technology Inc. [1] was used for data collection. This system was capable of acquiring data of 16 consecutive cycles of engine operation. For each of the engine conditions (class), three consecutive 16-cycle data were collected to provide sufficient number of data (48 sets) for the implementation of the algorithms introduced in the following chapters, and to evaluate their accuracy of classification results. We used the first 32 data cycles in each class as *training* and the remaining 16 as *testing* data sets.

The size of data for one complete engine cycle corresponding to two revolutions of crank shaft (720 degrees of crank angle) varied for different cycles and was 1975±15 sample points for data collected at a sampling rate of 25 KHz and 988±8 sample points for a sampling rate of 12.5 KHz. The variation in number of samples, caused by changes in engine RPM, was found to be insignificant (less than 1%). In this experiment, our concern was to study combustion event under different engine operating conditions as affected by different ignition timings. As such, data belonging to combustion zone only were considered for the analysis. This was done by defining a window of data, the size of which was carefully selected for investigating different ignition timings. The window was to be wide enough to cover all of the characteristics of the combustion event under different advance or retard spark timing conditions, but not too wide to overlap with other engine events.

Using a small window size was considered to be important to obtain low computational time. Here we use a segment of data belonging to -15 to +31 degrees of crank angle which covers the combustion event. For the acceleration data, this corresponds to data points ranging from 945 to 1072 for 25 KHz sampling rate and one half length for the 12.5 KHz sensor data. Accordingly, the size of data for 25 KHz sensor was 128 sample points, and for the 12.5 KHz sensor it was 64. The maximum heat-release of the engine as a measure of combustion quality, occurred at (-5.5) – (32), (-14.5) – (26.5), and (9) – (47) degrees of crank angle for -23, -33, and -10 degrees of ignition timing, respectively. The numbers in parenthesis correspond pairwise to the 5% and 95% of heat-release, which shows the amount of chemical energy of the fuel released by the combustion process at the specific crank angle degree [2]. With regard to the heat release and practical considerations where we need an identical interval for all of data categories, the above window size of -15 to +31 was found to be an appropriate selection.

Fig. 3.1 shows sample acceleration data acquired by the sensor in vertical direction with -23, -33, and -10 degrees of crank angle. We refer to these data as class 1, class 2, and class 3 engine conditions, respectively. The sections of acceleration signal from left to right with high magnitudes correspond to different machine events i.e. exhaust valve closing (EVC), intake valve closing (IVC), combustion, exhaust valve opening (EVO), and intake valve opening events (IVO), respectively.

Fig. 3.2 shows the Ricardo Engine valve timings in which TDC and BDC stand for top and bottom dead centers. The numbers above each event indicate crank angle in degrees and the corresponding sample point with respect to the TDC and BDC, respectively. For example, -(+)56 means 56 crank angle degrees before (after) BDC. Fig. 3.1 demonstrates that vibration signals are also affected by other events during engine operation. For instance, intake valve closing within the range of 650 and 820 data points shows two components of high amplitudes. However, comparison of figures 3.1 and 3.2 reveals that only the first high amplitude segment belongs to IVC; the second spike is generated by other machine components oscillating at their harmonic frequencies.

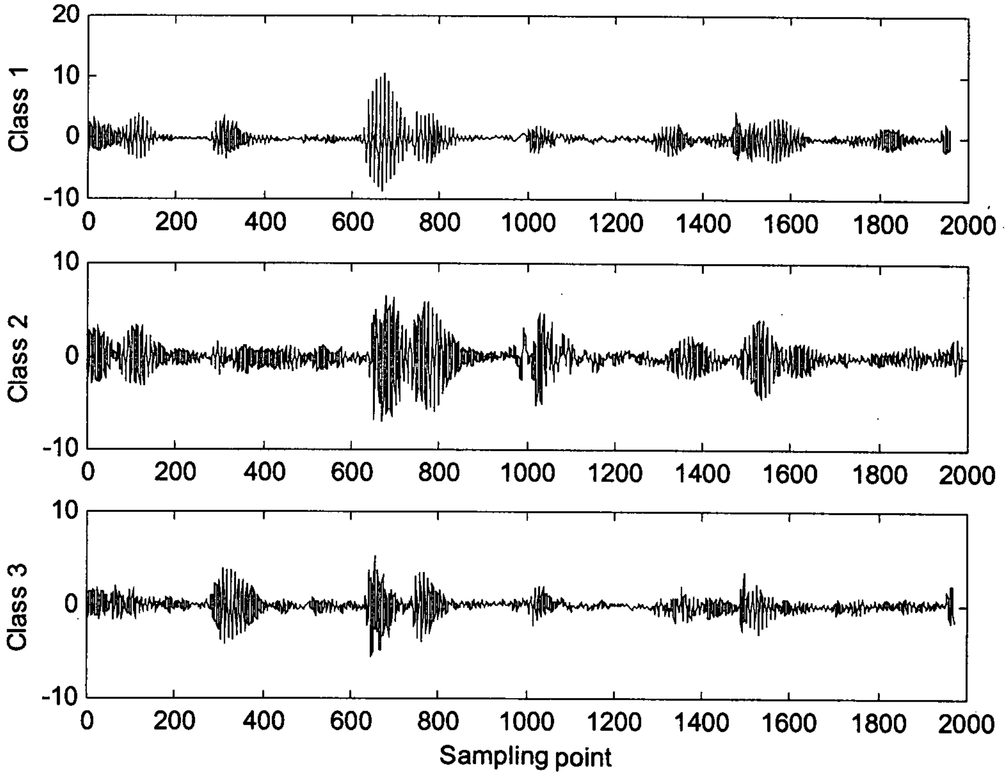


Fig. 3.1. One cycle of three classes of vertical vibrations with spark timings of: -23, -33, and -10 degrees of crank angle, in stoichiometric conditions, 1500 RPM, and 25KHz sampling rate. Vertical axis unit is "g".

Fig. 3.3 shows variations of in-cylinder pressure for the three classes of collected data. Using pressure signal, classes are distinctly differentiated from each other, indicating the usefulness of pressure signals for identifying different classes in engine diagnosis as compared with acceleration data. However, in practical applications using pressure transducers is neither feasible nor economical.

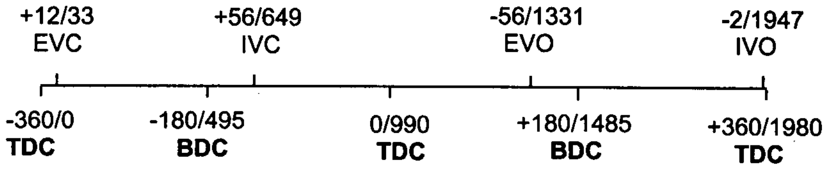


Fig. 3.2. Ricardo Research Engine valve timing characteristics.

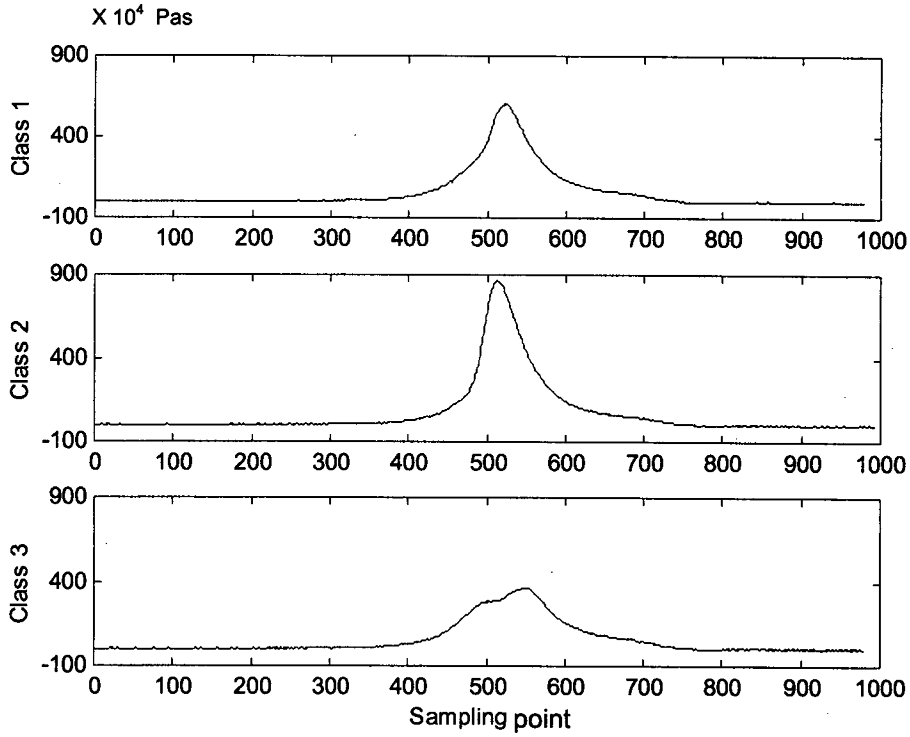


Fig. 3.3. Pressure signals for three classes of spark timing: -23, -33, and -10 degrees of crank angle and with 12.5 KHz sampling rate.

3.3.2. Data Analysis

Analysis of acceleration data indicates that horizontal vibration data do not carry useful information about the combustion event. The Fourier transform of horizontal vibration data sampled at 12.5 KHz (the solid graph in Fig. 3.4) indicates large spectral amplitudes in high frequencies as compared with vertical vibration data (at 25 KHz sampling rate) where high frequency components rapidly decay to zero (Fig. 3.5). An inspection of the spectrum (FFT) of vertical acceleration data at low sampling rate (the dotted graph in Fig. 3.4) exposes the nature of horizontal vibration data. They indicate that the reason horizontal accelerometer data do not carry the same valuable information is not due to the low sampling rate. We see that the dominant frequency band in vertical (2-4 KHz) and horizontal directions are not the same demonstrating that the information extracted from the horizontal direction is not related to the combustion event but are from other sources classified as noise interferences.

Spectral analysis of data for three classes indicates that frequency bandwidth with high spectral energy is almost identical for the all classes with some variations in spectral amplitudes (Fig. 3.5). This indicates that the spectral features cannot be directly used for the discriminatory classification of different engine conditions, specially for classes 1 and 3. An inspection of acceleration data showed that the engine was running relatively smoothly with minor background noise where noise reduction was not actually necessary. Investigating both vibration and pressure data indicated that engine operation was also relatively uniform with a low variability in the data among different cycles. Spectral characteristics shown in Fig. 3.5, however, will be exploited in chapter 7 in a different context.

A histograms of data indicates that distribution of the training data (Fig. 3.6) follows approximately the Gaussian distribution. Classes 1 and 3 exhibit highly clustered data patterns when the mean value of data in each class is drawn against their standard deviation.

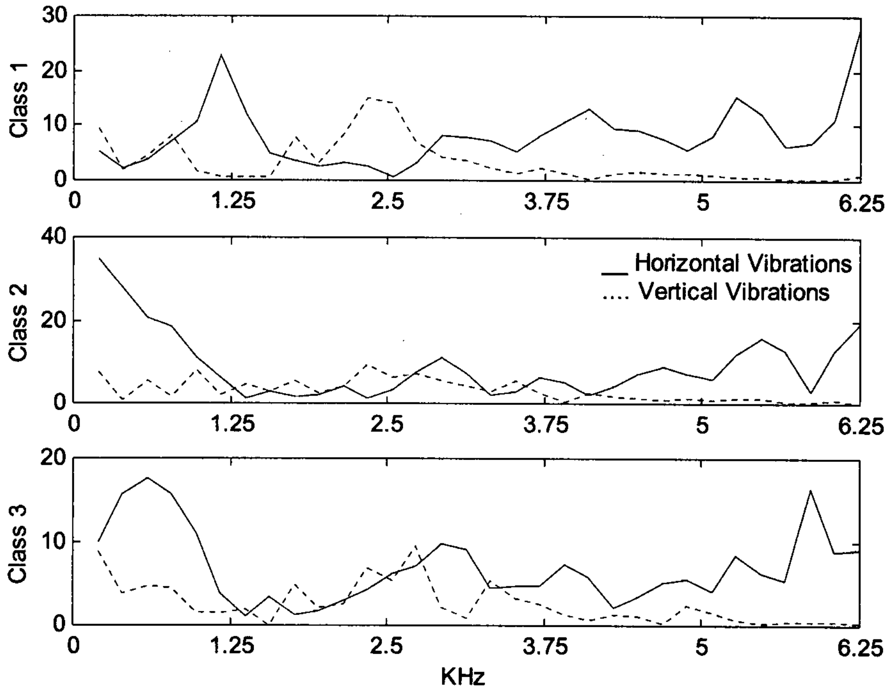


Fig. 3.4. FFT of horizontal and vertical vibrations with 12.5 KHz sampling rate, in combustion zone, three classes of spark timing: -23, -33, and -10 degrees of crank angle. Vertical axis unit is “g”.

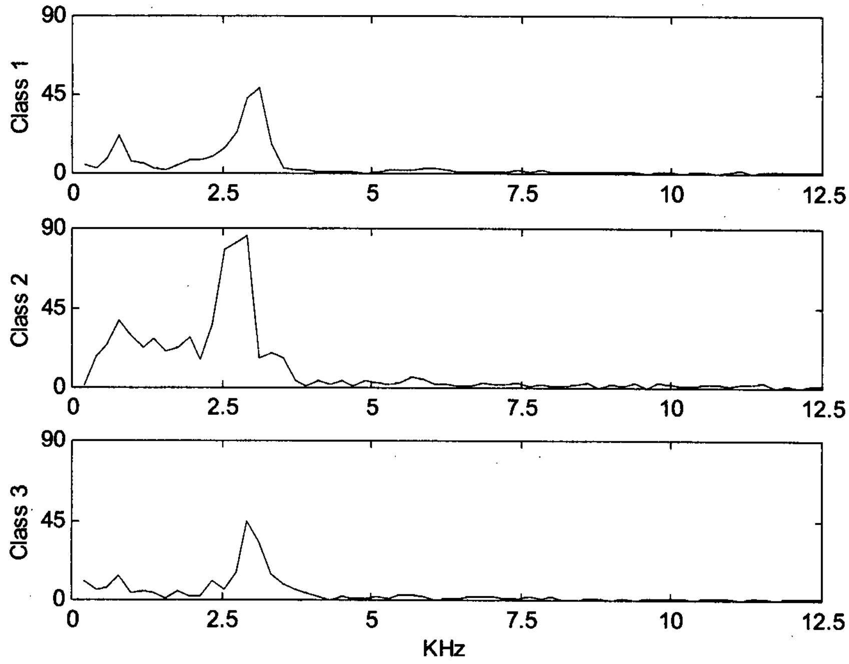


Fig. 3.5. FFT of vertical vibrations in combustion zone with 25 KHz sampling rate and three classes of spark timing : -23, -33, and -10 degrees of crank angle.

3.4. Mid-size Single Cylinder Diesel Engine Setup

The second test engine was a single cylinder dual mode unit operating either on diesel fuel or natural gas. Data presented here is from diesel mode operation. Acceleration data of the intake valve closing and combustion events from this engine were utilized for data analysis and algorithm testing. Two types of faults, namely intake loose valve and engine knock conditions each with varying intensity levels were considered. Engine knock condition was generated by judicious adjustment of load. Load changes were made in two incremental steps of approximately 15% above nominal load corresponding to 18, 22, 25 HP, respectively.

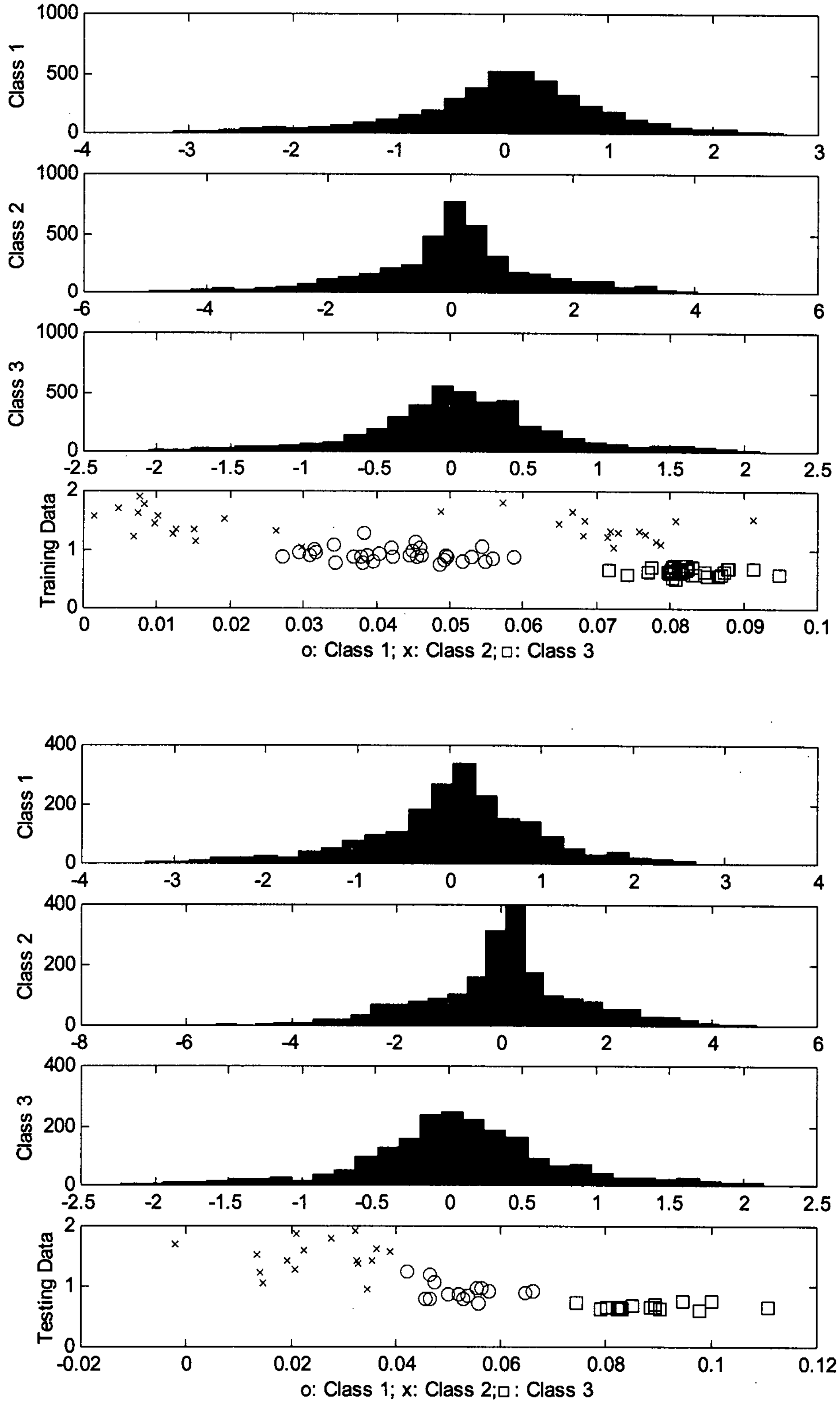


Fig. 3.6. Histograms and mean-std plots of training and testing data for three classes.

For loose valve experiments, a set of progressively increasing valve clearances, namely normal, 0.006 in. and 0.012 in. were set on the intake valve. Three categories of data were collected simultaneously: (a) cylinder pressure measured through a connecting tube to the cylinder, (b) block acceleration (vertical vibration) measured at a carefully chosen location on the cylinder head, and (c) engine RPM. Block vibration was actually measured at several places and the best location was found to be at the center of the upper part of the cylinder block which gave reliable signal intensities. Other supplementary data were also collected including engine power, peak cylinder pressure and peak pressure angle. For each test, data from sixteen consecutive cycle runs were acquired.

Fig. 3.7 shows sample cycle runs of the diesel engine in normal and knock conditions. In this figure, the high amplitude components from left to right correspond to exhaust valve closure, intake valve closure, combustion, exhaust valve opening, and intake valve opening. There were noticeable cycle-to-cycle changes in the signal patterns and intensities even under normal condition, which indicate the complexity and variability of the machine operation. An initial review of data, in which mean values vs. standard deviation of each training data were examined, showed that a certain degree of data clustering and class separation can be found (Fig. 3.8), though this could not be observed in all of the data sets. Separation of classes was more vivid in training data belonging to valve clearance conditions.

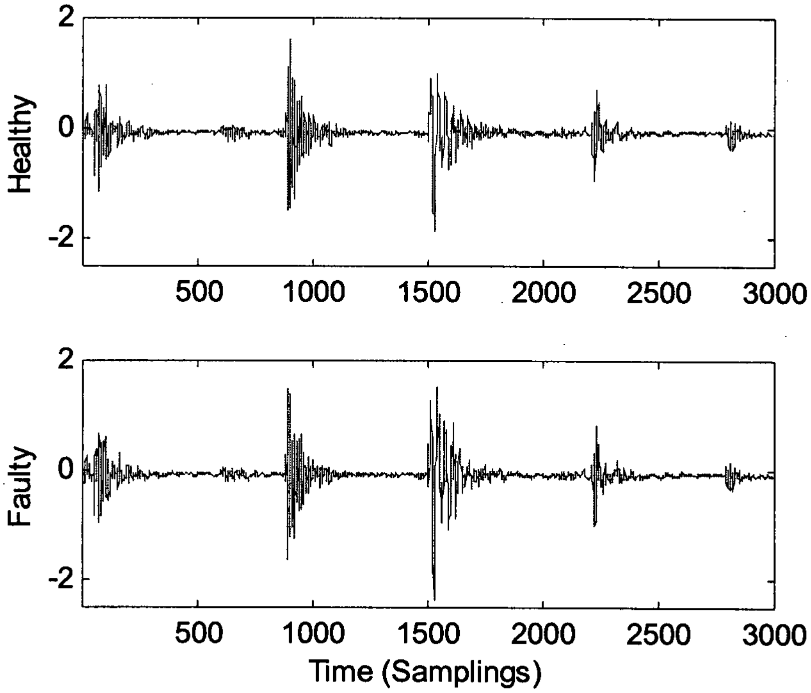


Fig. 3.7. Sample vibration signal of the single-cylinder diesel engine in normal and knock conditions.

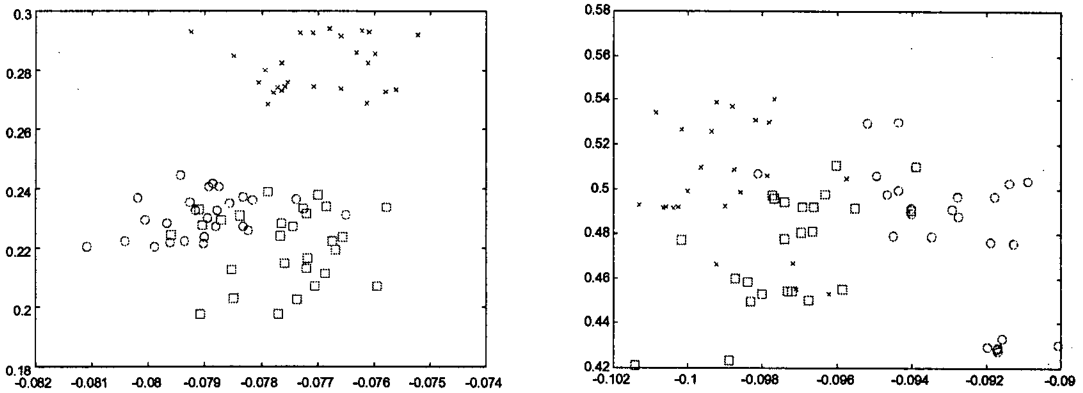


Fig. 3.8. Training data for valve clearance (left) & load change (right) for three classes of □: normal, x: mild fault, o: severe fault

In the data analysis, 28 training data were used in each of the three classes for both loose valve and knock conditions. Initially, two classes consisting of normal and one faulty condition were examined. At a later stage, we considered three classes of healthy, mild faults, and severe faults for two different intensity levels.

3.5. Synthetic Data

As a general example a set of synthetic data is also used in this thesis. This type of data, proposed in [54], has been very popular in waveform recognition and classification studies. We adopted a sample data from [22], in which because of dyadic dimensionality requirement in wavelet analysis the number of data points had been extended from 21 to 32. It is a three-class data based on triangular waveforms $h_1(t)$, $h_2(t)$, and $h_3(t)$ defined as

$$h_1(t) = max(6 - |t - 7|, 0)$$
$$h_2(t) = h_1(t - 8)$$
$$h_3(t) = h_1(t - 4)$$

where t = 1,..., 32. Each class of signal $\boldsymbol{x}$ is composed of a combination of two of the above triangles in addition to a uniform random number u over the interval (0,1), as well as 32 normally distributed random numbers ε (1),..., ε (32), with zero mean and unit variance. Three classes of synthetic signals, $x^{(1)}$, $x^{(2)}$, and $x^{(3)}$, are then generated by:

$$x^{(1)}(t) = u\, h_1(t) + (1 - u)\, h_2(t) + \varepsilon(t)$$
$$x^{(2)}(t) = u\, h_1(t) + (1 - u)\, h_3(t) + \varepsilon(t)$$
$$x^{(3)}(t) = u\, h_2(t) + (1 - u)\, h_3(t) + \varepsilon(t)$$

For each class 30 measurement vectors (data) were generated.

3.6. Closing Remarks

In this chapter the importance of data preprocessing for pattern recognition and classification was discussed and a number of methods were introduced. We gave a perspective on various sets of data that will be widely utilized throughout this thesis. Different concepts and parameters of machine data, data collection, and data preparation were presented, along with some necessary discussions on data representation and analysis.

CHAPTER 4

Mutual Information and Informative Wavelet

4.1. Introduction

This chapter deals with an application of wavelets for feature extraction and classification of machine faults. The statistical approach referred to as *informative wavelet* algorithm is utilized to generate wavelets and subsequent coefficients that are used as feature variables for the classification and diagnosis of machine faults. Informative wavelets are referred to classes of functions generated from elements of a dictionary of orthogonal bases, such as wavelet packet dictionary. Training data are used to construct probability distributions required for the computation of the entropy and mutual information. In our data analysis, we have used machine data acquired from a single cylinder engine under a series of induced faults in a test environment. The objective of the experiment was to evaluate the performance of the informative wavelet algorithm in classifying faults using real-world machine data and to examine the extent to which the results were influenced by different analyzing wavelets chosen for data analysis.

The correlation structure of the informative wavelets as well as coefficient matrix are also examined.

4.2. Informative Wavelets - Concept and Approach

Informative wavelets are classes of functions generated from a given analyzing wavelet in a wavelet packet decomposition structure in which for the selection of 'best' wavelets, concepts from information theory, i.e., mutual information [20] and entropy [8] are utilized. Entropy is a measure of uncertainty in predicting a given state of a system where a system state refers to different operating conditions such as normal or faulty. Computation of entropy requires calculating state probabilities from training data and supplying them as inputs to the algorithm. An iterative process to identify appropriate informative wavelets is used at each stage, whereby the algorithm selects a wavelet from a dictionary of orthogonal wavelets in a wavelet packet signal decomposition structure, which results in a maximal reduction in entropy. This is equivalent to obtaining maximal reduction in uncertainty of predicting a given system state. In this algorithm, reduction in uncertainty is expressed in terms of mutual information derived from the joint probability distributions of the training data and coefficients. Entropy of a system is defined as:

$$H(S) = -\sum_{i=1}^{M} P(S_i)\log(P(S_i))$$

where $S_1, S_2, \ldots, S_M$ are the states of the system with probability of occurrences given by $P(S_1)$, $P(S_2)$, …, $P(S_M)$. Entropy is a measure introduced for the quantification of the information. It can also be considered as a measure of complexity of prediction of the state of the system. The reduction in uncertainty can be regarded as the *quantity* of information about the original system contained in the measurement system, which is referred to as mutual information [51].

To derive the mathematical definition of mutual information we need to describe the states of the system. Such states can be observed by a measurement system with N possible outcomes $\{T_1,T_2,\ldots,T_N\}$ of a random variable T with a probability distribution $P(T_1)$, $P(T_2)$,..., $P(T_N)$. Mutual information between the states and measurements is defined as the difference between the uncertainty of predicting S before and after the observation of T:

$$J_S(\omega_\gamma) = H(S) - H(S/T) = \sum_{i=1}^{M}\sum_{j=1}^{N} P(S_iT_j)\log\frac{P(S_iT_j)}{P(S_i)P(S_j)}.$$

Here $H(S/T)$ and $P(S_iT_j)$ indicate conditional entropy of state S given measurement T and joint probability distribution of $S = S_i$ and $T = T_j$, respectively. ω_γ is the wavelet indexed by the triplet parameter $\gamma = (j, k, m)$, where j, k, m are the indices of scale, oscillation, and translation (time position) in a wavelet packet dictionary. When a given state of a system is independent of the measurements, i.e. $J_S = 0$, a change in the state of the machine will not cause any changes in the probability $P(S_iT_j)$. Then the algorithm selects wavelets that result in a maximal reduction of uncertainty i.e. maximal $J_S(\omega_\gamma)$. In informative algorithm, the measurement system is wavelet. Such wavelets are obtained iteratively where at each stage, the residual signal is considered for further signal expansion. These wavelets are referred to as *informative wavelets*. The iterative selection of the informative wavelets is very much similar to the classical matching pursuit algorithm [26]. Wavelet coefficients are then used as feature variables and as inputs to a neural network classifier for classification [20]. Fig. 4.1 illustrates the main stages of the algorithm. The next section explains different steps of informative wavelet algorithm in more details.

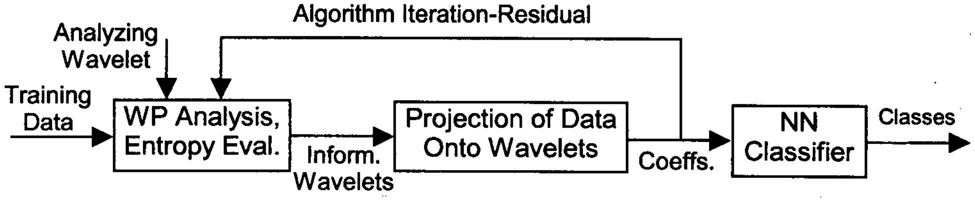

Algorithm Iteration-Residual
Analyzing Wavelet
Training Data
WP Analysis, Entropy Eval.
Inform. Wavelets
Projection of Data Onto Wavelets
Coeffs.
NN Classifier
Classes

Fig. 4.1. Block diagram of main stages of informative wavelet algorithm.

4.3. Informative Wavelet Algorithm

The algorithm has two stages: *training* stage and *class recognition* stage. In the training stage, Fig. 4.2, by using wavelet filters, training data are decomposed into low and high frequencies iteratively to form a wavelet packet (WP) for each training data. Then the collection of these wavelet packet decomposition coefficients is quantized into N fixed and equally-spaced sub-intervals. At this step probability distributions of S, T and joint probability distribution of S and T are obtained. Each wavelet is considered as a measurement system whose output is its decomposition coefficients obtained by projecting data onto the selected wavelet. These wavelet coefficients, which are in fact feature variables, are later fed to a neural network to classify the system state. Using the maximum mutual information ($J_S(\omega_\gamma)$) its corresponding informative wavelet is then selected. In the next step, the corresponding wavelet components are deducted from the residuals of entire training data, much in the same way as in the matching pursuit algorithm. As the informative wavelets are successively selected from these residuals at each iteration, they are less correlated with the ones selected previously.

At the final stage, the coefficients obtained above are used to train the neural network. Once the training is completed the NN weights are attained in order to memorize the main features of different classes. If three classes are considered, these can be, for example, severe fault, mild fault and healthy states.

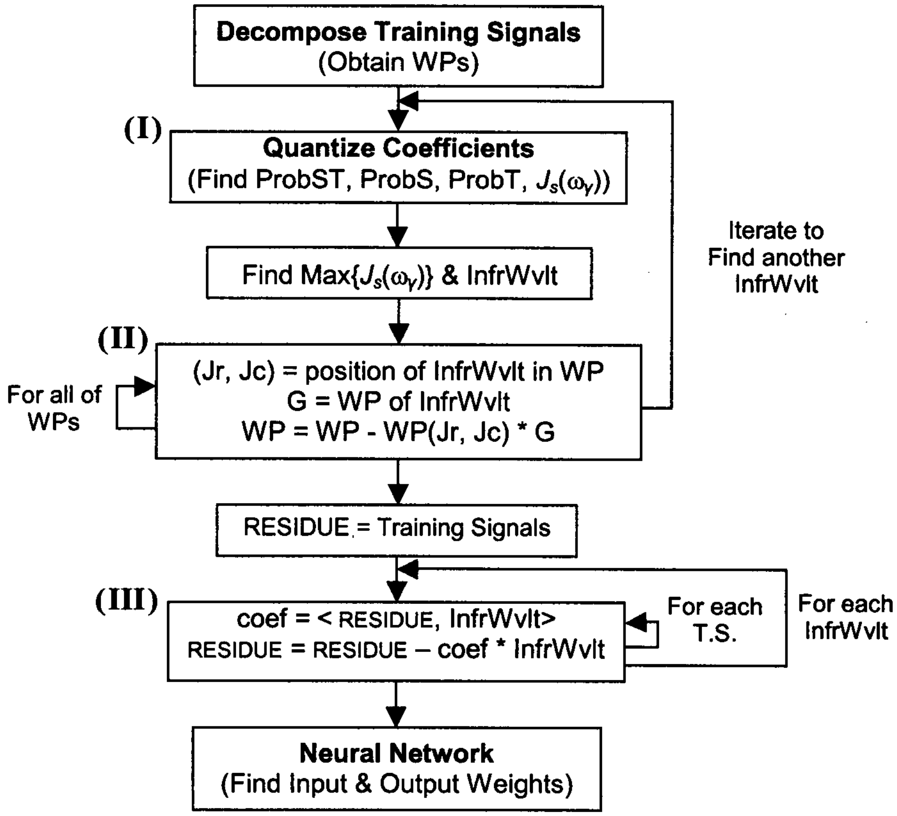

Decompose Training Signals
(Obtain WPs)
(I)
Quantize Coefficients
(Find ProbST, ProbS, ProbT, $J_s(\omega_\gamma)$)
Find Max{$J_s(\omega_\gamma)$} & InfrWvlt
Iterate to Find another InfrWvlt
(II)
For all of WPs
(Jr, Jc) = position of InfrWvlt in WP
G = WP of InfrWvlt
WP = WP - WP(Jr, Jc) * G
RESIDUE = Training Signals
(III)
coef = < RESIDUE, InfrWvlt>
RESIDUE = RESIDUE – coef * InfrWvlt
For each T.S.
For each InfrWvlt
Neural Network
(Find Input & Output Weights)

Fig. 4.2. Informative wavelet algorithm: training stage.

The input signal along with informative wavelet and neural network weights obtained from the previous stage are inputs to the second or class recognition stage. This stage consists of three steps: projecting the input signal – that we want to identify its class – onto selected informative wavelets, computing their feature vector, and classifying the state of machine (S_2). Fig. 4.3 shows the flowchart of this stage. This algorithm attempts to match joint state and measurement probability distribution of data with wavelet coefficients, the higher the probability the more the mutual information.

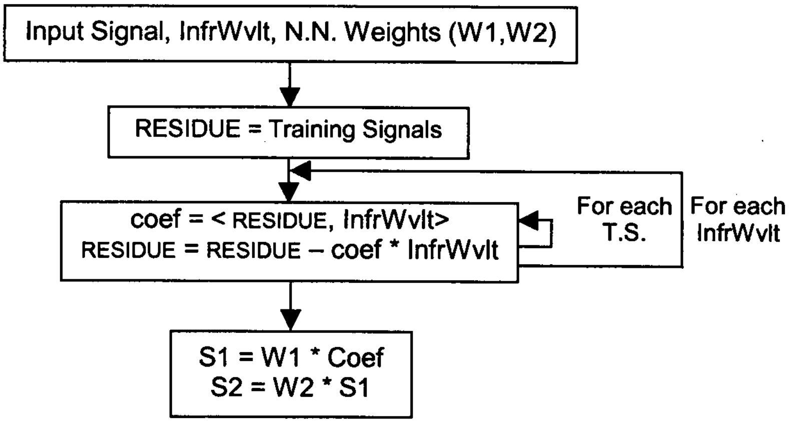

Input Signal, InfrWvlt, N.N. Weights (W1,W2)
RESIDUE = Training Signals
coef = < RESIDUE, InfrWvlt>
RESIDUE = RESIDUE – coef * InfrWvlt
For each T.S.
For each InfrWvlt
S1 = W1 * Coef
S2 = W2 * S1

Fig. 4.3. Informative wavelet algorithm: classification stage.

The major disadvantage of informative wavelet algorithm may be its computational complexity. The computational time for box (I) is $O(N)$, where N is the total number of training data in all classes. Since the loop of this box must iterate for the whole wavelet packet elements ($n\ log_2\ n$ times), and for the number of informative wavelets (W), the total cost is $O(WNn\,log_2\,n)$. This algorithm relies on the evaluation of probability density of training data; consequently, we usually need several training data. An empirical number is about the size of data (n), therefore, the total computational cost is $O(W n^2\ log_2\ n)$. If the time for decomposing each training data to wavelet packet coefficients is also added, i.e., $O(n\,log_2 n)$, along with other overhead computations, which is not insignificant in this algorithm, the real time cost will approach $O(n^3)$. We note that since probability density function must be evaluated in every iteration, calculation of probability density function is the most time consuming part of the algorithm.

4.4. Design of Experiments

To evaluate the performance of the algorithm and the accuracy of its classification results we used machine data from the single cylinder engine introduced in section 3.4. Two types of faults, namely engine knock, and loose intake valve conditions each with varying intensity levels were considered.

We note the following in the analysis:

- In informative wavelet algorithm, the "number of informative wavelets" corresponds to the number of feature variables used for the classification. In the absence of any á priori knowledge about a suitable number of feature variables, several values ranging from 1 to 50 were initially considered. At a later stage, the number was confined to a smaller set ranging from 4 to 10.
- Wavelets from orthogonal and biorthogonal wavelet families were used including Daubechies wavelets Db5, Db20, Db40 and Db45 as well as Coif5, Symlet5, Bior3.1, and Bior6.8.
- Multi layer perceptron backpropagation was used for the neural network classifier. For a three-class data set, five nodes of hidden layer were used in the network.
- We used 30 levels (bins) in quantification of coefficients and training data during construction of the probability distributions.

4.5. Data Analysis and Classification

As indicated earlier the informative wavelet algorithm is mainly a statistical approach for fault detection and classification in which probability distributions of training data are utilized to generate wavelets during signal expansion. In this algorithm, coefficients of the selected wavelet carry statistical properties that best matched those of the training data.

At the first glance, it may seem that classification results are determined jointly by capturing the statistical properties of the given training data as well as the analyzing wavelet used for data expansion. But our observations using different data and with several analyzing wavelets showed that the former has a higher influence on the classification results. In fact, different analyzing wavelets capture more or less the same amount of statistical information; therefore, the choice of analyzing wavelet does not significantly alter the correlation structure of coefficients, although Coiflet1 wavelet performed marginally better.

Using Coiflet1 we analyzed three load settings (leading to knock) as well as three valve clearance conditions. Mean values vs. standard deviations of the coefficients of training data for three classes as well as histogram of the coefficients were also examined (Fig. 4.4). Separation of classes in coefficient domain followed a similar pattern as those of training data. For both fault cases, classification errors were below 5%, which were considered to be acceptable. Classification errors for different load changes and knock conditions were influenced to a large extent by the uniformity of the training data in all classes.

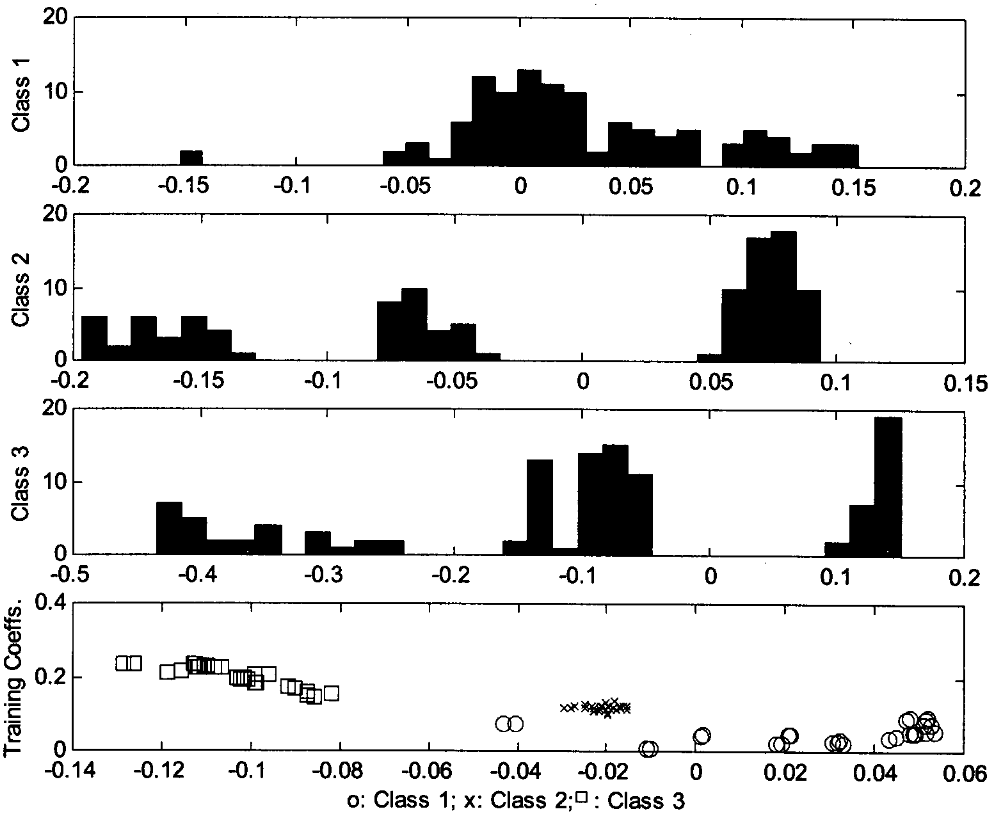


Fig. 4.4. Histograms of the coefficients as well as mean vs. standard deviations for three classes.

4.5.1. Selected Informative Wavelets

Informative wavelet algorithm is a nonorthogonal signal decomposition in which informative wavelets generated by the algorithm are in general correlated with each other and a certain degree of redundancy always exists in signal decomposition. Accordingly, the coefficients generated by projection of data onto informative wavelets follow the same pattern of correlation. Non-orthogonality of signal decomposition is mainly due to the iterative process of selecting informative wavelets where at each stage the residual signal is constructed and used for signal expansion. In our data analysis, we examined deviation from the orthogonality of the informative wavelet for several analyzing wavelets. We examined informative wavelets generated by orthogonal and biorthogonal analyzing wavelets. While informative wavelets in both categories deviated from orthogonality, which was measured by the inner product of the wavelets, the extent of the deviation varied for the two groups. Orthogonal wavelets such as Db family of wavelets or Coiflets, generate informative wavelets with a higher degree of orthogonality as compared with biorthogonal wavelets such as Bior3.1. The same trend can be seen in the correlation structure of coefficient matrix as well.

Correlation structure of coefficient matrix under several analyzing wavelets and for different number of iterations was examined for a given set of data. Differences were observed in correlation of the coefficients for different analyzing wavelets; however, such differences were insignificant to influence the classification results greatly.

4.5.2. Training Data and Number of Iterations

In our data analysis, a small change in training data resulted in a noticeable change in the informative wavelets selected. For example, a small increase in the number of training data (e.g. a simple repetition of data) caused a different set of informative wavelets to be selected. This could be attributed to the application of matching pursuit type approach in which a small change in the probability distribution of the coefficients leads to changes in mutual information value calculated. Often a small change in the training data caused a change in about half of the informative wavelets.

In the algorithm, the number of informative wavelets (iterations) is chosen á priori as an input. It was observed that increasing the number of iterations in a given data analysis does not alter informative wavelets derived from previous iterations. As a result, there were no changes in the corresponding coefficient values. The additional informative wavelets, selected with larger number of iterations, increased the number of feature variables and thus expanded the dimension of feature space.

In the experiments, mostly 5-10 iterations were used, although higher iterations were also selectively examined. It was observed that an increase in the number of iterations was not always accompanied by an increase in the accuracy of classification results. This could be traced to dilution of information, in which by selection of large number of features unnecessary information is added.

4.6. Conclusions

In this chapter results of an experimental study for an application of informative wavelet algorithm for the classification and diagnosis of machine faults were presented. Several prototype wavelets and different sets of machine data were used. Effectiveness of the algorithm for the classification of two categories of faults namely excess valve clearance and knock conditions each with varying intensity levels were examined. Accuracy of results under different parameters of the algorithm was also studied by employing different analyzing wavelets from both orthogonal and biorthogonal family of wavelets. Some notable results are summarized as follows.

- In majority of the experimental runs, using different analyzing wavelets, satisfactory classification results were obtained when sufficiently large number of training data with adequate uniformity was used. For load changes and knock condition, accuracy of results varied for different training data and different intensity levels of fault conditions.
- Informative wavelets generated by the algorithm varied significantly when small changes were introduced in the number of training data. This was also the case when minor changes were made in training data themselves. While classification results remained almost unaffected under minor changes in the training data, informative wavelets and subsequent coefficient values varied significantly. This was attributed to the particular structure of the algorithm in which minor modifications in the training data are followed by changes in probability distributions which in turn modify mutual information and informative wavelets derived by the algorithm. Changes in the informative wavelets are amplified by the application of matching pursuit algorithm. In the matching pursuit algorithm, wavelets generated at the later stages are highly sensitive to changes in the wavelets chosen at the early stages.
- The main problem in using informative algorithm is the high volume of computation, which makes the algorithm unsuitable for real time applications.

The above observations were considered to be major deficiencies for the application at hand, and as such, precluded further investigation and use of informative wavelets in this research.

CHAPTER 5

Dictionary Projection Pursuit

5.1. Introduction

To discover the outstanding and unspecified structure of a set of data, often visual representation such as histograms or scatterplots are utilized [55]. This can be easily done for low (one, two, or even three) dimensional data, however comprehensive visual tools for higher dimensions are not available.

Classical multivariate analysis provides powerful tools for gaining insight and understanding the nature of the phenomenon or the system that produced the data. These tools include a set of useful summary statistics (such as mean and covariance) as well as correlational structure of data. The summary statistics carries relevant information of the system if the data follow an elliptically symmetric distribution, such as the Gaussian (normal) distribution, in an n-dimensional variable space. However, there are numerous cases where real data deviates from a normal distribution, and consequently these summary statistics cannot represent the whole characteristics of data. Therefore, other appropriate methods need to be implemented to divulge the main characteristics of data.

In this chapter, projection pursuit (PP) algorithm as a method of revealing "interesting" structure of data is introduced, then an extension of PP which is a fast version of PP along with some results are presented.

5.2. Projection Pursuit Approach

Projection pursuit (PP) is a method for exploratory analysis of multivariate data sets which extracts remarkable linear projections of data to view them in a lower dimension; often onto a plane or a line. It numerically optimizes a certain criterion function or *projection index*. Friedman and Tukey [56] first used the term "projection pursuit", but the main idea was initially introduced by Kruskal [57]. Projection pursuit seeks a set of projections that are "interesting", in the sense of their deviation from Gaussian distribution [58]. PP is basically a method for revealing clusters among data.

There are several projection indices, among them, Friedman [59] proposed an index as the mean-squared difference between the projection score distribution and the Gaussian distribution, as the least structured density, to measure non-normality in the main body of the distribution (rather than in its entirety). His projection index basically measures departure from normality. Jones and Sibson [60] and Huber [58] set the PP idea in a more structured form and expanded it in a practical implementation. The approach involves an optimization process that starts at different random positions using the entropy concept from the information theory as the projection index to maximize the divergence of projected data from Gaussian distribution.

Projection pursuit was further elaborated in different applications such as regression [61], probability density approximation [58], and probability density estimation [62]. By employing a suitable projection index, PP technique can reveal an inherent structure or clustering in data. This can then be used in supervised [63,64,65], and unsupervised classification of high dimensional data [66,67], in detecting and classifying images [68] and in feature extraction of acoustic spectra [35].

The use of PP has been limited because of its high computational complexity. To resolve such a difficulty, Rutledge and McLean [35] employed wavelet packet decomposition during the search process of PP to introduce an extension of PP which is computationally more efficient. The procedure is described next.

5.3. Dictionary Projection Pursuit

Rutledge and McLean [35] proposed a method which looks for a set of basis functions from a dictionary of redundant wavelet packets in accordance with an orthogonality criterion, instead of optimizing a criterion as is done in PP. The search is performed in m iterations, where m is the required number of bases, and is decided upon empirically. In each iteration, they use a one-dimensional version of projection pursuit (which means $m=1$ in Eq. 2-1) to find the interesting features of acoustic waveforms. If A is a matrix consisting of all bases of a dictionary such as wavelet packets, then the first base is chosen from dictionary A according to a criterion described below. The dictionary is sometimes called *redundant*, since there are more than one set of basis functions which can span n-dimensional space. The one-dimensional version of projection pursuit is repeated m times where a procedure such as the one in matching pursuit [26] is applied until a set of bases $B=[b_1\ b_2\ ...b_m]$ is selected. Then the data are projected onto the selected basis functions to find an interesting view of the data. This is done by the linear projection $Z=B^TX$, where X is the data set and Z is the transformed data in the space of reduced dimension. The algorithm, so-called *dictionary projection pursuit* (DPP), is a greedy approach in a sense that after a given basis is selected in each iteration, its structure is eliminated from the data set before the next search for finding another basis is carried out in the subsequent iteration.

For finding a set of basis functions B, that contain desired characteristic information, a weight w is assigned to each basis function in the wavelet packet dictionary. The weight w is a measure of linear independence of the selected basis from all of the previously selected basis functions. For the initial condition, weight is set as a vector of ones at the beginning of the procedure, and is updated in each iteration. The weight vector is then changed in a manner that each selected basis is orthogonal to the subspace of previously selected basis functions, resulting in a set of orthogonal basis at the final stage of the algorithm . Here is the complete algorithmic procedure:

Step 1: Find wavelet packet coefficients of each training data in different classes, record them as a set of matrices of size n x (log n+1), where n is the signal dimension (call them *map*).

Step 2: Find the density (energy) of each packet by squaring each element in the wavelet packet matrices (*map*.^2 in Matlab notation) to obtain energy map of each training data.

Step 3: Sum all of the matrices in step 2 for each class. Normalize the resultant matrices by dividing them by the number of training data in each class (N_l) to find normalized total energy of training data (energy map) in each class:

$$C_l(j,k,m) \equiv \sum_{i=1}^{N_l} (\mathbf{b}_{j,k,m}^T \cdot \mathbf{x}_i^{(l)})^2 / N_l\,, \qquad \text{for} \qquad l = 1,\ldots,L. \tag{5-1}$$

C_l, the energy map of class l, is a table which can be rearranged in a matrix form (call it *e-map*). At this stage, there are L (number of classes) matrices.

Step 4: Find the relative entropy of *e-map* (call it *ent_map*) by applying Eq. (2-10) or the symmetric version Eq. (2-8).

Step 5: Repeat the following *m* times to find a set of orthogonal basis:

- Find the basis *b* corresponding to wavelet packet indices associated with *maxarg*(*w*.**ent_map*(:)), which is the maximum of element-by-element multiplication of vector *w* and vector form of *ent_map*.
- Compute the part of basis *b* which is orthogonal to basis functions already selected (call it residual), and normalize it to unity. (For more detail, please refer to [35].)
- Starting from left enter and save the residual of *b* in a matrix as a new column vector.
- Compute *coef*, wavelet packet coefficients of the new basis function (the residual), and accumulate the energy of coefficients:

  coef_e_sum = *coef_e_sum* + *coef*.^2.
- Update *w* as *w* =1- *coef_e_sum*.

In DPP, the projection index, which is the entropy of normalized sum energy of wavelet packet coefficients of all data set, is found in the beginning of the algorithm only once, contrary to PP in which the projection index must be calculated every time. This is the key feature of the algorithm which makes it faster compared to PP approach.

The normalization in step 3 will be trivial (inconsequential) if the number of training data in each class is the same. In chapter 7, a new normalization method is proposed to improve the classification results. In *step 5*, if n bases are selected a complete orthogonal basis is found. Then, another method such as *principal component analysis* (PCA) [46,55] can be applied for further dimensional reduction.

The above algorithm is a one-dimensional projection in the sense that matrix B is replaced by a single vector b. In each iteration, a basis function b is selected and added to the previously selected basis functions in the form of an expanding matrix B^T, where ^{T} is the matrix transpose operator. In this sense, the final projection is not one-dimensional, but a multi-dimensional projection.

In the next section, classification results of DPP obtained by applying the algorithm on a set of machine data are presented. In the classification stage a neural network (NN) classifier is used which was described in section 2.9.

5.4. Classification Results

To evaluate the algorithm, a single-cylinder engine data set is used, which contains 96 training data in three classes (32 data in each class). Data were normalized between –1 and 1 according to Eq. (3-4) in which the three classes correspond to three ignition timings of: -23 (normal), -33 (advance), and -10 (retard) degrees of crank angle.

Fig. 5.1 shows the first 8 bases selected by applying DPP on the data, using Coiflet 1 (6 tabs) as the analyzing wavelet.

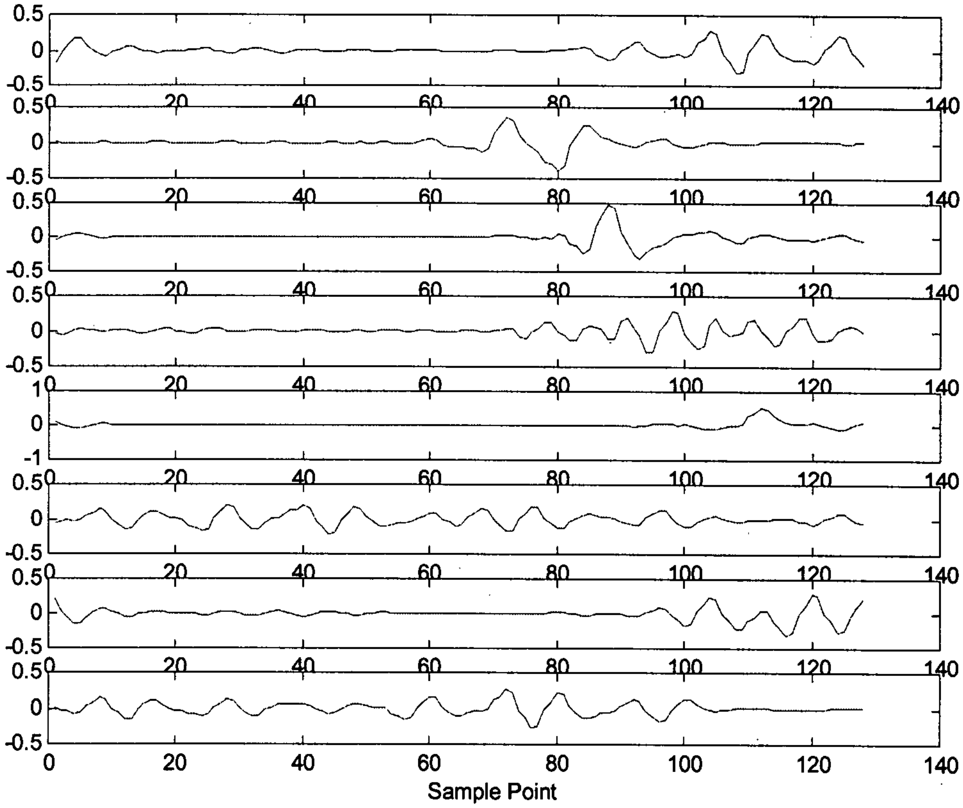


Fig. 5.1. The first 8 selected bases using Coiflet 1.

DPP chooses the first base from wavelet packet dictionary (which is optimal according to the criteria described in *Step 5*). The bases selected during the rest of the iterations are components of WP bases that are orthogonal to the bases selected in previous stages; they do not necessarily belong to the dictionary. Consequently, only the first base is certainly a wavelet packet base, the rest of bases are components of packet that are orthogonal to the previously selected bases. Fig. 5.2 shows the wavelet packet bases, determined by DPP, corresponding to the following wavelet packet indices:

5	4	3	5	2	7	4	6
6	2	1	9	0	25	3	12
0	6	11	0	25	0	5	0

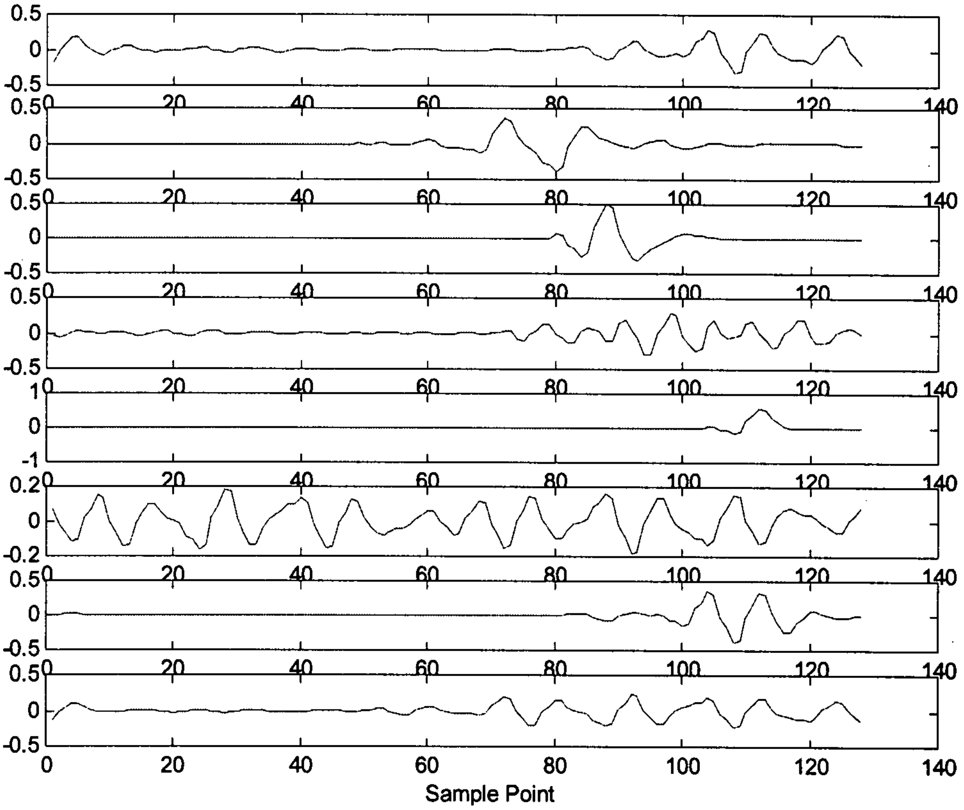


Fig. 5.2. The wavelet packet bases corresponding to the selected bases in Fig. 5.1.

First to third rows are scale, oscillation, and translation (position) indices, respectively. In this example, only bases 1 and 4 are packet bases, as can be seen by comparing Figs. 5.1 and 5.2. Further observation of the selected wavelets indicates similarity between the selected bases and corresponding packet bases that gradually decreases as we move to latter iterations (lower subplots in Figs. 5.1 and 5.2). In other words, as the number of selected basis increases (the number of columns in matrix B) the similarity of packet base and the corresponding selected base decreases. This can be explained nothing the fact that during the evaluation of the residuals in *step 5*, only a segment of packet basis that is orthogonal to the previously selected basis is chosen.

Fig. 5.3 illustrates training and testing coefficients for the first four bases before and after normalization and mean-centering, which were derived from projecting the training and testing data onto the first four bases. Four plots in each figure correspond to each of the four bases. Horizontal axis, segmented into three 32 training data (16 for testing data), corresponds from left to right to classes 1, 2, and 3. As it can be seen, the coefficient normalization applied through Eq. (3-4) magnifies the differences among classes, which is a suitable outcome for classification purposes. Fig. 5.4 shows histograms and mean-std plot of training coefficients in the same three classes, which indicates a sparse distribution with overlaps among different classes in the coefficient domain.

A common practice in classification is to preprocess data (here, the coefficients) before feeding them into the NN system. Such preprocessing may include a normalization method and mean-centering. The same preprocessing method should be applied to both training and testing data. To show the importance of preprocessing, a set of runs was carried out with different preprocessing schemes, including L^2-norm (Euclidean norm), norm defined by Eq. (3-4), both with and without centering. By normalizing coefficients through Eq. (3-4) the best classification result with a small error of 3% were obtained. While with no preprocessing the NN classification produces the worst results with about 13% error. Table 1 shows classification error under different preprocessing schemes.

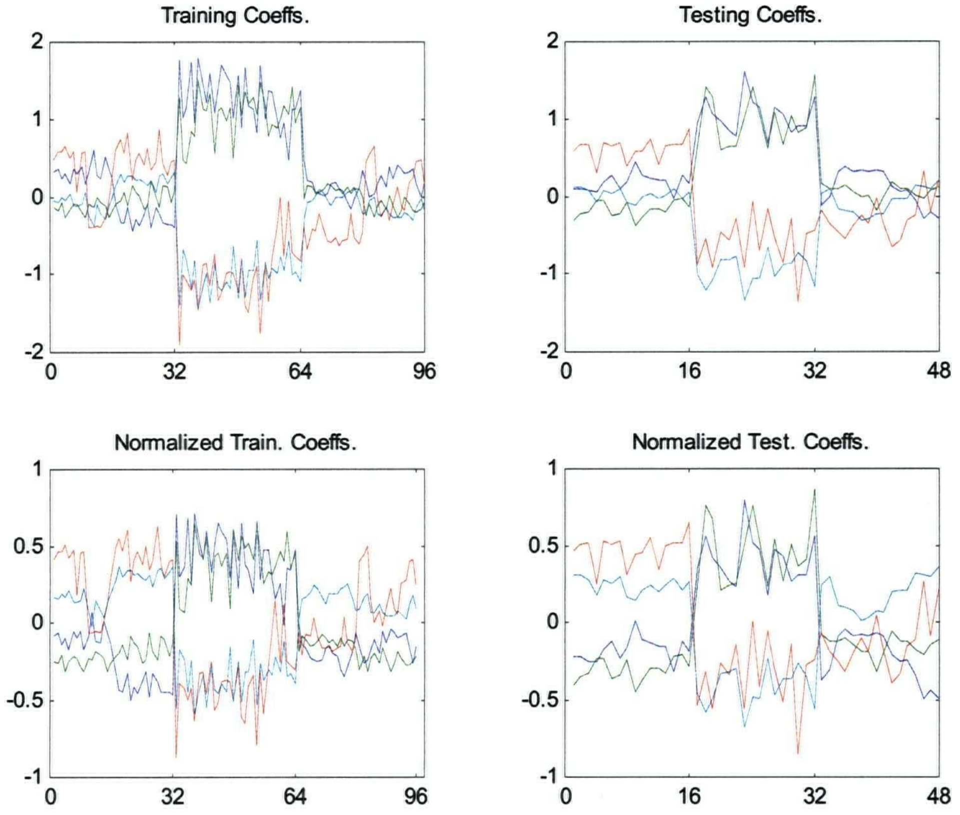


Fig. 5.3. Training and testing coefficients before and after normalization and mean-centering for bases 1-4: blue, green, red, and cyan, respectively. The horizontal axes are the number of training/testing data.

A moderate increase in the number of bases will slightly enhance the classification results but will escalate the computational cost.

DPP is still a time-consuming algorithm even though is much better than the original PP method. The computational cost is O($m\,n$ log n), where m and n are number of selected basis functions and signal size, respectively. If a complete set of basis functions is required, the computational cost will be $O(n^2 \log n)$, which is relatively high. In the next section a major drawback of DPP is highlighted.

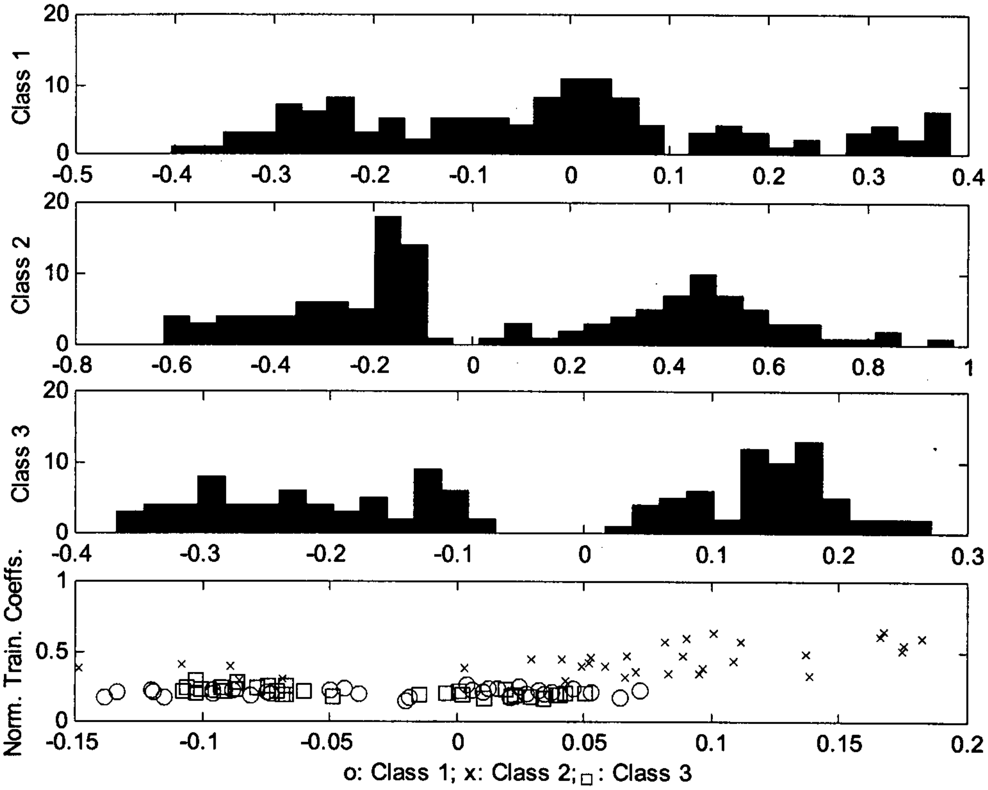


Fig. 5.4. Histograms and mean-std plot of normalized coefficients for three classes.

Table 5.1. Classification error using different preprocessing methods

Normalization Method	L^2-Norm & Centering	Eq. (3-4) Norm & Centering	L^2-Norm & No-Centering	Eq. (3-4) Norm & No-Centering	No Preprocessing
Error (%)	6	3	6	10	13

5.5. Disadvantages of DPP

Recalling the definition of Entropy in Eqs. (2-2) and (2-3) as

$$H(\mathbf{s}) = -\sum_{i=1}^{M} s_i \log s_i$$

where

$$\sum_i s_i = 1 \qquad \text{and} \qquad s_i \geq 0.$$

Entropy calculations require that each entry belongs to a probability density function (pdf). Meanwhile, DPP uses the relative entropy of normalized sum of the coefficient energies of all training data in L classes. It does not normalize *ent_map* to unity; instead, normalization is done by the number of training data in each class. (It is worth noting that normalizing *ent_map* to unity resolves the above problem; however, it adds another technical glitch: since we are comparing numbers rather than sequences, normalization to unity means that every element in the *ent_map* matrix must be one, which is trivial.)

As a result, and in accordance with the proof of *lemma* discussed in section 2.8, relative entropy will not necessarily be non-negative. Consequently, relative entropy as used in DPP, does not represent a theoretically acceptable measure for the separation of different distributions. However, while applying relative entropy measure under above condition is theoretically incorrect, it may still be considered as a measure (though not a robust one) for comparing different data and for the selection of wavelets for discriminatory classification.

The symmetric relative entropy measure (Eq. 2-8), because of its "symmetric" property, results in a non-negative value; whether or not the sum of sequence is unity (please refer to *Note* in section 2.8). Still, the symmetric version cannot provide a robust measure. In chapter 7 a method for resolving this problem is introduced.

5.6. Conclusions

The goal of projection pursuit for multivariate data analysis is to find low-dimensional projections, such as one or two dimensions, that provide the most revealing views of the full-dimensional data. In each iteration, DPP Finds the component of the selected basis that is orthogonal to the hyper-plane spanned by the previously selected basis functions. In this manner, an orthogonal set of basis is obtained. In this chapter, the usefulness of DPP in classification applications was shown.

Even though DPP is computationally faster than the projection pursuit algorithm, one may need a much more efficient method for on-line applications. It was also shown that DPP suffers from a technical deficiency in applying relative entropy on coefficients. To overcome this shortcoming, in chapter 7, DPP algorithm is modified to develop a new method for fault classification.

CHAPTER 6

Local Discriminant Bases

6.1. Introduction

In chapters 4 and 5, two different methods for pattern recognition and classification were introduced. In this chapter, another method, referred to as local discriminant bases (LDB), which is computationally faster is presented. Wavelets and LDB selection algorithm is applied to vibration signals in a single-cylinder spark ignition engine for feature extraction and fault classification. LDB selects a complete orthogonal basis from a wavelet packet dictionary of bases, which best discriminates the given classes, based on their time-frequency energy maps [48]. An appropriate normalization method in both data and wavelet coefficient domains, and a neural network classifier during the identification phase are used. By applying LDB to a real-world machine data the accuracy of the algorithm in machine fault diagnosis and classification is examined.

First "best-basis algorithm", which LDB's basic idea is based on, will be introduced. In order to make best-basis algorithm applicable to pattern recognition and classification problems the concept of discriminant measure, introduced in chapter 2, is employed. Then, classification results of applying LDB to machine data will be presented.

6.2. Best-Basis Algorithm

Coifman and Wickerhauser [21] employed entropy to efficiently represent a signal, mainly for data compression purposes. They introduced entropy as a real-valued cost function on sequences of coefficients and searched for its minimum over a dictionary of orthonormal bases. Entropy cost function can accurately quantify a sequence in the sense that entropy of a sequence is small when all but a few elements are negligible and is large when its elements are about the same size. Geometrically, best-basis algorithm minimizes the flatness of the energy distribution; the lower the entropy, the less flat the distribution is (see section 2.7 for more detail).

Best-basis algorithm selects a basis from a dictionary of orthonormal bases using the entropy criterion. It minimizes the entropy of the normalized signal energy after expanding a given signal or a collection of signals into a dictionary of orthonormal bases. This dictionary, which has a binary tree structure (Fig. 6.1), can be a set of redundant wavelet packet bases or local trigonometric bases.

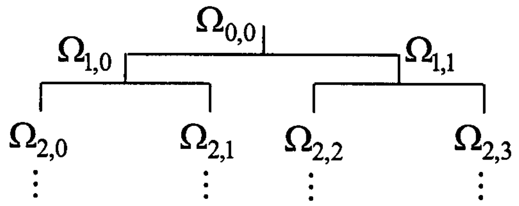


Fig. 6.1. Decomposition tree in a wavelet packet.

Best-basis algorithm uses a fast divide-and-conquer search. The algorithm is as follows:

Suppose $\mathbf{B}_{j,k} = (\mathbf{b}_{j,k,0},\ldots,\mathbf{b}_{j,k,2^{n_0-j}-1})^T$ be a set of basis vectors of $\mathbf{b}_{j,k}$ spanning the subspace $\Omega_{j,k}$ for $j = 0,1,\ldots,J$, $k = 0,1,\ldots,2^j - 1$, where n is the signal dimensionality, n_0 is the maximum level of wavelet packet signal decomposition, and J is the highest level of decomposition we would like to expand the signals to, with the upper limit of n_0 ($n_0 = \log_2 n \geq J$). $\Omega_{j,k}$ is the space of the basis vectors defined for node j,k of a binary tree constructed from a wavelet packet decomposition of the signal (Fig. 6.1). $\mathbf{B}_{j,k}$ can be written as a matrix corresponding to the subspace $\Omega_{j,k}$, which is the space of the basis vectors $\mathbf{b}_{j,k,m}$, defined for the node j,k of a binary tree constructed from a wavelet packet decomposition of the signal. In wavelet packet decomposition, basis vectors $\mathbf{b}_{j,k,m}$ are indexed by j, k, m representing scale, frequency band (oscillation), and time position, respectively. The number of basis vectors $\mathbf{b}_{j,k,m}$ equals to $n(1+\log_2 n)$. We note that $\mathbf{B}_{0,k}$ is the basis set of standard Euclidean coordinate system, which is the basis for the signal at its highest resolution level. Also, suppose that $\mathbf{A}_{j,k}$ is the best-basis for the signal $\mathbf{x} \in \mathbf{R}^n$ restricted by the span of $\mathbf{B}_{j,k}$.

Fig. 6.1 shows the subspace representation of a given signal in wavelet packet binary tree decomposition. Each node, which represents a subspace, is the orthogonal direct sum of the subspace of its two children. First, a time-frequency decomposition method such as wavelet packet transform or local trigonometric transform must be chosen. Then, the best-basis can be found by induction on j as following:

Step 1: Decompose the given signal $\mathbf{x}$ by expanding it into a dictionary of orthogonal bases to obtain coefficients.

Step 2: For the start of the algorithm, suppose that $\mathbf{A}_{J,k} = \mathbf{B}_{J,k}$ for $k = 0,...,2^J - 1$.

Step 3: Search for the best subspace $A_{j,k}$ starting from the last level of decomposition ($j = J-1,...,0$) and by the following ($k = 0,...,2^j - 1$):

if $H(\mathbf{B}_{j,k}\mathbf{x}) \leq H(\mathbf{A}_{j+1,2k}\mathbf{x} + \mathbf{A}_{j+1,2k+1}\mathbf{x})$ **set** $\mathbf{A}_{j,k} = \mathbf{B}_{j,k}$

otherwise $\mathbf{A}_{j,k} = \mathbf{A}_{j+1,2k}\mathbf{x} + \mathbf{A}_{j+1,2k+1}\mathbf{x}$.

To make the algorithm computationally efficient entropy map must be additive, which fortunately is, i.e. $H(0)=0$ and $H(\{s_i\}) = \sum_i H(s_i)$ (see the proof in section 2.7). This property of entropy makes the best-basis algorithm a fast divide-and-conquer search. To guarantee the condition $s_i \geq 0$ in the definition of entropy, algorithm chooses the entropy of normalized signal energy as $s_i=|x_i|/||\mathbf{x}||$, where $||.||$ is the Euclidean norm. The complexity of computation for a signal in $\mathbf{R}^n$ is $O(n)$.

Best-basis algorithm has shown its suitability for signal representation and data compression applications but it is not necessarily appropriate for classification problems. In classification one should search for those bases that give the most discriminant information about class separation.

6.3. Local Discriminant Bases Algorithm

Saito and Coifman [22] introduced local discriminant bases (LDB) algorithm as an extension of the best-basis algorithm by Coifman and Wickerhauser [21]. The best-basis algorithm uses the entropy of normalized signal energy to represent a signal efficiently, mainly for data compression applications. LDB approach, on the other hand, employs the relative entropy of normalized time-frequency energy map of all training signals in each class for classification purposes.

The time-frequency energy map of $\{\mathbf{x}_i^{(l)}\}_{i=1}^{N_l}$, a set of training signals belonging to class l, along the direction of a given basis $\mathbf{b}_{j,k,m}$ is a table defined by:

$$C_l(j,k,m) \equiv \sum_{i=1}^{N_l} (\mathbf{b}_{j,k,m}^T \cdot \mathbf{x}_i^{(l)})^2 / \sum_{i=1}^{N_l} \left\| \mathbf{x}_i^{(l)} \right\|^2 \qquad (6\text{-}1)$$

for $j = 0,1,\ldots,J$, $k = 0,1,\ldots,2^j - 1$, $m = 0,1,..,2^{n_0 - j} - 1$, where N_l is the number of training sets in class l. Here "." denotes the standard inner (dot) product in $\mathbf{R}^n$. The energy map defined by (6-1) is computed by summing the squares of expansion coefficients of the signals at each position and then normalizing them with respect to the total energy of the signals belonging to class l.

One can obtain expansion coefficients by decomposing a signal of length n into a tree-structured bases such as wavelet packet or local trigonometric dictionaries where the computational efficiency is $O(n\log n)$ and $O(n(\log n)^2)$, respectively. Here we have used wavelet packet dictionary.

Discriminatory power associated with a given wavelet packet node indexed by j,k is the sum of the discriminatory power of its constituent basis $\mathbf{b}_{j,k,m}$ measured in the coefficient domain. Additive property of discriminatory measure is used here as follows:

$$D(\{C_l(j,k,.)\}_{l=1}^{L}) \equiv \sum_{m=0}^{2^{n_0-j}-1} D(C_1(j,k,m),...,C_L(j,k,m))\,. \qquad (6\text{-}2)$$

As stated earlier, LDB selects a local orthogonal basis from a dictionary of basis in a wavelet packet, which properly categorizes the given classes, based on the discriminatory measures of their time-frequency maps. Suppose that $A_{j,k}$ represents the desired local discriminant basis restricted to the span of $B_{j,k}$, which is a set of basis vectors at (j,k) node, and $\Delta_{j,k}$ is the array containing the discriminant measure of the same node. The additive property of discriminant measure D is advantageous for a computationally fast algorithm.

The algorithm first chooses a time-frequency decomposition method such as wavelet packet transform or local trigonometric transform. Then, for a given training dataset consisting of L classes of signals $\{\{\mathbf{x}_i^{(l)}\}_{i=1}^{N_l}\}_{l=1}^{L}$, the local best-basis can be found by induction on j as following:

Step 1: Decompose the given signal $\mathbf{x}$ by expanding it into a dictionary of orthogonal bases to obtain coefficients and construct time-frequency energy maps C_l for l=1,...,L.

Step 2: For the start of the algorithm, suppose that $\mathbf{A}_{J,k}$ = $\mathbf{B}_{J,k}$ and set $\Delta_{J,k} = D(\{C_l(J,k,.)\}_{l=1}^{L})$ for $k = 0,...,2^J - 1$.

Step 3: Set $\Delta_{j,k} = D(\{C_l(j,k,.)\}_{l=1}^{L})$ and search for the best subspace $\mathbf{A}_{j,k}$ for $j = J-1,...,0$ and $k = 0,...,2^j - 1$:

If $\Delta_{j,k} \geq \Delta_{j+1,2k} + \Delta_{j+1,2k+1}$,

Then $\mathbf{A}_{j,k} = \mathbf{B}_{j,k}$,

Else $\mathbf{A}_{j,k} = \mathbf{A}_{j+1,2k} \oplus \mathbf{A}_{j+1,2k+1}$ and set $\Delta_{j,k} = \Delta_{j+1,2k} + \Delta_{j+1,2k+1}$.

Step 4: Rank in descending order the complete orthogonal basis functions found in *Step 3* according to their discrimination power.

Step 5: Use k (much less than n) most discriminant basis functions for constructing classifiers.

If we start from the last level of decomposition (which is usually the case), i.e., $J=n_0$, in step 3 there will be no summation and $\Delta_{J,k}$ is simply the elements of level n_0. During step 3, a complete orthogonal basis with a fast computation of $O(n)$ is built. Orthogonality of bases is imposed in the algorithm to ensure that wavelet coefficients used as features during classification are uncorrelated. After this stage, one can simply select k highest discriminant bases in step 5 and use the corresponding coefficients as features in a classifier. It is also possible to employ a statistical method such as Fisher's criterion to reduce the dimensionality of the problem first and then input them into a classifier [22].

In brief, LDB algorithm starts by comparing the discriminatory power of the nodes at the highest scale as "children" nodes with "parent" node residing one scale lower. For example, in a two level decomposition of wavelet packet tree (Fig. 6.1), algorithm first compares the discriminant power evaluated for coefficients of training data in different classes at node $\Omega_{1,1}$ with those nodes of $\Omega_{2,2}$ and $\Omega_{2,3}$. If the entropy of $\Omega_{1,1}$ is larger, the algorithm keeps bases belonging to this node and omits the other two, otherwise it keeps the two and disregards node $\Omega_{1,1}$ bases. This process is applied to all nodes in a sequential manner up to the scale $j = 0$. At this stage a set of complete orthogonal basis having the highest discriminatory power is obtained which can be sorted out further at the second stage and used for classification of designated classes.

6.4. LDB Classification

The Ricardo Hydra data, introduced in chapter 3, was used to evaluate the effectiveness of LDB algorithm for the classification. According to Eq. (3-4), 96 training data in three classes (32 data in each class) were normalized between –1 and 1, in which three classes correspond to three ignition timings of: -23 (normal), -33 (advance), and -10 (retard) degrees of crank angle.

Coefficients were also normalized between –1 and 1, and then mean-centered. This preprocessing method was found to yield the lowest classification error during testing.

Normalization done by Eq. (3-4) was found to be promising since under this scheme low values remain low and high values remain high. Normalizing the data reduces biasing among data, regulates the amplitude of the coefficients, and causes the algorithm to select wavelet bases which carry more discriminatory information. However, mean-centering of the data after normalization does not produce noticeable changes, since the mean value of machine vibration data, due to (almost) symmetric property of vibration with respect to neutral axis, is usually close to zero.

Fig. 6.2 illustrates the first 8 bases selected by the algorithm and sorted according to their discriminant power using Coiflet 1 (6 tabs) as the analyzing wavelet. Wavelet packet indices of the selected bases are given below in which first to third rows are scale, oscillation, and translation (position) indices, respectively:

5	4	6	5	4	6	7	5
0	2	12	9	2	16	34	7
3	6	0	0	7	0	0	3

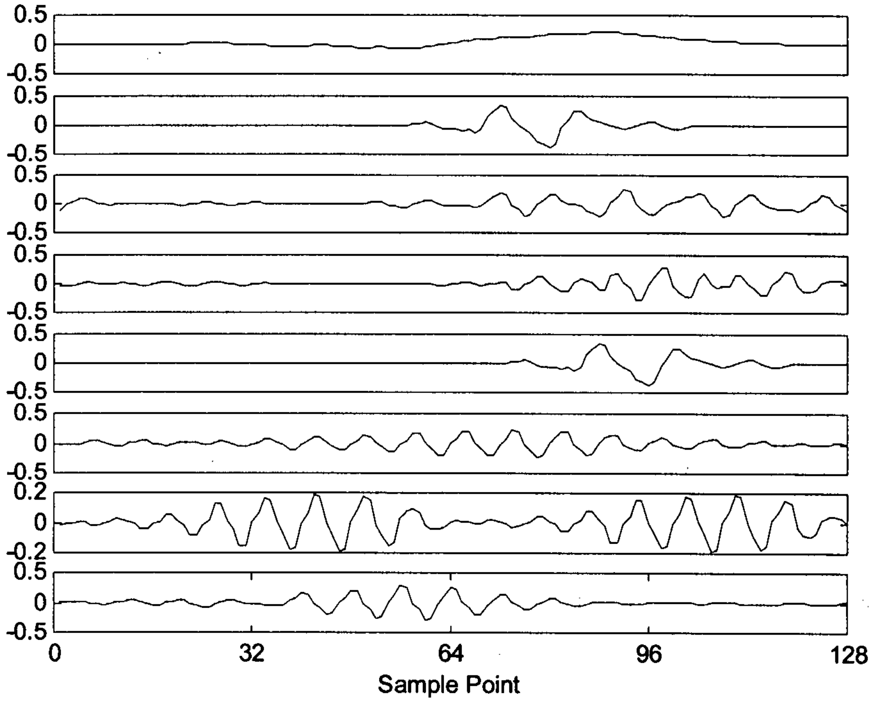


Fig. 6.2. The first 8 bases selected by the LDB algorithm using Coiflet 1.

Discriminant measures of all 128 complete orthogonal bases are ranked in decreasing order as shown in Fig. 6.3. The figure exhibits a sharp drop of the measure after the first few bases. Only bases with largest discriminatory measures may be considered for classification purposes, as the discriminant measures of other bases are insignificant.

Training coefficients were derived from projecting the training data on the first four bases; they were applied as feature variables and as inputs to a backpropagation neural network classifier after normalization and mean-centering (the specific method of NN used for classification were explained in chapter 2). For evaluating the algorithm, testing data were projected onto the selected bases to generate the coefficients followed by normalization and mean-centering as applied to the training data. Fig. 6.4 shows training and testing coefficients for the first four bases before and after normalization and mean-centering. Four graphs in each subplot correspond to each of the four bases. These plots indicate that the relative values of the coefficients among the classes have been preserved under the normalization method used here. Three classes can easily be identified by inspecting the values of the coefficients which are distinctly different in three segments of size 32 each corresponding to 32 training data in each class (and 16 for testing data).

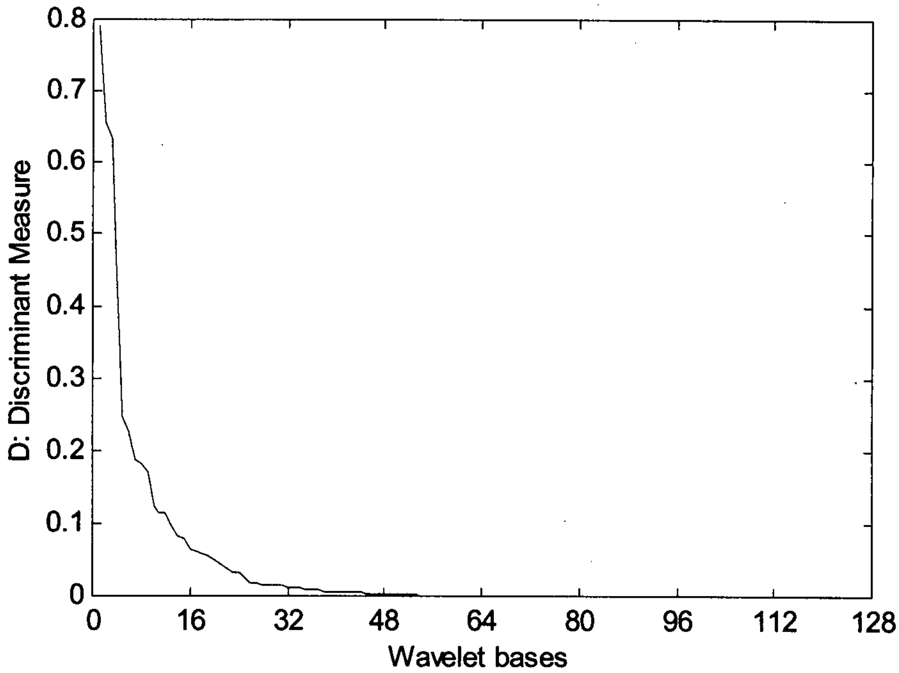


Fig. 6.3. Discriminant measure of all 128 selected bases.

Mean value vs. standard deviation of Ricardo Hydra data in each class, shown in chapter 3 (Figs. 3.6 and 3.7), indicates that classes 1 and 3 are highly clustered. However, applying raw data to a classifier is neither accurate because of dilution of information, nor time-efficient due to high dimensionality of the original data. On the other hand, training coefficients are sparsely distributed having a non-clustered mean-std pattern (Fig. 6.5) which is attributed to decorrelation effect caused by wavelet transform in the coefficient domain.

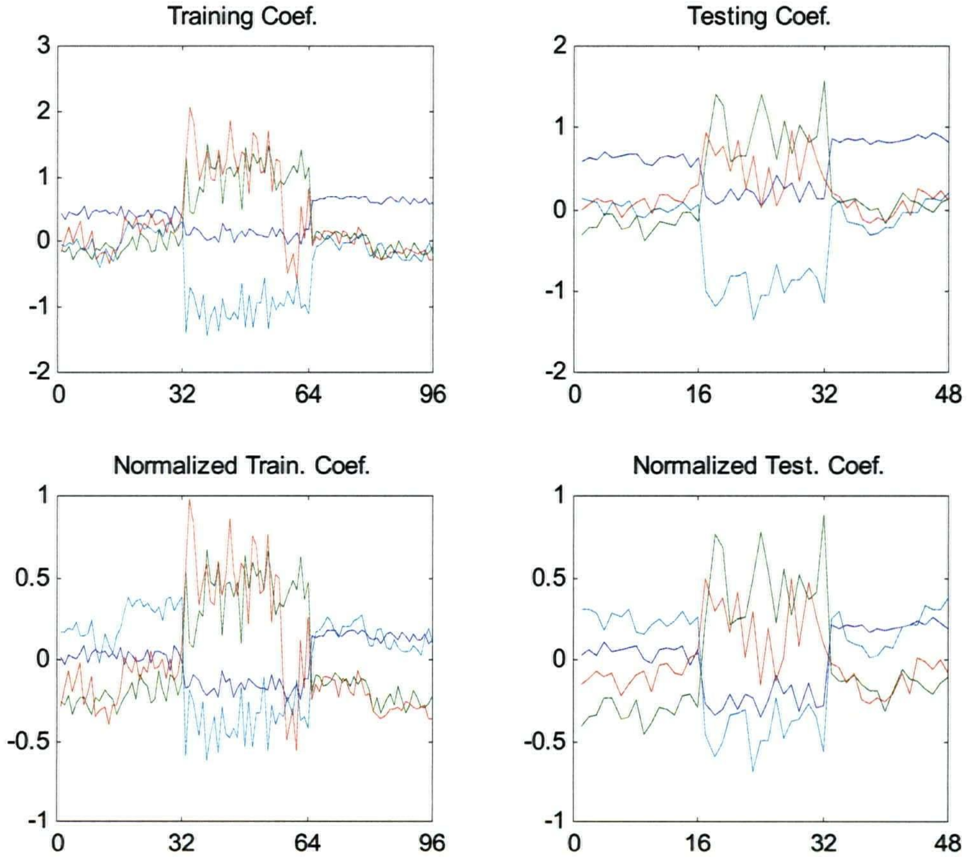


Fig. 6.4. Training and testing coefficients before and after normalization and mean-centering for bases 1-4: blue, green, red, and cyan, respectively. The horizontal axes are the number of training/testing data.

Training coefficients were applied to a two-layer NN backpropagation classifier with 5 neurons in hidden layer and tan-sigmoid transfer (activation) function in both hidden and output layers.

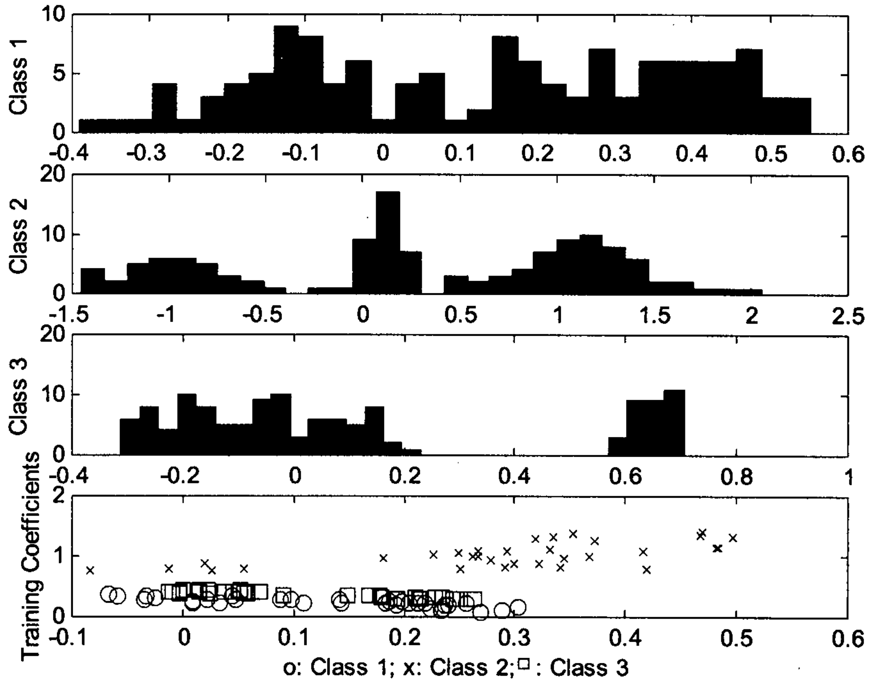


Fig. 6.5. Histograms and mean-std plot of training coefficients for three classes.

The number of selected bases was considered to be an important factor as it determined the number of features. In general, an increase in the number of bases, results in a more discriminatory information that are passed to the classifier. However, in classification applications, often a finite number of features is sufficient to describe different classes, beyond which dilution of the information deteriorates the classification results. Excessive number of features will also increase the computational cost during classification. It was observed that when the number of bases was changed from 4 to 128

– i.e. to include up to all of the orthogonal bases – the classification error increased from 1% for 4 bases to 6% for 128 bases. Generally, there is no analytical solution available for determining an optimum number of bases (features); often an empirical approach based on à priori information about the particular application is utilized.

Using wavelet transform and the coefficients as feature variables, considerable computational efficiency was gained during the classification. Noting that the computational order of backpropagation algorithm is of O(n^2), with n as the size of NN input, reducing the size of training data from $\{96 \times 128\}$ to coefficients of size $\{96 \times 4\}$ significantly contributed to computational efficiency. In order to compare the advantages gained from using wavelet transform, the original data were also applied to the NN. The classification error was observed to be higher (about 7%) as compared with 1% when using wavelet coefficients.

Something that we have to be aware of is that in both best-basis algorithm and LDB, the energy of every node is normalized by the norm of original signal; therefore, only in the root of the tree (the signal itself) condition (2-6), i.e. $\sum_i s_i = 1$, holds true. In fact, in both "best-basis" and LDB algorithms, we are comparing entropy of different node energies, while these energies may not be comparable since they are not normalized in the same basis/foundation. This may bring some inaccuracies.

6.5. LDB Shortcoming

Despite the potent capabilities of LDB, it encounters drawbacks somehow analogous to DPP. In step 2 of LDB algorithm (i.e., in the last level of wavelet packet decomposition), for example for a two-class case, we basically find the relative entropy of two positive scalars, instead of two sequences. It is noted that each node in the last level of decomposition includes only one base, thus each node contains one coefficient only; as a result, relative entropy is derived between two scalars. According to the definition (2-10), when $s^{(1)}/s^{(2)} <1$ (which is highly probable), relative entropy is negative. Then in step 3 of LDB, during comparison of the sum of discriminant measures of two children nodes with their parent node, we may add up a negative number with a positive number and compare the sum with the discriminant measure of the parent node. Under this situation, we will not have an effective measure of "distance" between two classes since distance is to be a positive number. This is the shortcoming of DPP as well, as mentioned in section 5.5.

On the other hand, in different levels of wavelet packet decomposition, other than the last level, relative entropy measure is not always positive, since condition 2-6 is not satisfied.

As it is described in *lemma* in section 2.8, the symmetric relative entropy measure is always non-negative, regardless of whether condition 2-7 holds (sums of sequences are one). However, this is not the case for the relative entropy, as mentioned above.

To overcome these shortcomings of LDB, a new methodology is outlined in the next chapter.

6.6. Conclusions

As the use of pressure signals for fault detection in internal combustion engines, due to their costly retrofit, is prohibitive, the suitability of using acceleration data for engine diagnosis was investigated. As well, due to the large size of data, direct application of NN classifier to data in different classes is extremely time consuming. Noting that sensor data usually include redundant information, direct use of sensor data, may also lead to dilution of information about engine faults and produce unacceptable classification results. Alternatively, transformation of the data into wavelet domain and use of wavelet coefficients as feature variables will reduce the data dimensionality considerably.

Transient nature of the machine vibration signals requires the use of basis functions that capture localized features of the signals. It was shown that wavelets with finite support width both in time and frequency domain are highly suitable for analysis of these signals in diagnostic applications. Along this line, using wavelet packet and redundant signal decomposition, local discriminant basis algorithm enabled us to select a subset of basis with highest discriminatory power to classify different engine operating conditions.

LDB algorithm attempts to select a set of orthogonal bases from a wavelet packet dictionary which best discriminates different states of the system. Wavelet coefficients constructed by projecting data onto the selected bases are used as feature variables and as inputs to a backpropagation neural network (NN) classifier. We employed LDB algorithm for data analysis with three classes of ignition timings. For acquiring wavelet coefficients, normalized training data were utilized.

It was shown that using proper normalization both in signal and feature domains, accurate classification results could be obtained. Furthermore, the use of orthogonal bases selection in LDB significantly contributed to the reduction of the number of bases where we used the first few bases with high discriminatory measure. This consequently enhanced the accuracy of classification results. Further, high computational efficiency of LDB lends itself to on-line performance monitoring.

Normalization was considered to be an important factor in the classification process. The particular normalization strategy was found to influence the accuracy of results considerably.

Choosing the number of features is also an important task; this number should neither be too large to dilute the information nor too small to miss important discriminant information. Roughly speaking, a value around log(n) (n is the length of data, or the space dimension) is considered optimum for the number of features.

CHAPTER 7

A New Approach for the Construction of Entropy Measure and Energy Map

7.1. Introduction

In chapters 5 and 6 details of two wavelet-based methods namely DPP and LDB for feature selection and classification were described. They were applied for machine fault diagnosis using real-world data. In sections 5.5 and 6.5 shortcomings of these methods for feature extraction were discussed. It was stated that a close examination of the DPP and LDB methods reveals that their interpretation of entropy is non-standard and this poses certain technical glitches. In both methods, relative entropy is applied on sequences of numbers that do not constitute a probability density function (pdf), in the sense implied by the condition (2-7).

Furthermore, in both DPP and in the last level of decomposition in LDB (as expanded in section 7.3), the relative entropy of two scalars instead of two sequences is evaluated. To satisfy condition (2-7), it is necessary that these two scalars be unity; however, such a circumstance corresponds to a trivial case of entropy evaluation where no useful information can be extracted. Therefore, both methods face a theoretical dilemma.

We note that for a sequence to be used in entropy measure, it must be expressed in the form of a pdf where the total sum of the sequence is unity. Nevertheless, one can still use entropy as a distance measure even if the sequences are not pdfs as is the case in LDB and DPP. This may result in negative relative entropies being calculated. But, in order to properly compare the relative entropies D, as a discriminant measure, it is essential that all Ds be non-negative. Necessary and sufficient condition for this is still relation (2-7) (please see *lemma* in section 2.8).

The use of symmetric relative entropy, which because of its "symmetric" property is always non-negative, can be considered as an option. (See the proof in the *note* of section 2.8 and Eq. 2-12). However, in symmetric relative entropy calculations some negative and positive terms cancel out. This poses a problem since the magnitude of discriminant measure has been reduced. This indicates that the measure cannot be considered as an effective measure for the discriminatory classification.

To resolve this dilemma, in this chapter, a novel method is presented that can be combined with other searching algorithms such as LDB and DPP. Also, the application of singular value decomposition (SVD), and its importance in statistical pattern recognition, along with some results and discussions are presented. SVD is a technique that is widely used for evaluating the correlation among experimental data composed of p sets where each set is a sequence of length q. Then, data can be expressed as a $p \times q$ matrix B. Singular values (SV) of matrix B are the eigenvalues of the correlation matrix B^TB ranked from high to low (for more detail and mathematical definition of SV readers can refer to statistical or linear algebra texts, such as [55]). In our data analysis, SVD of the coefficient matrix, constructed from projecting the data onto the selected bases, is employed to determine the extent of correlation among coefficients. We, first, introduce a modification to the normalization scheme used in DPP.

7.2. Class-Based Normalization

In DPP, coefficients are normalized as defined by Eq. (5.1), where the normalization is basically the average of the sum of the squared coefficients in each class. Consequently, the total energy of coefficients in each class is divided by the number of training data in that class. Under this scheme, normalization basically corresponds to scaling down the signal energy in each class. In the special case, where the number of training data is the same for all classes, energy values are scaled down by the same proportion which corresponds to a uniform scaling of the entropy map values, thus there will be no relative changes in the final outcome.

In the proposed approach, normalization as used in step 3 of DPP algorithm (section 5.3), is modified. Under the new approach, a class-based normalization is used in which each class is considered separately where the sum of squared *coefficient values* of different wavelet packet nodes, is adjusted by the sum squared values of *all the training data* in that class as:

$$\sum_{i=1}^{N_l}(\mathbf{b}_{j,k,m}^T.\mathbf{x}_i^{(l)})^2 / \sum_{i=1}^{N_l}\left\|\mathbf{x}_i^{(l)}\right\|^2 \tag{7-1}$$

where N_l is the number of training data in class l. Under normalization defined by Eq. (7-1), different classes are normalized with respect to their own factors, resulting in further class differentiation during feature extraction stage and an improved accuracy in the classification stage.

To examine the effectiveness of this modification, many trial runs were carried out in which the new normalization scheme was applied to Ricardo Hydra experimental data – the set of data that was introduced in chapter 3. In order to ascertain and generalize the effectiveness of the proposed method, a wide range of data analysis using different wavelets was planned and performed. These included the use of 32 different analyzing wavelets from the family of orthogonal, biorthogonal, symmetric as well as selected wavelets from Battle-Lamorie spline functions as follows:

1-Haar, 2-Beylkin, 3-Coiflet1, 4-Coiflet2, 5-Coiflet3, 6-Coiflet4, 7-Coiflet5, 8-Daubechies2 (Db2), 9-Db3, 10-Db4, 11-Db5, 12-Db6, 13-Db7, 14-Db8, 15-Db9, 16-Db10, 17-Db20, 18-Db40, 19-Db45, 20-Bior22, 21-Bior31, 22-Bior68, 23-Symmlet4 (Sym4), 24-Sym5, 25-Sym6, 26-Sym7, 27-Sym8, 28-Sym9, 29-Sym10, 30-Vaidyanathan, 31-Battle3, 32-Battle5.

Fig. 7.1 shows the classification results of DPP with the two normalization schemes, in which the horizontal axis indexed from 1 to 32 corresponds to the numbers used above to list the analyzing wavelets. For the majority of wavelets the proposed normalization scheme produced superior performance. For example, by applying Coiflet1 as the analyzing wavelet (number 3 in Fig. 7.1), misclassification rate was reduced from 4% with the N_l-normalization to 0.5% with the modified normalization scheme. This is considered as a notable improvement. In fact, with class-based normalization of signals additional separation of classes is induced. In the next section, a novel approach for using relative entropy in the construction of energy map is proposed.

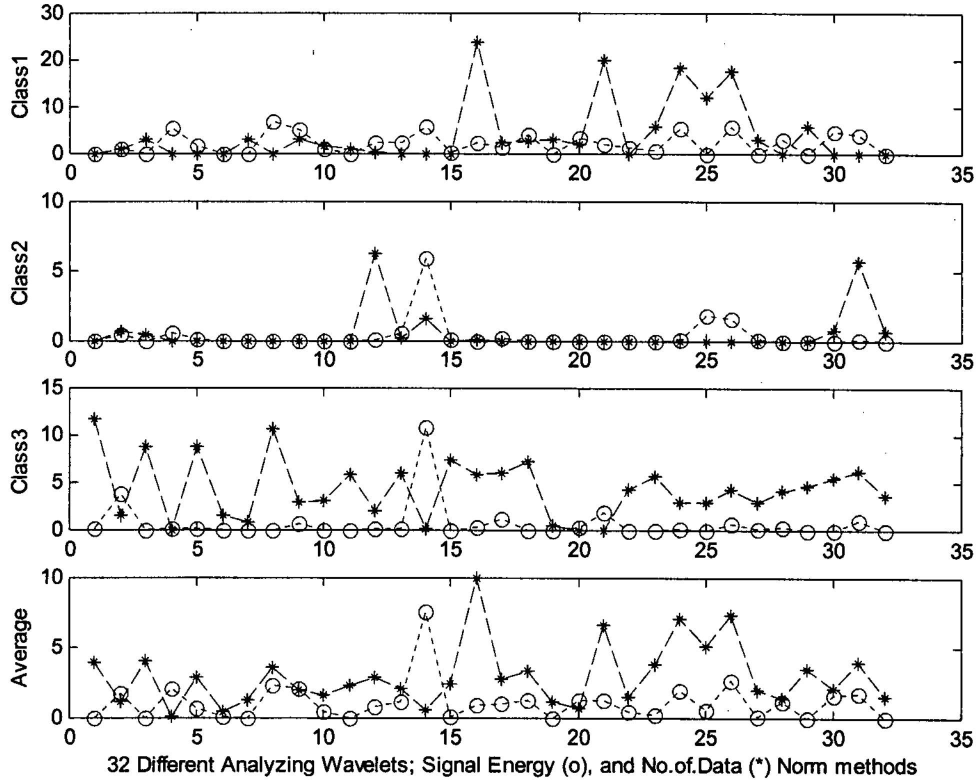


Fig. 7.1. Classification percentage error using DPP under two normalization schemes, with 32 different analyzing wavelets listed in section 7.2.

7.3. Cross-Data Entropy Approach

The use of entropy measure for feature extraction and classification requires that:

1) Entries to Eq. (2-10) for the evaluation of relative entropy be all non-negative,

2) Relative entropy among different classes in each node (in every base of the dictionary in the case of DPP), i.e., the outcome of Eq. (2-10), be also non-negative.

Condition (1) is met by considering sum of energy of coefficients as the entries to the relative entropy measure. However, the use of relative entropy of the above sum in each class, as prescribed by LDB and DPP methods, does not guarantee condition (2), i.e., we may not have non-negative relative entropies for each and every data at all times. (Refer to the explanation given in sections 5.5 and 6.5).

Another consideration for the use of relative entropy measure is the requirement that the sequence constitutes a pdf. While in LDB method we observed that in the last level of decomposition in wavelet packet tree (refer to step 3 of LDB algorithm described in section 6.3), relative entropy is applied on singular scalars. Moreover, in DPP approach, relative entropy is applied to singular scalars in all levels of the decomposition process.

To resolve these shortcomings, an approach is proposed here, in which training data are used to generate the required sequence of numbers for proper application and evaluation of entropy. It is proposed that in constructing the of entropy measure, instead of using the sum of coefficient energies of all training data in each class, and at each node, as used in LDB (and with some variations in DPP as described in 5.3), we consider individual coefficients for the evaluation of the entropy measure. As a result, the role of every single data is taken into account in the sense that the relative entropies of each element in the wavelet packet matrix are used to find the appropriate bases. Under this approach we deviate from the concept of "averaging of data" as is the case in both LDB and DPP methods. Two advantages are gained using the proposed scheme.

1) Averaging of all training data as used in LDB and DPP methods essentially utilizes the *first order* statistics only. By not involving a second order statistics, such as standards deviation, the dispersion of data is masked. This is considered a limitation of the LDB and DPP methods. The proposed scheme eliminates this limitation by considering all training data where coefficients are obtained and used for each and every training data.

2) In the proposed algorithm, each coefficient is evaluated for all training data and thus at all nodes including the last level of wavelet packet tree, evaluation of entropy is carried-out on a sequence of scalars rather than on a single scalar. The scheme can then be interpreted as a cross-data entropy evaluation or cross-data energy map approach. Since we still use relative entropy, discriminatory bases will be derived as before. Under this scheme the relative entropy of distributions of the coefficients in different classes is taken into account, that is, discriminant information of every data (mutual discrimination among all data) is considered. For this reason, we will refer to this method as a *cross-data entropy or mutual-based approach.* The cross-data entropy approach alleviates the shortcoming of standard relative entropy measure used in LDB and DPP methods.

The proposed method can also be used in conjunction with the main idea of other searching algorithms. In the following sections, formalization of the extended versions of DPP and LDB methods referred to as *cross-data or mutual dictionary projection pursuit* (MDPP), and *mutual local discriminant bases* (MLDB) is given. We define the following notations before describing the methods.

Let "*map*" be the wavelet packet coefficients of each training data $\mathbf{x}_i$, for i = 1,..., N, which can be demonstrated as a set of N matrices of size n x ($\log_2 n$ + 1), where n is the signal length, N is the number of training data, n_0 is the maximum level of wavelet packet signal decomposition, and J is the scale index of decomposition level, with $n_0 = \log_2 n \geq J$. Let $C^{i=1:N}(j,k,m) \equiv (\mathbf{b}_{j,k,m}^T . \mathbf{x}_{i=1:N})^2$ be the energy map of each training data derived by squaring each element of the *map* matrices, where $C^{1:N}$ is used to denote N energy map matrices each of the size of *map*. (Matrices $C^{1:N}$ can also be viewed as a 3D-array, *e-map*, of size n x ($\log_2 n$ + 1) x N.)

Recall that N_l is the number of training data in class l, where $N = \sum_{l=1}^{L} N_l$ is the total number of training data in all classes. If $C_l^{1:N_l}$ are energy maps of each training data in class l then $[C_l^{1:N_l}(j,k,m)]$ can be defined as a vector consisting of N_l number of element (j,k,m) of $C_l^{1:N_l}$:

$$[C_l^{1:N_l}(j,k,m)] = [C_l^1(j,k,m),...,C_l^{N_l}(j,k,m)]^T \qquad \text{for } l=1,\ldots,L$$

Similarly, we can think of $C_l^{1:N_l}$ as a 3D-array *e-map*$_l$.

The process used in the new mutual dictionary projection pursuit (MDPP), and mutual local discriminant bases (MLDB) is described next.

Consider a time-frequency dictionary such as wavelet packet transform. For a training dataset consisting of L classes of signals $\{\{\mathbf{x}_i^{(l)}\}_{i=1}^{N_l}\}_{l=1}^{L}$, MDPP and MLDB can be implemented by induction on the scale j, as follows.

7.3.1. Mutual Dictionary Projection Pursuit

Step 1: Expand each training signal into a dictionary of orthogonal bases (*map* matrices) to obtain coefficients.

Step 2: Find energy map of the coefficients, $C^{1:N}$, composed of squared values of each element of *map* matrices.

Step 3: Normalize matrices $C^{1:N}$ according to the new method proposed in chapter 3.

Step 4: Find discriminant power (by applying Eqs. 2-8 or 2-10) amongst L vectors $[C_l^{1:N_l}(j,k,m)]_{l=1:L}$ for $j=0,1,...,J$, $k=0,1,...,2^j-1$, and $m=0,1,..,2^{n_0-j}-1$, where n_0 is the maximum level of signal decomposition. Call the resultant matrix *ent_map*.

Step 5: Apply step 5 as outlined in section 5.3.

7.3.2. Mutual Local Discriminant Bases

The first four steps of MLDB and MDPP are the same; here we show the rest of the procedure:

Step 5: As initial values for the algorithm, suppose that $\mathbf{A}_{J,k}$ = $\mathbf{B}_{J,k}$ and set $\Delta_{J,k} \equiv \sum_{m=0}^{2^{n_0-J}-1} ent\_map(J,k,m)$ for $k=0,...,2^J-1$ (if we start from the last level of decomposition, i.e., $J = n_0$, there is no summation and $\Delta_{J,k}$ is simply the elements of level n_0).

Step 6: Set $\Delta_{j,k} = \sum_{m=0}^{2^{n_0-j}-1} ent\_map(j,k,m)$ and search for the best subspace $\mathbf{A}_{j,k}$ for $j=J-1,...,0$ and $k=0,...,2^j-1$:

If $\Delta_{j,k} \geq \Delta_{j+1,2k} + \Delta_{j+1,2k+1}$,

Then $\mathbf{A}_{j,k} = \mathbf{B}_{j,k}$,

Else $\mathbf{A}_{j,k} = \mathbf{A}_{j+1,2k} \oplus \mathbf{A}_{j+1,2k+1}$ and set $\Delta_{j,k} = \Delta_{j+1,2k} + \Delta_{j+1,2k+1}$.

***Step** 7*: Rank in descending order, the complete orthogonal basis functions found in *Step 6* according to their discriminant power.

***Step** 8*: Use k most discriminant basis functions for constructing the classifiers. Note that k is much less than n.

Let us reiterate that the mutual-based approach is not founded on the "sum" of energy map of data in each class as LDB and DPP are. Instead, by employing the energy maps of "every" data in each class, it finds a set of discriminant values at every base of wavelet packet dictionary that "truly" represents the discriminant power of every data.

Computational efficiencies of MDPP and MLDB are similar to those of DPP and LDB methods, respectively. Expectedly, the new methods have more data storage requirements since energy map of each training data must be saved for the evaluation of the entire relative entropy map.

In the following, some results and analysis of MDPP and MLDB are presented.

7.4. Data Analysis Results

To assess the effectiveness of the new proposed algorithms, MDPP and MLDB, and to compare them with each other and against DPP and LDB, the algorithms are applied on Ricardo Hydra test data. The new normalization method employed in sections 5.4 and 6.4 is also utilized.

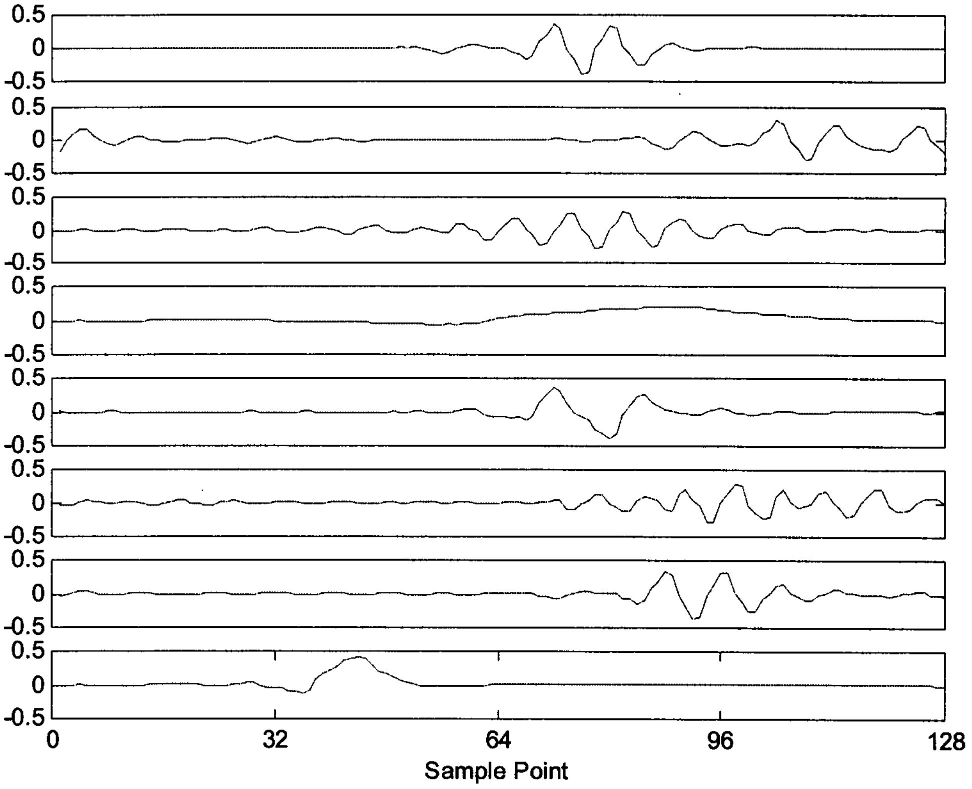


Fig. 7.2. The first 8 bases selected by MDPP using Coiflet1.

7.4.1. MDPP Classification

Applying MDPP algorithm on Ricardo Hydra test data with Coiflet1 wavelet results in selecting the bases plotted in Fig. 7.2 (only the first 8 bases have been shown). Their corresponding wavelet packet indices are as shown below. From first to third row are scale, oscillation, and translation indices, respectively.

4	5	5	5	4	5	4	3
3	6	8	0	2	9	3	0
3	0	0	3	6	0	4	2

If the above indices are compared with indices obtained from DPP (section 5.4) we observe that the second, fifth, and sixth bases of MDPP have also been selected by DPP. The rest of the bases selected by MDPP and DPP belong to similar frequency bands; as a result, the classification errors of both methods are almost identical. Fig. 7.3 shows DPP and MDPP classification errors for 32 different analyzing wavelets listed in section 7.2. Even though the general performance of MDPP is acceptable (a maximum classification error of 6% for most of the analyzing wavelets) DPP's performance is slightly better.

As the classification results of MDPP and DPP were close, we also examined the synthetic data set introduced in chapter 3. Fig. 7.4, the classification error evaluation of two methods, shows that the performance of MDPP is slightly superior with the synthetic data. Overall, the comparison of results from both techniques with two different types of data indicates that their performance is similar.

7.4.2. MLDB Classification

The wavelet packet indices associated with bases selected by applying MLDB algorithm on the same set of data are as follows:

4	7	4	4	4	5	7	7
3	33	3	2	3	9	35	34
3	0	4	6	5	0	0	0

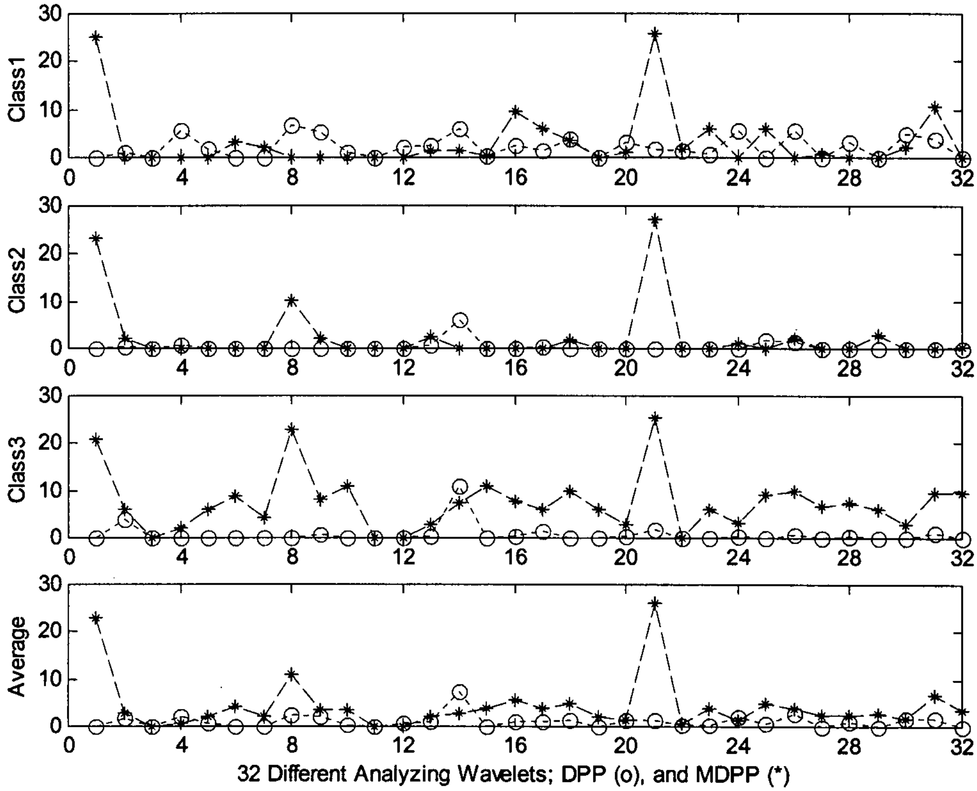


Fig. 7.3. DPP classification percentage error vs. MDPP for 32 different analyzing wavelets listed in section 7.2.

where the first to third row are again scale, oscillation, and translation indices of wavelet packet, respectively. The corresponding wavelet bases are plotted in Fig. 7.5. By comparing the above indices with indices obtained from LDB (section 6.4), we observe that the first basis of LDB indexed by (5,0,3) has not been selected by MLDB. This basis belongs to the interval (0, π/32) frequency band of wavelet packets which corresponds to (0, 0.4) KHz frequency band of the signal as

$$12.5 \text{ KHz} / 32 = 0.4 \text{ KHz}.$$

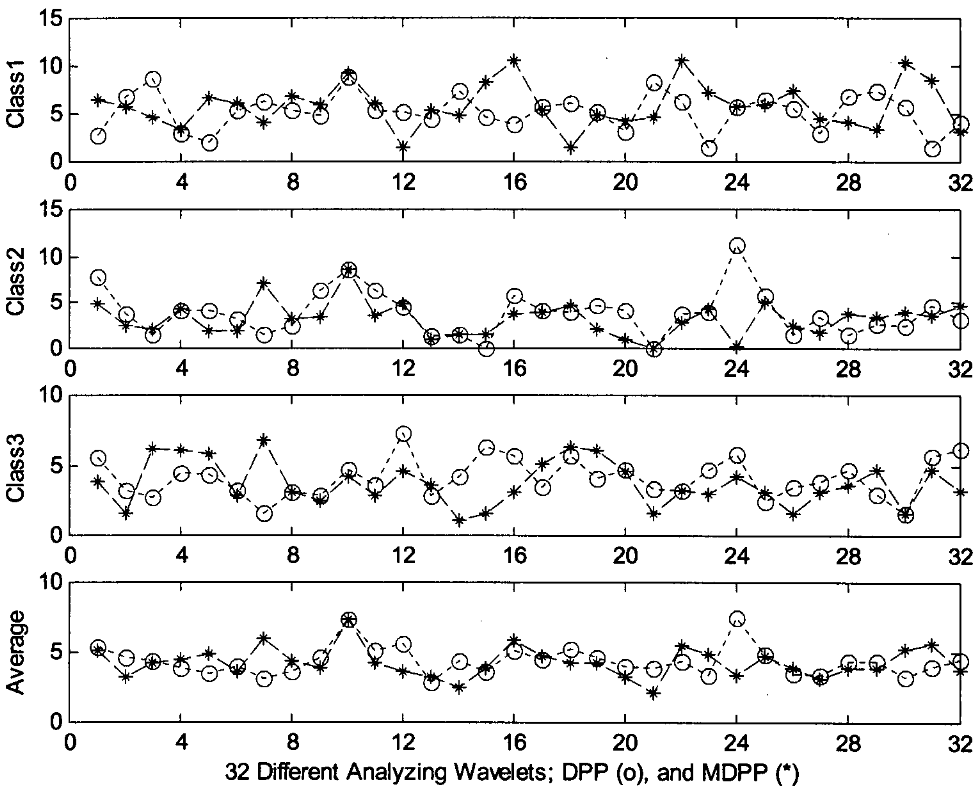


Fig. 7.4. DPP classification percentage error vs. MDPP for 32 different analyzing wavelets listed in section 7.2 with synthetic data set defined in chapter 3.

By examining typical signal spectrum shown in Fig. 3.5, one can notice that this frequency band is not located in a dominant frequency band of the signal; thus, it does not carry considerable energy. Since a combustion event can be viewed as a set of impulses, it is accompanied by a high energy release; a superior searching algorithm should readily pick those bases of wavelet packets, which are located in a high-energy node. As a result, the selection of this basis by LDB has not been a good choice.

On the other hand, the forth and eighth bases of MLDB have also been selected by LDB. The above frequency analysis shows that these bases along with other first three bases chosen by MLDB carry significant amount of energy and have been successful selections. In fact, the frequency band of wavelet packet nodes, which the bases 2 and 4 belong to, is located in the middle of the dominant frequency band of the signal (around π/4 or 3.1 KHz) and the one associated with bases 1 and 3 is placed at the beginning of the dominant frequency band (around 3π/16 or 2.3 KHz).

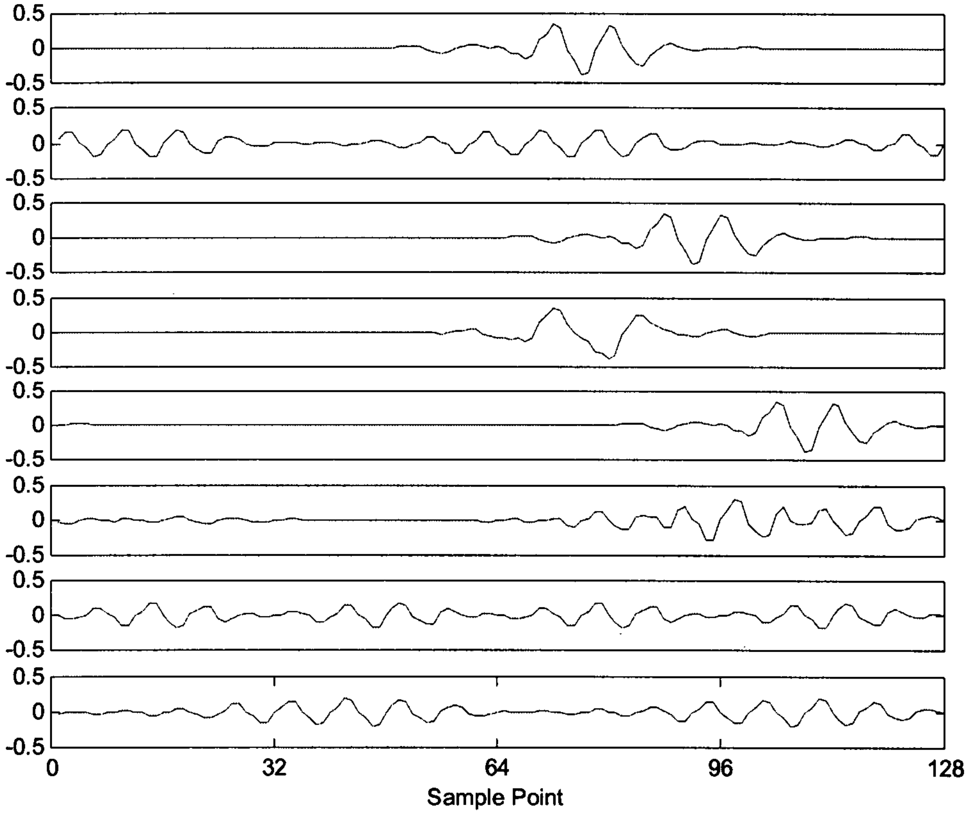


Fig. 7.5. The first 8 bases selected by MLDB using Coiflet1.

Fig. 7.6 illustrates discriminant measure of all 128 complete orthogonal bases selected by MLDB. By comparing this figure with Fig. 6.3, we can see that the discriminant measures of bases selected by MLDB possess higher values, with almost the same drop rate. Part of the large magnitude of discriminant measures may be attributed to the way that mutual approach calculates the relative entropy of the energy maps, in which all training data in each class are encountered, not just their average. Nevertheless, the error analysis shows that the mutual approach has a greater discriminant power.

Fig. 7.7 compares LDB and MLDB classification errors for the 32 different analyzing wavelets. It shows that for most of the analyzing wavelets MLDB performs better or as good as LDB, which shows the overall superiority of the proposed approach.

To assess the accuracy of the classification results we use the singular value decomposition (SVD) of the coefficient matrix. Its application, results and discussions are presented next.

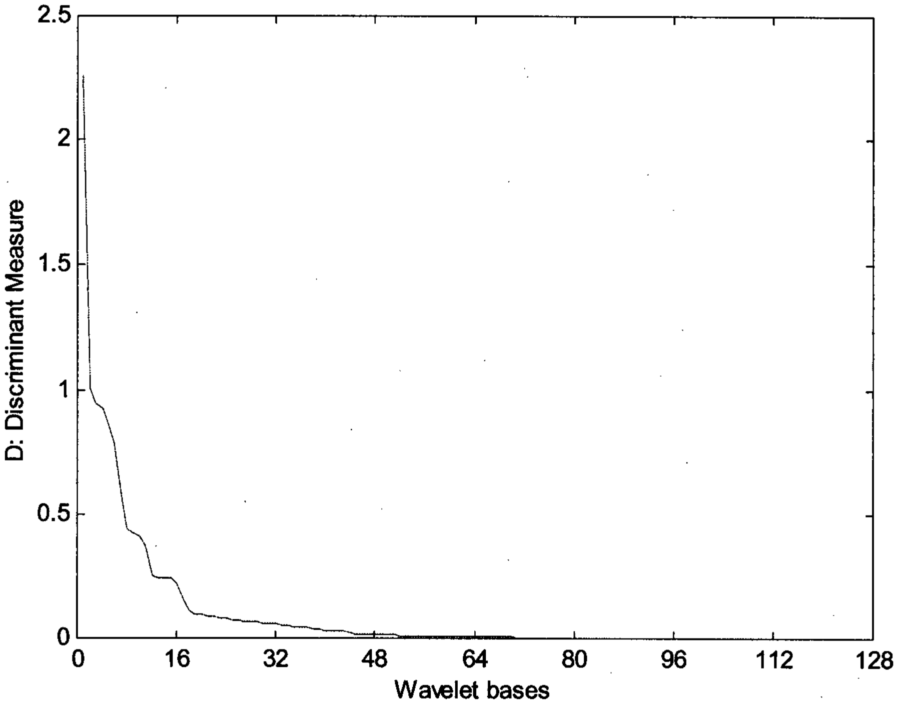


Fig. 7.6. Discriminant measure of the complete orthogonal 128 bases.

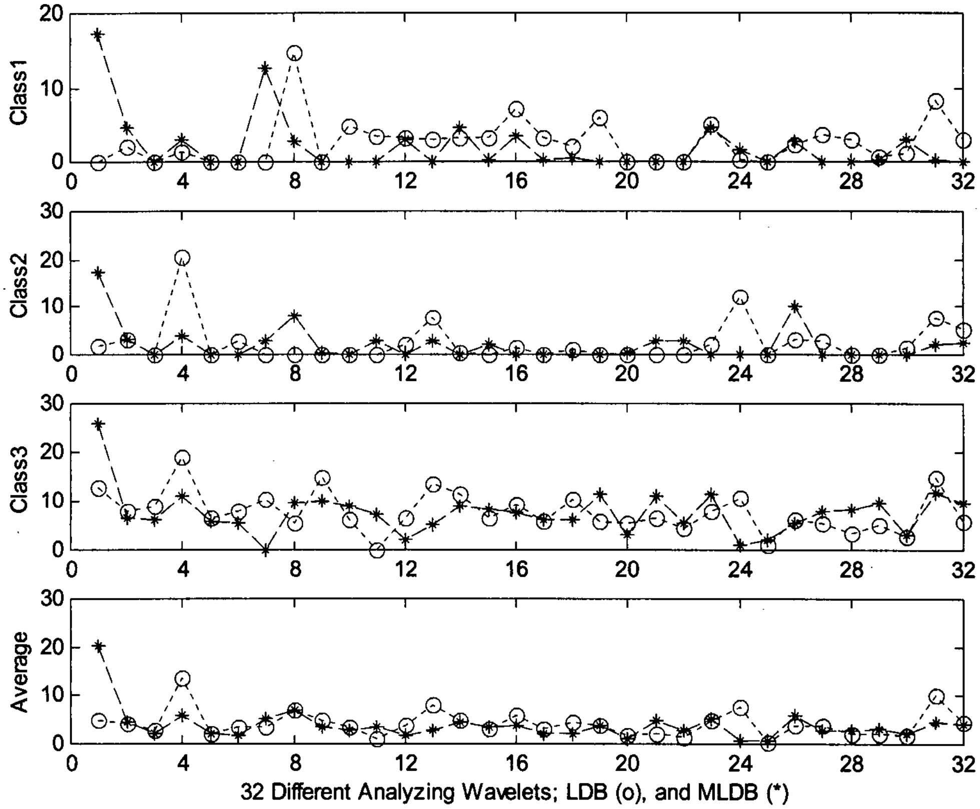


Fig. 7.7. LDB classification percentage error vs. MLDB for 32 different analyzing wavelets listed in section 7.2.

7.4.3. Analysis of Coefficient Correlation using Singular Values

We used SVD of the coefficient matrix, obtained by projecting data onto the selected bases, to determine the extent to which the feature variables, i.e., the coefficients, are correlated.

SV decay, Fig. 7.8, which is the drop rate from first to second, second to third, so on, is an important parameter in statistical analysis. For a matrix with large rank, usually the decay of the first few SVs is of interest; where rank is the maximum number of linearly independent rows (columns) [69]. In our case, where the coefficient matrix is of size 96 x 4, there are four SVs (the rank is only four); therefore, the decay from first to second SV is of great importance.

To obtain acceptable classification results, SV decay of coefficient matrix must be neither too large nor too small. Large SV decay shows that the useful information of coefficients is just in one direction (i.e., the direction of the eigenvector corresponding to that SV); therefore, the rest of selected directions, which are the bases found by the algorithm, contain redundant information. In other words, they are correlated with the first direction. On the other hand, choosing one direction to represent a multi-dimensional data is not usually a reasonable approach, specially noting that dimension reduction has already been implemented in the wavelet coefficient domain. Such a case shows that selected bases, from which coefficients are derived, do not contain all of the essential information about the system performance. Indicating that the respective algorithm has found a set of bases where only one of them includes useful information, the rest of the bases do not disclose anything new. As a result, a coefficient matrix with very high SV decay is not actually desirable. Similarly, having low SV decay means that all of bases (directions) have more or less the same information content, since there is a high correlation among them.

To support the above argument, in Fig. 7.8, we have shown four singular values of coefficient matrix corresponding to the first four bases selected by MDPP when applied on Ricardo Hydra test data. The figure demonstrates the SVs in each class and the average for all classes, along with the corresponding classification results. The horizontal axis numbered from 1 to 32 corresponds to the same analyzing wavelets given in section 7.2. Db2 analyzing wavelet (number 8 in Fig. 7.8) maintains a coefficient matrix with large SV decay in all of the classes, which leads to a relatively large classification error of over 10%. Conversely, Bior3.1 wavelet (number 21) produces a very low SV decay with mutually very close values, but still attains a high classification error of over 25%. Haar wavelet is also in the same category. Wavelets other than these three extreme cases have almost the same SV pattern with proper decay rate and produce an acceptable classification result – typically with less than 7% error. With highly variable data and the performance errors observed with other classification methods, 7% is considered as a relatively low and acceptable error.

Another interesting observation is that SVs in different classes shown in Fig. 7.8 follows a discernible pattern, in which SVs of different classes are quite distinct. For instance, the first singular values in each class associated with Coiflet1 (wavelet number 3) are 4.5, 3.6, and 2.2 for classes 1, 2, and 3, respectively, which show at least 20% difference among different classes. This is considered as an important aspect of the algorithm which extracts the information that makes different classes distinguishable from each other.

7.4.4. Application of Different Analyzing Wavelets

To attain a comprehensive view of the effect of using different analyzing wavelets in classification, several wavelets from various wavelet families, such as Daubechies, Coiflet, Symlet, and biorthogonal were used and tested in multiple runs. By examining Fig. 7.8 and our experience on other SVD graphs related to various data sets, we conclude that the use of different wavelets has no significant influence on the correlation structure of the coefficients. For this reason, the classification errors for most of the wavelets are almost the same with only minor deviations. We can then postulate that the improved classification is due to the algorithm.

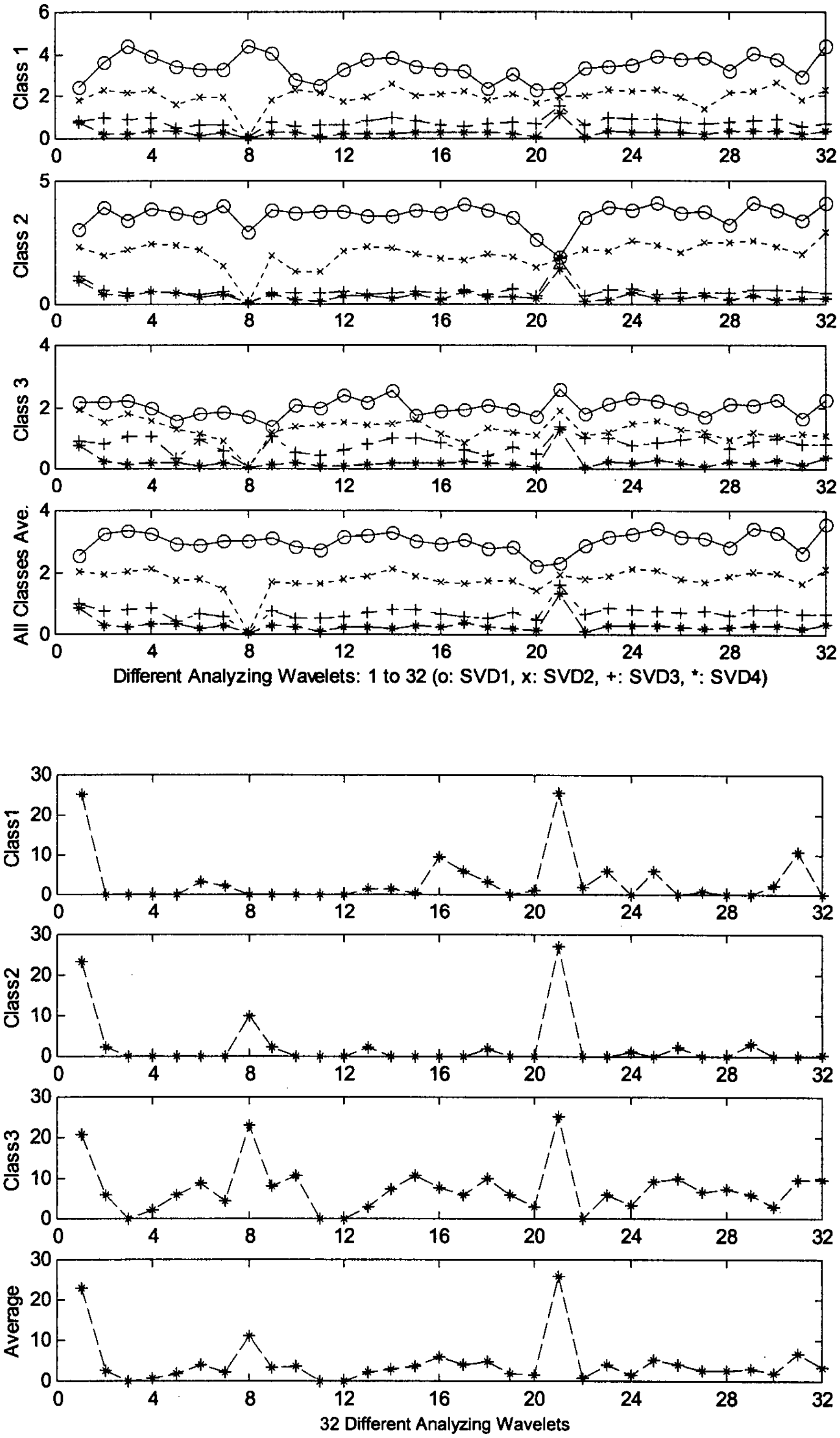


Fig. 7.8. Singular values of coefficient matrix corresponding to the first 4 bases selected by MDPP for 32 analyzing wavelets, along with consequent classification results.

7.4.5. LDB versus DPP

LDB and DPP methods were introduced and examined in previous chapters. During numerous simulations conducted while developing and examining MDPP and MLDB algorithms, a closer observation of LDB and DPP algorithms revealed that often the outcome of LDB and DPP methods were the same. For instance, three out of four of the selected bases in these two algorithms were the same when Coiflet1 was used as an analyzing wavelet on Ricardo Hydra data set. This can be traced to rather similar process they use in searching for the best set of basis. Both search methods are based on the first order statistics, in which the sum-squared coefficients in each class are used for the construction of the relative entropy measure. However, there are some differences in details which were explained in chapters 5 and 6. Classification results with different analyzing wavelets are also very similar for LDB and DPP, and the differences are within a few percentage points.

7.4.6. Different Neural Network Algorithms

An overview of different neural network backpropagation algorithms was introduced in section 2.9. It was mentioned that among 12 different backpropagation learning rules, Levenberg-Marquardt algorithm was found to be both fast and accurate. This was the overall observation using multiple runs with different analyzing wavelets. Table 7.1 shows the classification errors of the proposed cross-data entropy algorithm for each of these 12 methods, in which the Ricardo Hydra engine data (introduced in chapter 3) with Coiflet1 as analyzing wavelet was used.

Table 7.1. Classification error using different backpropagation algorithms.

Algorithm	Error (%)
Basic gradient descent	12.77
Gradient descent with momentum	11.18
Adaptive learning rate	5.22
Resilient backpropagation	3.81
Fletcher-Reeves conjugate gradient	4.49
Polak-Ribiére conjugate gradient	4.77
Powell-Beale conjugate gradient	4.05
Scaled conjugate gradient	4.24
BFGS quasi-Newton	4.08
One step secant	4.12
Bayesian regularization	4.84
Levenberg-Marquardt	2.95

7.5. Conclusions

The importance of normalization method in the wavelet domain was shown in this chapter. As one of the steps during preprocessing and to improve classification results, we modified DPP algorithm by applying an appropriate normalization, as defined in chapter 3. It was demonstrated that with the new normalization scheme a more accurate classification results can indeed be obtained.

A novel method, referred to as mutual or cross-data entropy approach, was then presented. Using this approach, two well-known discriminant algorithms in classification were modified. It was shown that the new methods are as efficient as the previous methods, and that for MLDB the classification results are consistently superior under a wide range of analyzing wavelets with both real and synthetic data.

The classification results were influenced by the selection of appropriate bases; but the question was how we could relate the accuracy of classification to the selected bases. In this respect, we related the accuracy of the classification results to the correlation of coefficient matrix through singular value decomposition. We detailed the assessment process and interpreted the relevance and meaning of various drop rates.

CHAPTER 8

Conclusions

8.1. Introduction

In this chapter, a summary of work accomplished in this thesis is given, and an overall conclusion of the research conducted is presented. Particular contributions to the pattern recognition, classification, and fault diagnosis systems are specified. To conclude the chapter and the dissertation, possible future research directions are proposed.

8.2. Synopsis and Conclusions

With the objective of engine performance diagnosis, we employed wavelets for the analysis of vibration data collected at the cylinder head position of a number of internal combustion engines pertaining to normal operation and operation under faulty conditions. An outline of the work carried out in this thesis may be stated as follows.

The importance of condition monitoring for fault detection and prevention in modern industrial settings was emphasized. The specific case of internal combustion engines and possible engine faults were introduced. As obtaining in-cylinder pressure information is prohibitive due to high cost and difficulty in retrofitting of pressure sensors in the existing engines, the significance of using acceleration data (vibration signals) of cylinder block in detecting these faults were highlighted. A literature review of previous works on machine fault diagnosis was also presented.

We reviewed basic concepts of statistical pattern recognition and classification, in addition to fundamentals of wavelet theory and its comparison to Fourier and short time Fourier transforms. We noted that acquired acceleration data exhibit variations from cycle to cycle, which is caused by variation in the combustion process, fuel-air mixture and variation in inertial system response of the cylinder head assembly to the combustion events. As such, information about a particular state of the combustion process for fault diagnosis is to be sought in the statistical behavior of acceleration data. Accordingly, the use of statistical pattern recognition methods using training data was judged to be suitable for fault detection in the application considered.

It was shown that the expansion of the concept of entropy into the domain of wavelet transform and its application on the sequence of wavelet coefficients, demonstrate a very promising approach for tackling fault diagnosis problems. Significant order reduction was achieved using wavelet transform where a countable number of feature variables in the coefficient domain were considered to be sufficient for discriminatory classification of different engine faults. Along the same line, two well-known approaches in pattern recognition and classification, referred to as dictionary projection pursuit (DPP) and local discriminant bases (LDB) were introduced, in which relative entropy – as a logical extension of entropy for using in classification – as well as wavelets were utilized.

Normalization and preprocessing are essential parts of a classification system. An appropriate normalization method, which can successfully be employed in both time domain and wavelet coefficient domain, was introduced. A neural network classifier was needed in the training and classification stage. For the training phase, embedded in the algorithm, many exploratory runs with different neural network algorithms showed that for the nature of machine data used, Levenberg-Marquardt backpropagation learning rule performed consistently and reliably. This was maintained as the main classifier for all the results reported in this thesis.

DPP and LDB were applied on a set of real world machine data and their classification results were given. By critically examining both methods, their shortcomings were revealed. We showed that none of these methods were using relative entropy in a theoretically correct manner. To overcome this dilemma, a novel method, referred to as cross-data entropy map or mutual-based approach was presented, in which a correct interpretation and application of relative entropy were made. In this approach, each training data contributes to the evaluation of relative entropy for discrimination purposes, contrary to the previous methods, which only use the first order statistics. Consequently, a superior classification results were obtained with a range of engine test data.

8.3. Contributions

Contributions of this thesis are as follows:

- Successfully applying wavelets to engine diagnosis using statistical pattern recognition methods and neural network classification.
- Utilizing wavelet packets for signal decomposition in which redundancy of signal decomposition by wavelet packets was used to select wavelet basis for a discriminatory classification of different faults. Furthermore, effectively associating singular value decomposition and correlation of coefficient matrix with the classification error.
- Successfully relating signal energies in wavelet packet nodes to the node(s) selected by the search algorithm. Demonstrating that a superior search algorithm can indeed select the nodes with high energy.
- Introducing a normalization method in DPP that improved classification results considerably – from 5% error in the original method to 1% in the new approach.
- Introducing a novel method, referred to as cross-data entropy approach for discriminatory classification and demonstrating the effectiveness of the method for entropy-based feature extraction using dictionary of orthogonal bases.
- Employing the new cross-data entropy approach in DPP search algorithm, and conducting tests with promising results.
- Employing the proposed approach in LDB search algorithm, and demonstrating superior results with various real-world data.

8.4. Future Research Direction

As a result of this research, a novel methodology for fault diagnosis was developed. It would be possible to extend the application of this method for the detection of malfunctions in multi-cylinder industrial engines. The initial results (conducted in [70], but not reported in this thesis) indicate that the extension as proposed above, is promising although there are still a number of obstacles to overcome. In industrial engines, due to several signal disturbance sources, that exhibit themselves as background noise, as well as the cross-talk effects generated from events which are simultaneous or are at the proximity of combustion in a given cylinder, detection of malfunction in combustion event in one cylinder poses a very challenging problem. In other words, differentiating abnormal from normal operation cannot be easily achieved. Fault diagnosis of multi-cylinder engines is considered a natural extension of this project for future research.

Another area as an extension of this research is the application of the proposed methodology to other faults that occur in internal combustion engines such as cylinder and ring wear, injection system problems, cracked teeth in gear train, and loose or cracked bearings.

Thus potential opportunities for the expansion of the project and application of the modified DPP and LDB methods to other classification problems in other areas do exist. One may continue exploring the possible extension of fault diagnosis for developing new techniques/methodologies in areas dealing with biomedical applications and analysis of signal in EEG (Electroencephalogram or brain waves) and EMG (Electromyography or electrical activity of the muscles).

In the course of this work we had collaboration with other research groups in our department and benefited from accessing their research engines for data collection. Similar collaborations with other research teams can be considered in future which can lead to a multi-disciplinary research project in the areas of vibration analysis, signal processing, automatic control, internal combustion engine research, and clean energy.

References

[1] H. Malm, *Fundamentals of reciprocating machinery analysis*, REM Technology Inc. (1994).

[2] J. B. Heywood, *Internal combustion engine fundamentals*. McGraw-Hill Inc. (1988).

[3] K. Englehart, B. Hudgins, P. A. Parker, and M. Stevenson, *Improving myoelectric signal classification using wavelet packets and principal component analysis.* Proceedings, IEEE (1999).

[4] B. Samimy G. Rizzoni, *Mechanical signature analysis using time frequency signal processing: Application to Internal Combustion Engine Knock detection.* Proc. of IEEE, Vo. 84 No.9 (Sep. 1996).

[5] G.T. Zheng, P.D. McFadden, *A time-frequency distribution for analysis of signal with transient components and its application to vibration analysis*. Trans. ASME Vol 121 (July 1999).

[6] B. Samimy G. Rizzoni, A. M. Sayeed, D. L. Jones, *Design of training data–based quadratic detectors with application to mechanical systems*. Proc. of ICASSP-96, Atlanta, GA (1996).

[7] F. Millo, C. V. Ferrro, *Knock in S.I engines: a comparison between different techniques for detection and control*. SAE technical paper series, # 982477 (1998).

[8] Q. Huang, Y. Liu, H. Liu, L. Cao, *A new vibration diagnosis method based on the neural network and wavelet analysis*. SAE technical paper series, 2003-01-0363 (2003).

[9] R. Tafreshi, F. Sassani, H. Ahmadi, and G. Dumont, *Local discriminant bases in machine fault diagnosis using vibration signals*. Journal of integrated computer-aided engineering, Vol. 9 (2004). In Press.

[10] S. Ortmann, M. Rychetsky, M. Glesner, *Engine knock detection using multi-feature classification by means of nonlinear mapping*. ISATA Conference (1997).

[11] R. Worret, S. Benhardt, F. Schwarz, and U. Spicher, *Application of different cylinder pressure based knock detection methods in spark ignition engines*. SAE papers, 2002-01-1668 (2002).

[12] S. Carstens-Behrens, and J. F. Bohme, *Fast knock detection using pattern signals*. IEEE international conference on acoustics, speech, and signal processing, Vol. 5, 3145-3148 (2001).

[13] R. Tafreshi, H. Ahmadi, F. Sassani, and G. Dumont, *Informative wavelet algorithm in diesel engine diagnosis*, The 17th IEEE international symposium on intelligent control (ISIC'02) 361-366 (2002).

[14] R. Rubini, and U. Meneghetti, *Application of the envelope and wavelet transform analyses for the diagnosis of incipient faults in ball bearings*, Mechanical systems and signal processing 15(2-3) 287-302 (2001).

[15] P. W. Tse, Y. H. Peng, and R. Yam, *Wavelet analysis and envelope detection for rolling element bearing fault diagnosis- their effectiveness and flexibilities*. Journal of vibration and acoustics, ASME, Vol. 123 (2001).

[16] H. Ahmadi, G. Dumont, F. Sassani, and R. Tafreshi, *Performance of informative wavelets for classification and diagnosis of machine faults*. International journal on

wavelets, multiresolution and information processing (IJWMIP), Vol. 1, No. 3 (2003).

[17] W. J. Wang, P. D. McFadden, *Application of wavelets to gearbox vibration signals for fault detection.* Journal of sound and vibration, 927-939 (1996).

[18] D. Chen, and W. J. Wang, *Classification of wavelet map patterns using multi-layer neural networks for gear fault detection.* Mechanical systems and signal processing 16(4), 695-704 (2002).

[19] H. Zheng, Z. Li, and X. Chen, *Gear fault diagnosis based on continuous wavelet transform.* Mechanical systems and signal processing 16(2-3) 447-457 (2002).

[20] B. Liu, S. F. Ling, *On the selection of informative wavelets for machinery diagnosis.* Mechanical systems and signal processing, Vol. 13, No 1, 1999.

[21] R. R. Coifman, M. V. Wickerhauser, *Entropy-based algorithm for best basis selection.* IEEE Transactions on information theory, 38,713-718 (1992).

[22] N. Saito, R. R. Coifman, *Local discriminant bases.* Wavelet applications in signal and image processing II. Proc. SPIE, Vol. 2303 (1994).

[23] K. Englehart, B. Hudgins, P. A. Parker, and M. Stevenson, *Classification of the myoelectric signal using time-frequency based representations.* Medical engineering and physics, special issue: Intelligent data analysis in electromyography and electroneurography, Vol. 21, pp. 431-438 (1999).

[24] N. Saito, R. R. Coifman, Frank B. Geshwind, Fred Warner, *Discriminant feature 24 using empirical probability density estimation and a local basis library.* Pattern Recognition Volume: 35, Issue: 12, December, pp. 2841-2852 (2002).

[25] K. Englehart, B. Hudgins, P. A. Parker *A wavelet based continuous classification scheme for multifunction myoelectric control.* IEEE Transactions on biomedical engineering, Vol. 48, Issue 3, Page(s): 302 –311 (March 2001).

[26] S. Mallat, Z. Zhang, *Matching Pursuit with Time Frequency Dictionaries.* IEEE Transactions on signal processing, 41, 3397-3415 (1993).

[27] J. H. Friedman, and W. Stuetzle, *Projection pursuit regression*, Amer. Statist. Asso., Vol. 76, pp. 817-823 (1981).

[28] J. Yang, L. Pu, Z. Wang, Y. Zhou, X. Yan, *Fault detection in a diesel engine by analyzing the instantaneous angular speed.* Mechanical systems and signal processing, 549-564 (2001).

[29] M. K. Zavarehi, D. R. Schricker, *Method and system for determining and absolute power loss condition in an internal combustion engine*, Patent No. US 6,199,007B1, Mar. 6 (2001).

[30] M. R. Dellomo, *Helicopter gearbox fault detection: a neural network based approach.* Journal of vibration and acoustics (July 1999).

[31] W. J. Wang, *Wavelets for detection mechanical faults with high sensitivity.* Mechanical systems and signal processing, 685-696 (2001).

[32] J. Shiroishi , Y. Li, S. Liang, T. Kurfess, S. Danyluk, *Bearing condition diagnostics 32 vibration and acoustic emission measurements.* Mechanical systems and signal processing, 693-705 (1997).

[33] J. Ma, L. Luo, Q. Wu, *A filter design method based on combination wavelets.* Mechanical systems and signal processing, 767-772 (1997).

[34] G. Luo, D. Osypiw, M. Irle, *Real-time condition monitoring by significant and natural frequencies analysis of vibration signal with wavelet filter and autocorrelation enhancement*, Journal of sound and vibration, 413-430 (2000).

[35] G. Rutledge, G. Mclean, *Dictionary project pursuit: a wavelet packet technique for waveform feature extraction*. PhD dissertation, University of Victoria (2001).

[36] G. Strang, T. Nguyen, *Wavelets and filter banks*, Wellesley-Cambridge Press (1996).

[37] S. Theodoridis, and K. Kourtroumbas, *Pattern recognition*. Academic Press (1999).

[38] L. Fausett, *Fundamentals of neural networks, architectures, algorithms, and applications*. Prentice Hall (1994).

[39] A. V. Oppenheim, R. W. Schafer, *Discrete-time signal processing*. Prentice Hall (1999).

[40] D. Gabor, *Theory of communication*. J. IEE, 93, 429-457, (1964).

[41] C. S. Burrus, R. A. Copinath, and H. Guo, *Introduction to wavelets and wavelet transforms*. Prentice Hall (1998).

[42] I. Daubechies, *Ten lectures on wavelets*. SIAM CBMS-NSF conference series (1992).

[43] S. Mallat, *A wavelet tour of signal processing*. Academic Press (1999).

[44] R. R. Coifman, Y. Meyer, and M. V. Wickerhauser, *Wavelet analysis and signal processing, Wavelets and their applications*. Jones, Barlett, Ruskai, editors, 153-178, Boston (1992).

[45] C.E. Shanon, *A Mathematical Theory of Communication*. The Bell systems technical journal, 27, 379-423 (1948).

[46] David W. Scott, multivariate density estimation : theory, practice, and visualization, Wiley, New York, 1992.

[47] C. K. Chui, An introduction to wavelets, volume 1, Academic Press (1992).

[48] J. Buckheit and D. Donoho, Improved linear discrimination using time-frequency dictionaries, Proceedings of SPIE Wavelet Applications in Signal and Image Processing III Vol 2569, 540–551 (1995).

[49] G. N. Saridis, *Stochastic processes, estimation, and control, the entropy approach.* John Wiley & Sons Inc. (1995).

[50] R. M. Gray, *Entropy and information theory.* Springer-Verlag (1990).

[51] Karmeshu, N. R. Pal, *Uncertainty, entropy and maximum entropy principle- and overview.* Studies in fuzziness and soft computing, Vol. 119, Entropy measures, maximum entropy principle and emerging applications, Karmeshu (editor), Springer (2003).

[52] S. Watanabe, T. kaminuma, *Recent developments of the minimum entropy algorithm.* Proceedings of the international conference on pattern recognition, IEEE, New York, pp. 536-540 (1988).

[53] M. T. Hagan, H. B. Demuth, M. Beale, *Neural network design.* PWS Publishing Company (1996).

[54] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stones, *Classification and regression tree*, Chapman and Hall, Inc, New York (1993), previously published by Wadsworth Inc. (1984).

[55] W. Härdle, L. Simar, *Applied multivariate statistical analysis.* Springer-Verlag Berlin Heidelberg (2003).

[56] J. H. Friedman and J. W. Tueky, *A projection pursuit algorithm for exploratory data analysis*. IEEE Transactions on computers, Vol. 23, 881-889 (1974).

[57] Joseph B. Kruskal, *Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which Optimizes a new "Index of* Condensation. Statistical Computation, R. Milton and J. Nelder (editors), Academic Press, New York, pp. 427-440 (1969).

[58] P. J. Huber, *Projection pursuit (with discussion)*. Annals of Statistics, 13(2):435–525 (1985).

[59] J. H. Friedman, *Exploratory projection pursuit*. Journal of American statistical association, Vol. 82, No. 397, 249-266 (March1987).

[60] M. C. Jones and R. Sibson, *What is projection pursuit? (with discussion)*, Journal of the royal statistical association, 150(1):1–36 (1987).

[61] J. H. Friedman and W. Stuetzle, *Projection pursuit regression*. Journal of the American statistical association, 76(376):817–823 (1981).

[62] J. H. Friedman, W. Stuetzle, and A. Shroeder, *Projection pursuit density estimation*. Journal of the American statistical association, Vol. 79, No. 387, 599-608, (1984).

[63] L. O. Jimenez, D. A. Landgrebe, *Hyperspectral data analysis and supervised feature reduction via projection pursuit*, IEEE Transactions on geoscience and remote sensing, Vol. 37, No. 6, 2653-2667 (Nov. 1999).

[64] H. Lin, L. M. Bruce, *Parametric projection pursuit for dimensionality reduction of hyperspectral data*. Proceedings of IEEE international geoscience and remote sensing symposium, Vol. 6, 21-25, 3483-3485 (July 2003).

[65] B. Kuo and D. A. Landgrebe, *Hyperspectral Data Classification Using Nonparametric Weighted Feature Extraction.* International geoscience and remote sensing symposium, Toronto, Canada (June 2002).

[66] A. Ifarraguerri, Chein-I Chang, *Unsupervised hyperspectral image analysis with projection pursuit*, IEEE transactions on geoscience and remote sensing, Vol. 38, Issue 6, 2529–2538 (Nov. 2000).

[67] S.-S. Chiang; C.-I. Chang, I.W. Ginsberg, *Unsupervised target detection in hyperspectral images using projection pursuit.* IEEE transactions on geoscience and remote sensing, Vol. 39, Issue 7, 1380–1391 (July 2001).

[68] D. B. Trizna, C. Bachmann, M. Sletten, N. Allan, J. Toporkov, R. Harris, *Projection pursuit classification of multiband polarimetric SAR land images.* IEEE Transactions on Geoscience and Remote Sensing, Vol. 39, No. 11, 2380–2386 (Nov. 2001).

[69] K. Fukunaga, *Introduction to statistical pattern recognition*, Academic Press, Inc. (1972).

[70] R. Tafreshi, H. Ahmadi, F. Sassani, and G. Dumont, *Malfunction detection in multi-cylinder engines using wavelet packet dictionary*, to be presented in SAE 2005 Noise & Vibration Conference and Exhibition (May 2005).

[71] B.D. Ripley, *Pattern recognition and neural networks*, Cambridge University Press (1996).