University of Alberta

Data-Driven Fault Detection, Isolation and Identification of Rotating Machinery: with Applications to Pumps and Gearboxes

by

Xiaomin Zhao

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Mechanical Engineering

©Xiaomin Zhao
Fall 2012
Edmonton, Alberta

Permission is hereby granted to the University of Alberta Libraries to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only. Where the thesis is converted to, or otherwise made available in digital form, the University of Alberta will advise potential users of the thesis of these terms.
The author reserves all other publication and other rights in association with the copyright in the thesis and, except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatsoever without the author's prior written permission.

Abstract

Fault diagnosis plays an important role in the reliable operation of rotating machinery. Data-driven approaches for fault diagnosis rely purely on historical data. Depending on how a diagnosis decision is made, this thesis divides data-driven fault diagnosis approaches into two groups: signal-based approaches and machine-learning-based approaches. Signal-based approaches make diagnosis decisions directly using signal processing techniques. Machine-learning-based approaches resort to machine learning techniques for decision making. There are three main tasks in fault diagnosis: fault detection (detect the presence of a fault), fault isolation (isolate the location/type of the fault), and fault identification (identify the severity of the fault). This PhD research studies signal-based approaches for fault identification and machine-learning-based approaches for fault detection, isolation and identification.

In signal-based approaches for fault identification, generating an indicator that monotonically changes with fault progression is a challenging issue. This thesis proposes two methods to generate such indicators. The first method uses multivariate signal processing techniques to integrate information from two sensors. The second method uses fuzzy preference based rough set and principal component analysis to integrate information from any number of sensors.

In machine-learning-based approaches, feature selection is an important step because it improves the diagnosis results. For fault detection and isolation, classification is often used as the machine learning algorithm. In this thesis, a feature selection method based on neighborhood rough sets is proposed for classification. Coming to fault identification, classification is not suitable because classification does not utilize the ordinal information within the different fault severity levels. Therefore, this thesis proposes to use another machine learning algorithm, ordinal ranking, for fault identification. A feature selection method based on correlation coefficient is proposed for ordinal ranking as well. Moreover, an integrated method which is capable of conducting fault detection, isolation and identification is proposed by combining classification and ordinal ranking.

The proposed methods are applied to fault diagnosis of impellers in slurry pumps and

fault diagnosis of gears in planetary gearboxes. Experimental results demonstrate the effectiveness of the proposed methods.

Acknowledgements

I would like to express my sincere appreciation to my supervisor, Dr. Ming J Zuo, for his support, care, encouragement and enthusiastic supervision throughout my study and research. His extensive discussions around my work and interesting explorations are of great value to this thesis. Dr. Zuo taught me how to question thoughts and express ideas. Under his guidance, I have gained not only valuable academic training but also the logical way of thinking to deal with problems effectively.

My warm thanks are due to Dr. Carlos Lange who directed me in Computational Fluid Dynamic (CFD) studies. His patient guidance and valuable advice are very helpful for my research.

I am very grateful to my PhD committee members, Dr. Jin Jiang (the University of Western Ontario), Dr. Witold Pedrycz, Dr. Yongsheng Ma, and Dr. Mike Lipsett for providing their precious time to examine my thesis.

Many thanks to my colleagues and friends: Dr. Yaguo Lei, Dr. Tejas H Patel, Dr. Qinghua Hu, Zhiye Zhang, Jia Wang, Vera Dickau, Hoifen Xu, Jihuan Yin, Xingyu Li, Yuning Gao, former and current members in our Reliability Research Lab, and many others for their various forms of help and support.

I am indebted, forever, to my parents and sisters, for their understanding, unconditional love and endless encouragement.

Contents

List of Tables

List of Figures

List of Abbreviations

ANN: Artificial Neural Network

BEPQ: Best Efficient Point Flow Rate

CFD: Computational Fluid Dynamics

EMD: Empirical Mode Decomposition

FDI: Fault Detection and Isolation

FDII: Fault Detection, Isolation and Identification

FFNN: Feed Forward Neural Network

FFT: Fast Fourier Transform

IMF: Intrinsic Mode Function

KNN: K-Nearest Neighborhood

LED: Leading Edge Damage

MZ: Mean Zero-one error (classificaiton error)

MA: Mean Absolute error

PC: Principal Component

PCA: Principal Component Analysis

PNN: Probabilistic Neural Network

PRM: Revolution Per Minute

SVM: Support Vector Machine

TED: Trailing Edge Damage

Chapter 1

Introduction

1.1 Background

Rotating machines are widely used in various industries including power, mining, aerospace and oil sands. Pumps and gearboxes are important types of rotating machinery. A pump is a device used to move fluids such as liquids or slurries. A gearbox is a device used to provide speed and torque conversions from one rotating power source to another device. With the increase of operation time and/or the change of working conditions, the performance of rotating machines might inevitably encounter an unexpected degradation.

Take slurry pumps in oil sands as an example. As slurries contain abrasive and erosive solid particles, the slurry pump impellers are subjected to harsh direct impingement [18]. Figure 1.1 shows a damaged impeller of a slurry pump after certain period of operation. With this damaged impeller, the designed function can not be achieved, e.g. the head of the pump would drop. If the fault is not detected in advance, it will continue progressing. As a result, the performance of the pumping system is affected, even unexpected downtime might be caused along with the economic loss.

Fault diagnosis checks the health condition of rotating machines. More specifically, it detects a fault, isolates the fault location/type, and identifies the fault severity. The diagnosis information is helpful in scheduling preventive maintenance or other actions to prevent serious consequences. Therefore, fault diagnosis plays an important role in the operation of rotating machinery, including improving safety, increasing efficiency and lifetime, and reducing downtime and total cost. The above benefits can be achieved only when fault diagnosis provides reliable information on machine health conditions. If incorrect diagnosis results are generated, ineffective maintenance would be arranged which would result in system failure or shutdown. An alarming fact is that one-third to one-half of the maintenance cost is wasted because of ineffective maintenance [19]. Therefore, there is a need to develop and improve the quality of fault diagnosis.

Next, the terminology relevant to fault diagnosis is formalized in Section 1.1.1.

Figure 1.1: A damaged slurry pump impeller (provided by Syncrude Canada Ltd.)

1.1.1 Basic Definitions

In 1991, a Steering Committee called Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS) was created within the International Federation of Automatic Control (IFAC). In 1993, SAFEPROCESS became a technical committee within IFAC. One important initiative of this committee was to define a common terminology for fault diagnosis [20]. In this thesis, the terminology presented in [5, 14, 20, 21] is followed, as described below.

A **"fault"** is considered as an unexpected change of system behavior such that it either deteriorates the performance of the system or demolishes the normal operation of the system. The term **"failure"** indicates a serious breakdown of the whole system. A failure is usually the result of the progression of a fault over time and could lead to hazardous consequences.

"Fault diagnosis" refers to detecting faults and diagnosing their locations/types and severities. In another word, fault diagnosis consists of three tasks which are defined as follows [5, 14].

- **Fault detection**: to make a binary decision whether everything is fine (no fault exists) or something has gone wrong (fault exists).

- **Fault isolation**: to pinpoint the mode of the fault or its location. Take gears for example. Some common fault types for gears are pitting, crack and missing teeth; different fault locations can be gear A, gear B and gear C. Fault isolation is to make decision whether the fault is pitting, crack or missing teeth; and/or to determine whether the fault occurs on gear A, gear B or gear C. It can be seen that fault isolation needs to

make decision from multiple options (when more than two fault types/locations are considered), whereas fault detection makes decisions from two options only.

- **Fault identification**: to estimate the fault severity. The fault severity can be specifically described by the fault size or more generally, the fault level (e.g. slight fault, moderate fault or severe fault). Fault identification needs to make decision from multiple options (when more than two fault sizes/levels are considered) as does fault isolation.

The first two tasks are the first step in fault diagnosis, and are often considered together as fault detection and isolation (FDI) [5, 22]. The third task (fault identification) assesses the severity of an identified fault and is also a very important aspect in fault diagnosis. For the convenience of description, "diagnosis of fault types" is used to indicate fault detection and isolation, and "diagnosis of fault levels" is used to indicate fault identification. In the following section, different approaches to conduct fault diagnosis are summarized.

1.1.2 Fault Diagnosis Approaches

There are mainly two approaches for fault diagnosis: model-based and data-driven [5].

"The model-based fault diagnosis can be defined as the determination of the fault in a system by comparing available system measurements with a priori information represented by the system's analytical/mathmatical model, through generation of residual quantities and their analysis. A residual is a fault indicator that reflects the faulty condition of the monitored system." [23]. Figure 1.2 shows the two main stages for model-based fault diagnosis: residual generation and residual evaluation (also called decision making). In the first stage, a residual is obtained by comparing available system measurements with priori information represented by the system's analytical model. In the second stage, the generated residual is inspected and fault condition is determined by applying a decision rule, e.g. a threshold test.

The advantage of model-based fault diagnosis is its ability to incorporate the physical understanding of a system for diagnosis. The disadvantage, however, is that it is difficult to build a mathematical model for the system with a large number of inputs, outputs and state variables. A detailed review of model-based fault diagnosis is referred to [21].

Data-driven fault diagnosis relies purely on historical data which include information on different known health conditions of machinery. There are three categories of historical data [6]:

- Value type - Data collected at a specific time epoch is a single value, e.g. oil analysis data and temperature data.

- Waveform - Data collected at a specific time epoch is a time series, e.g. vibration data and acoustic data.

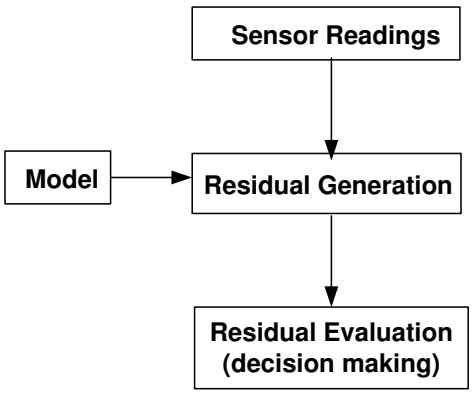


Figure 1.2: Model-based fault diagnosis [5, 6]

- Multidimension type - Data collected at a specific time epoch is multidimensional, e.g. X-ray images.

Among them, waveform data, especially vibration data, are most regularly measured in fault diagnosis of rotating machinery. A survey [24] shows that, of the 18 commercially available condition monitoring systems for wind turbines, 12 systems provide vibration monitoring. In this thesis, only vibration data will be used.

This thesis further divides the data-driven fault diagnosis into two groups, depending on how a diagnosis decision is made using the historical data. The first group makes decisions directly based on data analysis results using signal processing techniques, and is called the signal-based approach in this thesis. The second group resorts to machine learning techniques for decision making, and is called the machine-learning-based approach in this thesis. Their block diagrams are given in Figure 1.3.

Data processing is the first step in both groups of methods. In signal-based approaches, a few fault-sensitive indicators are firstly generated using signal processing techniques. **An indicator** is a parameter that represents the health condition of a machine. Then the fault indicators are checked and compared with certain thresholds for fault diagnosis. If the indicator values are above certain thresholds, then a fault is said to be presented. This method is straightforward, and the results are easy to be interpreted. The generation of effective indicators is the key issue of this method. To achieve this, proper signal processing techniques need to be used depending on specific objectives. A review on signal processing techniques will be given in Section 1.2.1.

In machine-learning-based approaches, data processing step usually produces a large number of features to ensure that health information is fully obtained. **A feature** describes the information on machine health conditions. This step is also called feature extraction. The next step, called feature selection, is used to reduce the number of features so as to improve the performance of machine learning. Details are given in Section 1.2.2.2. Finally, a machine learning algorithm is utilized to build a model which maps features in the feature space to fault conditions in the fault space. This model is used for fault diagnosis.

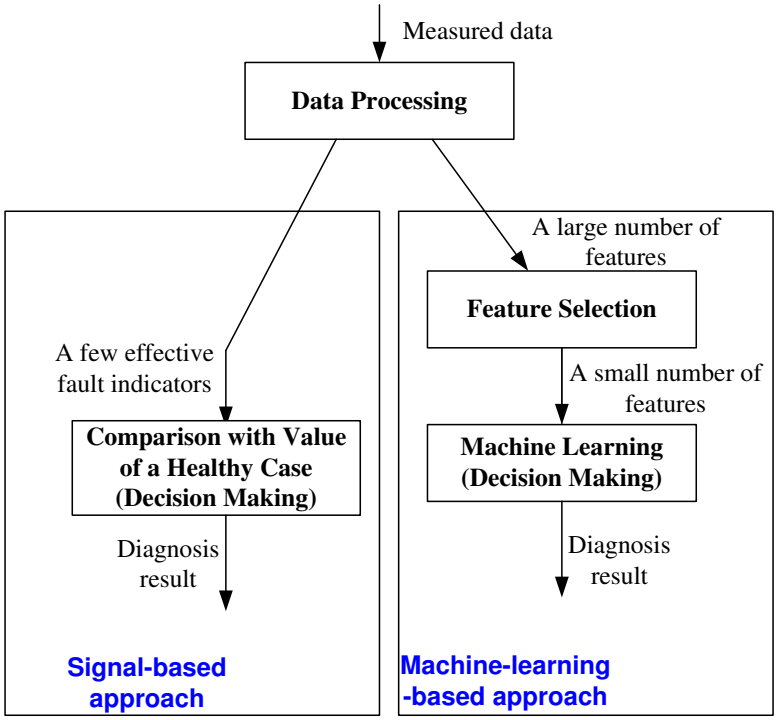


Figure 1.3: Data-driven fault diagnosis

Machine-learning-based approaches can achieve automatic fault diagnosis. The disadvantage, however, is that a large number of training samples are needed in order to build a good model.

This thesis focuses on data-driven fault diagnosis. A literature review on the two kinds of approaches for data-driven fault diagnosis is provided next.

1.2 Literature Review on Data-Driven Fault Diagnosis

In this section, signal-based approaches and machine-learning-based approaches in data-driven fault diagnosis are reviewed.

Data processing is used in both signal-based and machine-learning-based approaches, as shown in Figure 1.3. In Section 1.2.1, the signal-based fault diagnosis is reviewed and signal processing techniques are discussed. These techniques can be applied for feature extraction within the machine-learning-based approaches as well.

1.2.1 Signal-Based Fault Diagnosis

There are three main categories of vibration data analysis: the time-domain analysis, the frequency-domain analysis, and the time-frequency domain analysis. Each of them is discussed in separate sections next.

1.2.1.1 Time-domain Analysis

Time-domain analysis is directly based on the time waveform (e.g. vibration on a bearing casing with respect to time) itself. The most straightforward technique is simply to visually inspect portions of the time waveform. However, vibration signals produced by a large machine containing many components may be very complicated when viewed in the time domain, making it unlikely that a fault be detected by a simple visual inspection.

Instead of visual inspection, the signal can be characterized using some statistical parameters. These parameters can be compared with predefined thresholds for exceeding (if exceeding, then a fault is indicated), and/or trended against time for tracking the deterioration of a machine. Statistical measures such as mean, standard deviation, peak, root mean square (RMS), crest factor, impulse factor, clearance factor, kurtosis, skewness are commonly used descriptive statistics [6, 25]. Their definitions are given in Equations (1.1)-(1.9).

The mean, standard deviation, peak and RMS values define the central tendency, the dispersion, the amplitude spread and the energy of the vibration signal, respectively. RMS has been used to perform fault detection, e.g. the ISO 2372 (VDI 2056) norms define three different velocity RMS alarm levels for four different machine conditions. These mentioned parameters have the same units as the vibration signal, and thus are dimensional.

Crest factor (CF), impulse factor (IF), clearance factor (CLF), kurtosis, and skewness are dimensionless statistics. The advantage of dimensionless statistics is that they are less sensitive to the variation of load and speed [26]. Crest factor, clearance and impulse factor are sensitive to the existence of sharp peaks, so they are often used for faults that involve impacting, e.g. rolling element bearing wear, gear tooth wear or cavitation in pumps [25]. High-order statistics, such as kurtosis and skewness, describe the shape of the amplitude distribution of the waveform. Kurtosis expresses an aspect of spikiness of the signal, i.e. how peaked/float the distribution is. A normal distribution has a kurtosis value of 3. If a signal contains sharp peaks, then its kurtosis will be higher [14]. Skewness is a measure of the asymmetry of the distribution. A symmetric distribution results in a skewness of 0. A machine in good condition usually has a Gaussian distribution, whereas a damaged machine usually has a non-Gaussian distribution [14]. Thus kurtosis and skewness can be used to indicate the presence of a fault as well.

$$x_{\text{mean}} = \frac{1}{N}\sum_{k=1}^{N} x(k) \tag{1.1}$$

$$x_{\text{std}} = \left(\frac{1}{N-1}\sum_{k=1}^{N}\left(x(k) - x_{\text{mean}}\right)^2\right)^{\frac{1}{2}} \tag{1.2}$$

$$x_{\text{peak}} = \max_k \left|x(k)\right| \tag{1.3}$$

$$x_{\text{RMS}} = \sqrt{\frac{1}{N}\sum_{k=1}^{N} x(k)^2} \tag{1.4}$$

$$\text{CF} = \frac{x_{\text{peak}}}{\sqrt{(\frac{1}{N}(x(k) - x_{\text{mean}})^2)}} \tag{1.5}$$

$$\text{CLF} = \frac{x_{\text{peak}}}{(\frac{1}{N}\sum_{k=1}^{N}\sqrt{|x(k)|})^2} \tag{1.6}$$

$$\text{IF} = \frac{x_{\text{peak}}}{\frac{1}{N}\sum_{k=1}^{N}|x(k)|} \tag{1.7}$$

$$\text{kurtosis} = \frac{\frac{1}{N}\sum_{k=1}^{N}(x(k) - x_{\text{mean}})^4}{\frac{1}{N}(\sum_{k=1}^{N}(x(k) - x_{\text{mean}})^2)^2} \tag{1.8}$$

$$\text{skewness} = \frac{\frac{1}{N}\sum_{k=1}^{N}(x(k) - x_{\text{mean}})^3}{\frac{1}{N}(\sum_{k=1}^{N}(x(k) - x_{\text{mean}})^2)^{3/2}} \tag{1.9}$$

In the above equations, $x(k)$ is the vibration amplitude at time point k, x_{mean} is the mean value of the signal $x(k)$, x_{peak} is the peak value of the signal and N is the total number of data points in the signal.

The above statistics are computationally easy. In signal-based approaches, they have been used as indicators for fault detection and isolation. However, coming to fault identification (diagnosis of fault levels), they demonstrate poor performance. For example, Sass et al. [27] reported that the values of statistics including RMS, kurtosis, impulse factor and crest factor decreased to the level of an undamaged case when the damage was severe. In other words, they can not distinguish the severe fault and the normal ("no fault") conditions.

By themselves, these statistics may not be effective for all faults. But they can be used to reflect certain health information, which makes them useful in machine-learning-based fault diagnosis. In this thesis, they are calculated and used, together with other features, in identification of gear pitting levels in Chapter 6.

1.2.1.2 Frequency-domain Analysis

Frequency-domain analysis is based on the transformed signal in the frequency domain. The frequency-domain analysis reflects a signal's constituent frequencies. The most widely used analysis is the spectrum analysis by means of Fast Fourier Transform (FFT).

The amplitude information of the Fourier transform (also known as the Fourier spectrum, or frequency spectrum) is mostly investigated in spectrum analysis, and can be used as fault indicators. Generally, if the amplitudes of characteristic frequencies are below certain thresholds, then the machine would be considered as normal [28]. Otherwise, if the amplitudes of characteristic frequencies are above certain thresholds, then these frequencies can be analyzed and related to certain fault types for fault isolation. Besides the amplitudes at

specific frequencies, the amplitudes can also be checked for particular frequency bands of interest [29].

The phase information of the Fourier spectrum, though not as popularly used as the amplitude information, provides important and complementary information. For example, McFadden [30] used the phase angle of the meshing harmonics to determine the location of a gear crack.

Based on the Fourier spectrum, other spectra such as power spectrum, cepstrum [31], bispectrum [32, 33], trispectrum [32], and holospectrum [34] have also been reported for fault diagnosis of rotating machinery [32, 35]. Power spectrum is the Fourier transform of a signal's autocorrelation function. It reflects the energy at a specific frequency. Cepstrum has several versions of definitions. Among them, power cepstrum (the inverse Fourier spectrum of the logarithmic power spectrum) is the most commonly used one. It can reveal harmonics and sideband patterns in power spectrum. Bispectrum and trispectrum are the Fourier transforms of the third- and the fourth- order statistics of the waveform. Details on these spectra are available [6].

The above mentioned spectra deal with one-dimensional signals only. Full spectrum [8, 36] considers two signals measured from two orthogonal directions together when doing Fourier transform and thus keeps the directivity information of a planar motion. Holospectrum [34] considers two or three signals together and integrates information of phase, amplitude and frequency of a waveform. Full spectrum and holospectrum have both been used for diagnosis of rotor faults, and are found to outperform the conventional Fourier spectrum [34, 37].

In this thesis, Conventional Fourier spectrum and full spectrum are used. Descriptions on conventional Fourier spectrum and full spectrum will be presented in Section 2.1.2. Full spectrum is employed for signal-based fault identification in Chapter 5. Conventional Fourier spectrum is used for feature extraction in machine-learning-based fault diagnosis in Chapters 4, 6 and 7.

1.2.1.3 Time-frequency Analysis

Time-frequency analysis investigates the vibration signals in both the time domain and the frequency domain. This analysis captures frequency contents at different time points. Wavelet Transform (WT) and Empirical Mode Decomposition (EMD) are modern time-frequency analysis tools.

Wavelet transforms are inner products between signals and the wavelet family, which are derived from the mother wavelet by dilation and translation. Let $\psi(t)$ be the mother wavelet which is a zero average oscillatory function centered around zero with a finite energy. The daughter wavelet is $\psi_{a,b}(t) = \psi((t-b)/a)$, where a is positive number and defines the scale and b is any real number and defines the time shift. By varying the parameters a and b, different daughter wavelets can be obtained to constitute a wavelet family. The

continuous wavelet transform is to perform the following operation:

$$W(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\psi^*(\frac{t-b}{a})\,\mathrm{d}t, \tag{1.10}$$

where ψ^* is the complex conjugate of ψ. Lin and Zuo [38] used wavelet analysis for the detection of early tooth crack in a gearbox. A review of the application of wavelet transform in fault diagnosis is given in [39].

Empirical mode decomposition (EMD) decomposes a signal into a set of intrinsic mode functions (IMFs). Each IMF stands for a generally simple oscillatory mode. A detailed introduction of EMD is presented in Section 2.2. EMD has the advantage over WT in that it does not need a predefined mother wavelet and the decomposition totally depends on the data itself. Gao et al. [40] investigated an EMD-based fault diagnosis for rotating machinery, and found that EMD could extract the fault indicators and identify the fault effectively.

In this thesis, EMD is utilized in signal-based fault identification in Chapter 5.

1.2.1.4 Summary

In literature, majority of signal-based fault diagnosis methods focus on fault detection and isolation [26, 35, 41], and recently more attention is given to fault identification [34, 42, 43].

To conduct fault identification, the signal-based approaches generate a fault indicator that varies monotonically with the fault progressive. Thus by comparing the value of this indicator with pre-determined thresholds, the fault severity level can be estimated. How to generate such an indicator is a challenging issue, especially for complex systems [44]. This issue will be further illustrated and two indicator generation methods will be proposed in Chapter 5 .

1.2.2 Machine-Learning-Based Fault Diagnosis

In machine-learning-based fault diagnosis, machine learning algorithms are employed for decision making (as shown in Figure 1.3). There are two categories in machine learning: supervised learning and unsupervised learning.

Supervised learning infers a function from a set of training samples. Each sample is a pair consisting of an input (i.e. a feature vector $\mathbf{x}$) and a label (i.e. a fault condition d). A supervised learning algorithm produces an inferred function between $\mathbf{x}$ and d. The inferred function can be used to predict the label (d) for any new input ($\mathbf{x}$).

Unsupervised learning discovers the particular patterns that reflect any kind of structure of the training samples, such as properties of the distribution and relations between samples. Each training sample is described by a feature vector ($\mathbf{x}$) only, and there is no target label provided for the sample. The training samples are used to explore the underlying structure of the data. Clustering is a typical example of unsupervised learning, in which data is grouped into clusters and each cluster stands for one health condition.

Supervised learning and unsupervised learning are two ways for solving diagnosis problems. The choice of the two depends on the data available. If historical information on health conditions, e.g. fault types and fault levels, are available, supervised learning can be used. In some cases, it may be extremely difficult to acquire fault data from real systems. To handle such situations, unsupervised learning can be used. In this thesis, supervised learning is employed. Unsupervised learning is not discussed; however, it does not mean that unsupervised learning is not important.

As stated in Section 1.1, machine-learning-based fault diagnosis consists of three steps: feature extraction, feature selection, and machine learning. The techniques for each of them are reviewed next.

1.2.2.1 Feature Extraction

Feature extraction is a process of obtaining features, either by direct reading from sensors (e.g. shaft rotational speed read from a tachometer) or by calculating through signal processing techniques. With the development of signal processing techniques, many features can be extracted. For example, the statistics in the time-domain, the amplitudes at some frequencies in the frequency-domain, and decomposition results (e.g. wavelet coefficients) in the time-frequency domain can all be used as features. Signal processing techniques, from the time-domain, the frequency-domain or the time-frequency domain, have already been covered in Section 1.2.1, and are not repeated here.

Feature extraction is not the focus of this thesis. Existing techniques from the time-domain and the frequency-domain are directly used for feature extraction in Chapters 4, 6 and 7.

1.2.2.2 Feature Selection

In machine-learning-based fault diagnosis, the diagnosis results highly rely on the performance of the model that is trained using training samples. Each sample is described by a set of features. The number of features is the dimension of the sample. It might be expected that the inclusion of an increasing number of features would include more information for building the model. Unfortunately, that is not true if the number of the training samples doesn't also increase rapidly with each additional feature included. This is the so called curse of dimensionality [45]. The basic idea of the curse of dimensionality is that high-dimensional data is difficult to work with for several reasons [46]: (1) adding more features can increase the noise and hence the error; (2) there are not enough samples to get good estimates for parameters in the model. For example, Silverman [1] provided Table 1.1 illustrating the difficulty of density estimation in high dimensional space. It can be seen that with the increase of the number of dimension, the required sample size (number of samples) increases sharply. Dimension reduction is thus desired. There are two ways to conduct dimension reduction. The first one is transformation-based reduction, which involves an irreversible

Table 1.1: Required sample size in density estimation [1]

No. of dimensions	Required sample size
1	4
2	19
5	786
7	10700
10	842000

transformation (linear or nonlinear) in the reduction process. Principal component analysis (PCA) can be used as a dimension reduction tool by choosing the first few components and leaving the rest unimportant principal components. Other commonly used transformation-based reduction tools include Project Pursuit (PP) [47], Isomap [48], etc. Because of the linear or nonlinear transformation, the transformation-based reduction cannot preserve the physical meaning of the original data set. So it is often used in situations where the semantics of the original data set are not needed by any future process.

The second one, called feature selection, keeps the physical meaning of the original feature set. Instead of creating new features as does transformation-based reduction, feature selection chooses a feature subset from the original feature set. It removes redundant features and irrelevant features from the original feature set, so as to improve the learning ability of a machine learning algorithm. In this thesis, feature selection techniques are used for dimension reduction.

Methods for feature selection generally fall into three categories: filter, wrapper and embedded methods [49]. Filter methods firstly rank the features with an ad hoc measure and then find a feature subset based on that ordering. In wrapper methods, a machine learning algorithm is used to score features. In embedded methods, feature selection is performed in the process of training. Compared with the other two categories of feature selection methods, filter methods are independent of any machine learning algorithms, and thus are computationally simple and fast. Therefore, this thesis works on filter methods.

Generally speaking, there are two important components in filter methods: (1) a measure to evaluate the performance of a feature or a feature subset, and (2) a search strategy to find the best feature subset as defined by the corresponding measure. The two components are reviewed next.

1.2.2.2.1 Evaluation Measure Measures in filter methods evaluate the performance of a feature or a feature subset in distinguishing different labels (i.e. fault conditions). The reported measures can be grouped into distance-based (e.g. Fisher criterion [50] and Relief [51]), correlation-based (e.g. correlation coefficient [52]), information-based (e.g. mutual information [53]), and consistency-based (e.g. FOCUS [54, 55] and rough set [55, 56]) measures. In the following, each of them is briefly discussed (see [55,57] for more details).

(1) Distance-based measures evaluate the difference between samples having different labels induced by a feature. A feature x_i is preferred to another feature x_j if x_i induces a greater difference than x_j. If the difference is zero, then x_i and x_j are indistinguishable. Distance-based measures are employed in [50, 51, 58].

(2) Correlation-based measures quantify the ability of one variable to predict another variable. Correlation coefficient is a classical correlation measure and can be used to find the correlation between a feature and the label. If the correlation of feature x_i with label d is higher than the correlation of feature x_j with d, then x_i is preferred to x_j. Correlation-based measures are employed in [52, 59].

(3) Information-based measures typically determine the information gain from a feature. The information gain from a feature x_i is defined as the difference between the prior uncertainty and the expected posterior uncertainty in distinguishing different labels using x_i. Feature x_i is preferred to x_j if the information gain from feature x_i is greater than that from x_j. Information-based measures are employed in [53, 60].

(4) Consistency-based measures are characteristically different from the above three measures because of consistency-based measures' heavy reliance on the data itself and use of Min-feature bias in selecting a subset of features [61]. Min-feature bias prefers consistent hypotheses definable over as few features as possible. These measures find out the minimal size of the feature subset that satisfies an acceptable consistency. Consistency-based measures are employed in [10, 54, 55, 62, 63].

The first three types of measures (i.e. distance-based, information-based and correlation-based) are univariate measures. That is, they check one feature at a time. The consistency-based measures are multivariate measures which check a set of features at a time. The consistency-based measure can help remove both redundant and irrelevant features; whereas the other measures may not be able to do so by themselves [55]. In order to do so, the maximum relevance and minimum redundancy scheme was proposed in [53]. Under this scheme, the relevance between a feature and the label, as well as the redundancy between two features, are evaluated based on the first three types of measures listed above. The selected feature subset is the one that maximizes the relevance and minimizes the redundancy. Formula (1.11) [53] shows an example to realize this scheme, where C is the original feature set, S is the selected feature subset with m features, $V(x_i, d)$ is the relevance between feature x_i and the label variable d, $M(x_i, x_j)$ is the redundancy between features x_i and x_j, and x_i and x_j are two features in S. By employing this scheme, the first three types of measures can remove both redundant and irrelevant features.

$$\max_{x_i \in S, S \subset C} \quad \frac{1}{m} \sum_{x_i \in S} V(x_j; d) - \frac{1}{m^2} \sum_{x_i, x_j \in S} M(x_i, x_j) \tag{1.11}$$

Correlation coefficient and rough set are two powerful and popularly used tools for features selection. The basic ideas of these two techniques are briefed as below. A detailed introduction on these two techniques will be given in Section 2.4 and Section 2.5, respec-

tively.

The absolute value of the Pearson correlation coefficient is the most commonly used correlation-based measure. It evaluates the correlation between two variables. If the two variables are features, then it evaluates the redundancy between the two features (i.e. $M(x_i, x_j)$ in Formula (1.11)); if one variable is a feature and the other is the label, then it evaluates the relevance between a feature and the label (i.e. $V(x_i, d)$ in Formula (1.11)). The absolute value of the Pearson correlation coefficient takes values between 0 and 1. The value of 1 means the two variables are perfectly correlated; the value of 0 means the two variables are not related. The optimum feature subset is the one that maximizes the objective function in Formula (1.11).

Rough set, first described by Zdzislaw I. Pawlak [64], is a formal approximation of a crisp set (e.g. the set of labels, D) in terms of a pair of sets which give the lower and the upper approximations of the original set (D). A parameter, called Approximation quality (also called dependency), is defined to evaluate the approximation ability. A feature subset which approximates the label set (D) better has higher values of approximation quality. The selected feature subset is the one that has the highest approximation quality and minimum number of features.

The Pawlak rough set model is best suited for nominal features which are discrete variables. In fault diagnosis, it is often the case that numerical features (i.e. real-values), must be dealt with. To consider this situation, Hu et al. [10] extended the Pawlak rough set model to a neighborhood rough set model. One problem with applying the neighborhood rough set model in fault diagnosis is determining the neighborhood size. This problem will be further discussed and a modification to the neighborhood rough set model will be proposed in Chapter 4.

Another issue with the existing feature selection methods is that most of them are proposed for classification problems which make them work efficiently for fault detection and isolation (FDI), but inefficiently for fault identification. The reason lies in the fact that the types of fault conditions (labels (d)) in fault identificaiton and FDI are different. In FDI, the labels are nominal variables, whereas in fault identification, the labels (e.g. slight fault, moderate fault, severe fault) are ordinal variables. This results in two different machine learning algorithms: the former is classification and the latter is ordinal ranking, as will be discussed in Section 1.2.2.3.

1.2.2.2.2 Search Strategy Besides evaluation measure, search strategy is the other important component for filter methods. The aim of the search is to optimize the value of an evaluation function, such as Formula (1.11). Techniques of search strategy are categorized into three groups by Doak [65]:

- Exponential (e.g. exhaustive and Branch & Bound (BB) [66]).
- Randomized (e.g. heuristics such as simulated annealing [67], probabilistic hill-

climbing [68], genetic algorithms (GA) [69], and partial swarm optimization (PSO) [70]).

- Sequential (e.g. sequential backward, sequential forward and bidirectional [71]).

In exhaustive search, all $2^{m(m-1)} - 1$ possible combinations of m input features are used in the search process. Hence, although exhaustive search ensures an optimal solution, it is the most computationally expensive approach. The BB approach decreases the search time by pruning some branches of the search tree. Randomized heuristics further speed up the search process. These algorithms incorporate randomness into their search procedure to escape local maxima. For different seeds of the random number generator, randomized heuristics may give different results. Sequential search adds or removes features sequentially and is computationally fast, but has a tendency to become trapped in local maxima.

The widely used sequential search algorithms are sequential forward selection (SFS), sequential backward selection (SBS), and bidirectional search (BS) [65]. SFS starts with an empty set of features ($S = \emptyset$). As search starts, features are added into S one at a time. At each time, the best feature among unselected ones is chosen based on the evaluation measure (i.e. the criteria in filter methods). S grows until it meets a stopping criterion. The stopping criterion can be that the number of selected features (say q) is reached, or the value of the evaluation measure has reached a threshold. SBF begins with a full set of features and removes one at a time. At each time, the least important feature is removed based on the evaluation measure. S shrinks until it meets a stopping criterion, e.g. q features are left or a certain value of the evaluation measure is reached. BS starts in both directions, i.e. two searches proceed concurrently. SFS is used in the case where the most important feature is easy to find. SBS is used when the least important feature is easy to find. BS is used when both the most important and the least important features are easy to find.

The search strategy is not the focus of this thesis. The SFS sequential search is directly adopted because it is the most computationally economic one [65].

1.2.2.3 Machine Learning

After the feature selection in Section 1.2.2.2, now the third step is machine learning. As stated in the beginning of Section 1.2.2, this thesis focuses on supervised learning only. A supervised learning algorithm analyzes the training data and infers a function which represents the relationship between the input (i.e. a feature vector $\mathbf{x}$) and a label (i.e. fault condition d). According to the label's type, a supervised learning problem can be grouped into three categories. (1) If the label is continuous, the problem is called regression. Otherwise the label is discrete, which results in the last two categories: (2) if the label is a nominal variable, the problem is called classification; (3) if the label is an ordinal variable, the problem is called ordinal ranking [72]. A detailed description on different types of variables are given in Section 2.3. For the convenience of description, we call the label in a classification problem, **class**, and the label in an ordinal ranking problem, **rank**.

In fault diagnosis, classification algorithms are often used, such as K-nearest neighborhood (KNN) [73], Artificial Neural Network (ANN) [74] and Support Vector Machine (SVM) [75]. They are briefly described as follows.

K-nearest neighborhood (KNN) is a type of learning algorithm where the inferred function is only approximated locally. In KNN, a sample is classified by a majority vote of its neighbors, with the test sample being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). Figure 1.4 illustrates the concept of KNN. The circle point is a test sample. The rectangles and the triangles are training samples, of which five triangles are from Class A and the six squares are from Class B. If $k = 3$, then we examine the 3 nearest neighbors of the test sample. The circle is assigned to Class A because there are 2 triangles and only 1 square inside the inner circle including the test sample and its 3 nearest neighbors. If $k = 5$, it is assigned to Class B (3 squares vs. 2 triangles inside the outer circle). KNN is amongst the simplest of all machine learning algorithms. One drawback of KNN is that the classes with the more frequent samples tend to dominate the prediction of the test sample, as they tend to come up in the k nearest neighbors when the neighbors are computed due to their large number. Application of KNN to fault diagnosis is reported in [73, 76]

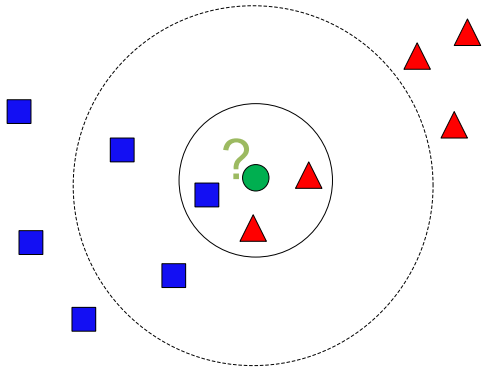


Figure 1.4: An example of KNN classification

Artificial neural network (ANN) is a mathematical model that mimics the human brain structure. A neural network is composed of a large number of highly interconnected processing elements (called neurons or nodes). This structure allows the approximation of an underlying nonlinear function between the input and the output.

There are various types of neural networks. The feed forward neural network (FFNN) is the first and arguably simplest type of artificial neural network devised. In this network, the information moves in only one direction, forward, from the input layer, through the hidden layers (if any) and to the output layer. At each neuron, a function (called an activation function) of the inputs is computed. Figure 1.5 shows a schematic representation of an FFNN with one hidden layer where the circles stand for nodes. In this example, the input

layer contains m nodes: one for each of the m input features of a feature vector. The output layer has one node meaning that the output d is a scalar. The links carry weights w_{ij} and w'_{ij}. The weights can be found in the training process by minimizing the mean-square error (Equation (1.12)) using the Backpropagation (BP) training algorithm.

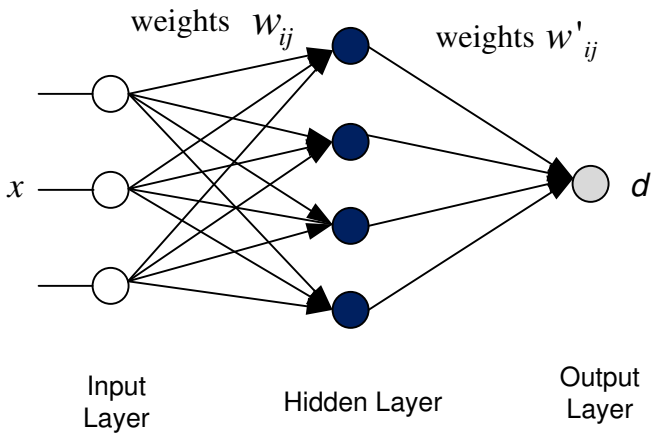


Figure 1.5: Illustration of FFNN

$$\text{Error} = \frac{1}{n}\sum_{i=1}^{n}(d_i - d'_i)^2 \tag{1.12}$$

In Equation (1.12) n is the total number of training samples, d_i is the desired output of the i^{th} sample and d'_i is the output by the network. This network is powerful for problems where the relationships between the input and the output may be quite dynamic or non-linear. However, it has slow convergence speed and may result in a local minimum value instead of the global optimal solution.

Probabilistic neural network (PNN) [77] is another type of feed forward neural network. It has four layers: input layer, pattern layer, summation layer and output layer. Figure 1.6 displays the architecture of a PNN that recognizes 3 classes, but it can be extended to any number of classes. The input is a three-dimensional feature vector. The pattern nodes are divided into groups: one group for each of the 3 classes. Each pattern node in the group for class i corresponds to a Gaussian function (i.e. $e^{-\frac{\|x-x_{ij}\|^2}{\sigma^2}}$) centered on its associated feature vector in the i^{th} class. All of the Gaussian functions in a class feed their values to the same summation node. There are 3 summation nodes, and each node represents one class. The output is the class that has the maximum function value. The training process of a PNN is essentially the act of determining the value of the smooth parameter σ. The PNN offers the following advantages [78]: rapid training speed and robustness to noisy training samples. The disadvantage of PNN is that it requires representative training samples.

Application of ANNs to fault diagnosis is reported in [79–81], and it is often used when many training samples are available.

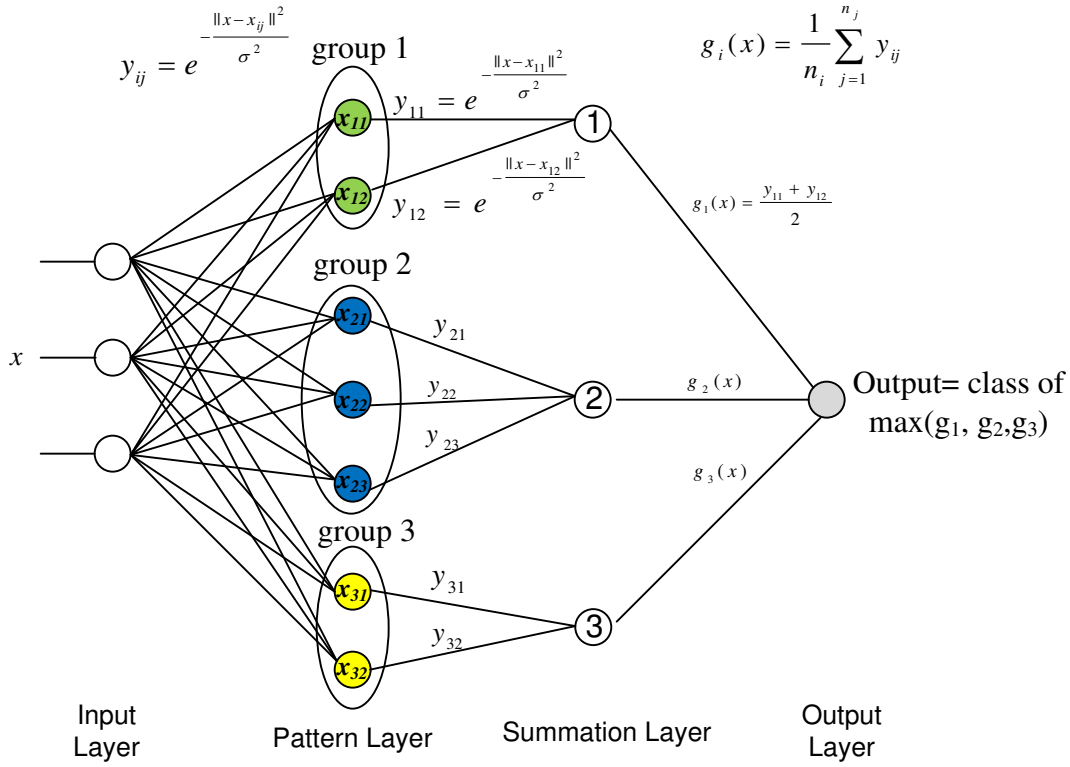


Figure 1.6: An example of PNN classification

Support Vector Machine (SVM) maps the original feature space into a higher dimensional space, in which a hyperplane (or a set of hyperplanes) is constructed to achieve linear separation. Figure 1.7 illustrates the concept of SVM classification. SVM can achieve the global optimal solution and has good ability in generalization. Application of SVM to fault diagnosis is reported in [82, 83]. Detailed mathematical explanation of SVM will be given in Section 2.6.1.

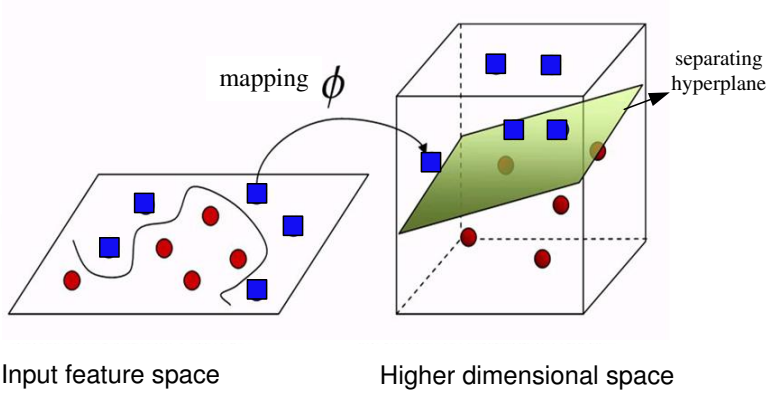


Figure 1.7: Illustration of SVM classification [7]

In this thesis, KNN, PNN and SVM are adopted as classification algorithms for decision making in fault detection and isolation (diagnosis of fault types).

These classification algorithms, when applied to fault identification (diagnosis of fault levels), ignore the ordinal information among different fault levels [84]. For example, a moderate fault is worse than ("<") a slight fault but is better than (">") a severe fault. In classification, however, the severe, moderate and slight faults are parallel to each other and are not compared using ">" and "<" operations. The above mentioned ordinal information is the main characteristic of the fault levels, which makes the fault identification (diagnosis of fault levels) more complicated than fault detection and isolation (diagnosis of the fault types). In order to express the ordinal information, this thesis uses another machine learning technique, ordinal ranking, for fault identification.

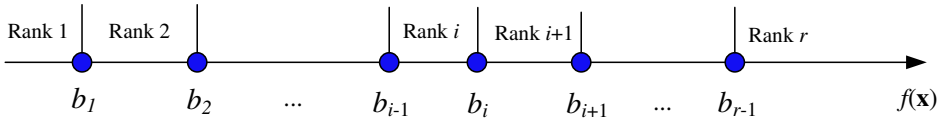


Figure 1.8: Illustration of ordinal ranking

Ordinal ranking generates a ranking model that expresses the ordinal information contained in the training data. Figure 1.8 shows an example of a ranking model trained by ordinal ranking. Ordinal ranking finds a $f(\mathbf{x})$ which is calculated using the input features $\mathbf{x}$, and the value of $f(\mathbf{x})$ changes monotonically with the increase of the fault level. The rank of a sample can be estimated by checking the value of $f(\mathbf{x})$, that is, if $b_{i-1} < f(\mathbf{x}) < b_i$, the sample belongs to rank d_i. A detailed description on ordinal ranking will be given in Section 2.6.2. Ordinal ranking has been proved to perform better in ranking documents in the information retrieval field [85], but its application to fault diagnosis hasn't been reported yet. This thesis applies ordinal ranking for fault identification in Chapter 6 and Chapter 7.

1.2.2.4 Summary

In machine-learning-based fault diagnosis, classification is often employed for decision making. In fault detection and isolation (FDI), different fault types are diagnosed. Classification is able to distinguish different fault types, so it is suitable for FDI. In fault identification, fault levels are diagnosed. Classification can not express the ordinal information contained in fault levels, so it is not suitable for fault identification. Ordinal ranking, a recently studied machine learning algorithm, is able to keep ordinal information, and therefore is adopted in this thesis for fault identification in Chapter 6.

Moreover, FDI and fault identification are usually considered separately. An integrated technology which is capable of conducting fault detection, isolation and identification (FDII) is more helpful in fault diagnosis [21]. Researchers have used classification algorithms for FDII [83, 86]. However, as stated earlier, classification ignores the ordinal

information. So there is a need to combine classification and ordinal ranking techniques for FDII. This will be present in Chapter 7.

Feature selection plays an important role in improving the performance of machine learning. There are two issues to be studied for the existing feature selection methods. The first issue is the determination of certain parameters for some feature selection methods (e.g. neighborhood size in neighborhood rough set, as stated in Section 1.2.2.2). The second issue is that the majority of the existing feature selection methods are for classification problems, which work less efficiently for ordinal ranking problems. In this thesis, two feature selection methods are proposed for classification and for ordinal ranking in Chapter 4 and Chapter 6, respectively.

1.3 Objectives and Contributions of the Thesis

The objective of the PhD research is to improve the performance of signal-based and machine-learning-based techniques in fault detection, isolation and identification.

In signal-based approaches, as discussed in Section 1.2.1, generating an indicator for fault identification is a challenging issue. This indicator needs to show a monotonic trend with the fault level. To tackle this challenge, two methods of integrating information from two or more sensors are developed to generate an indicator for fault levels. Specifically, the two methods are listed below and are detailed in Chapter 5.

- Generating an indicator by processing signals from two sensors together
- Generating an indicator by processing signals from each individual sensor or from several sensors together and then combining information from different sensors

In machine-learning-based approaches, the problem of fault detection and isolation (FDI) can be regarded as a classification problem. Feature selection largely affects the performance of the machine learning algorithms. In this thesis, neighborhood rough set is adopted for feature selection for a classification problem. The problem of neighborhood size selection, as discussed in Section 1.2.2.2.1, is solved. Fault identification (i.e. diagnosis of fault levels) has an important characteristic, that is, there is ordinal information among different fault levels. In order to preserve the ordinal information, this thesis proposes to do fault identification using ordinal ranking as a machine learning algorithm. Moreover, most feature selection methods are for classification problems, and they work less efficiently for ordinal ranking problems [60]. In this thesis, a feature selection method based on correlation coefficients is proposed for ordinal ranking. Specifically, the following two topics are investigated and the results are provided in Chapter 4 and Chapter 6, respectively.

- A feature selection method based on neighborhood rough sets for fault detection and isolation
- A feature selection method based on correlation coefficients for fault identification

Furthermore, FDI (diagnosis of fault type) and fault identification (diagnosis of fault level) are usually considered separately in literature. An integrated methodology which is capable of conducting fault detection, isolation and identification (FDII) is more useful in fault diagnosis [21] and worths studying. This thesis proposes a machine-learning-based method for FDII (diagnosis of both fault type and fault levels). In this method, classification is used for diagnosis of fault type and ordinal ranking is employed for diagnosis of fault levels. Thus the last topic of this thesis addressed in Chapter 7 is:

- Fault detection, isolation and identification by combining classification and ordinal ranking.

1.4 Outline of the Thesis

The organization of this thesis is as follows. Chapter 2 presents the preliminary knowledge on techniques used in this thesis. Chapter 3 describes the experimental data collection for the verification of the proposed diagnosis methods. Chapter 4 studies the feature selection method based on neighborhood rough sets for fault detection and isolation. Chapter 5 presents two signal-based methods to generate indicators for fault identification. Chapter 6 applies ordinal ranking to the machine-learning-based fault identification. A feature selection method for ordinal ranking is proposed, and the advantage of ordinal ranking over classification is discussed. Chapter 7 proposes a scheme for fault detection, isolation and identification. Fault detection and isolation is conducted through classification, and fault identification is achieved through ordinal ranking. Finally, Chapter 8 summarizes the contributions and introduces the possible directions for future work.

Chapter 2

Background Information for Techniques Used in This Thesis

As reviewed in Chapter 1, signal processing, feature selection and machine learning are important components in fault diagnosis. In this chapter, fundamentals of these techniques are introduced. Two signal processing techniques, Fourier transform and empirical mode decomposition (EMD), are introduced in Section 2.1 and Section 2.2, respectively. The selections of feature selection algorithm and machine learning algorithm depend on the measurement scale of features and fault conditions. So in Section 2.3, different measurement scales are presented first. Then two techniques, correlation coefficient and rough set, that are often used in feature selection are introduced in Section 2.4 and Section 2.5, respectively. Two machine learning algorithms (classification and ordinal ranking) based on support vector machine (SVM) are introduced in Section 2.6. The uses of these techniques in later chapters are summarized in Section 2.7.

2.1 Fourier Transform

Fourier transform is a mathematical operation that decomposes a signal $x(k)$ which is a function of time into its constituent frequencies, known as a Fourier spectrum $X(n)$. The discrete Fourier transform of $x(k)$ is defined in Equation (2.1), where N is the total length of $x(k)$. The amplitude of the frequency spectrum is given by the magnitude of $X(n)$, i.e. $|X(n)|$.

$$X(n) = \sum_{k=0}^{N-1} x(k) e^{-j\frac{2\pi}{N}nk}, \text{ where } n = 0, 1, \ldots, N-1. \tag{2.1}$$

Let $x(k)$ be a vibration signal from direction X only. $x(k)$ is represented by real numbers (i.e. $x(k)$ is 1-dimensional), then $X(n)$ is the conventional Fourier spectrum (also called half spectrum). Let $y(k)$ be the other vibration signal measured from direction Y which is orthogonal to X. $x(k)$ and $y(k)$ together can be represented by complex numbers, i.e. $z(k) = x(k) + j \cdot y(k)$ where j is the imaginary unit. The Fourier spectrum of $z(k)$, $Z(n)$, is

called the full spectrum of $x(k)$ and $y(k)$. Half spectrum and full spectrum are explained in this section. Materials in this section are from [87] and [8].

2.1.1 Conventional Fourier Spectrum (Half Spectrum)

When Fourier transform is applied to real-valued data (i.e. data measured from one direction only), the conventional Fourier spectrum is obtained. In this spectrum, the negative frequency component is the complex conjugate of the positive frequency component. The positive and negative parts of the spectrum are mirror images of each other. Thus the amplitudes at a positive frequency (e.g. f_n) and its corresponding negative frequency (e.g. $-f_n$) are the same. This is shown in a simple example below.

Let $x(k) = 0.5\cos(40\pi k) + 0.2\cos(100\pi k) + \cos(-100\pi k)$. The amplitude of its conventional Fourier spectrum ($|X(n)|$) is plotted in Figure 2.1. It can be seen that the amplitude values at -20 HZ (respectively -50 Hz) and 20 Hz (respectively 50 Hz) are exactly the same. Because of this, only the positive frequency components need to be analyzed. Therefore, the conventional Fourier spectrum is also called **half spectrum**.

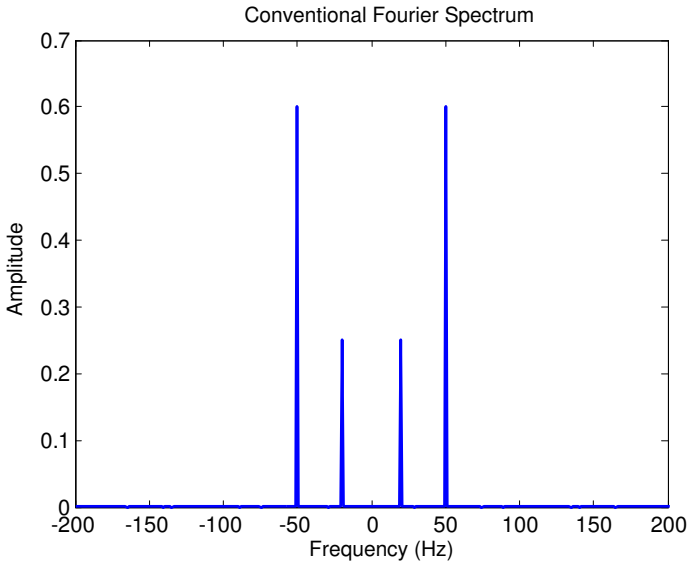


Figure 2.1: The amplitude of a conventional Fourier spectrum

Conventional Fourier spectrum can be used to analyze a vibration signal measured from one direction only. If the vibration motion occurs in a plane instead of a single direction, then two vibration signals can be measured from two orthogonal directions (X and Y). If the conventional Fourier spectrum is conducted on X and Y motions separately, the phase correlation between the X and Y motions can not be revealed. To overcome this limitation, full spectrum analysis needs to be applied.

2.1.2 Full Spectrum

Full spectrum analyzes vibration signals measured from two orthogonal directions in the plane of vibration motion together. Let $x(k)$ and $y(k)$ be two signals simultaneously measured from two orthogonal directions, and $z(k) = x(k) + j \cdot y(k)$ where j is the imaginary unit. The full spectrum of $z(k)$ can be obtained by replacing $x(k)$ by $z(k)$ in Equation (2.1). Figure 2.2 shows the detailed procedure. The input of the "FFT" module has two parts (the direct buffer and the quadrature buffer). The output of the "FFT" module also has two parts (the direct output and the quadrature output). Simultaneously sampled signals are put into the direct buffer and the quadrature buffer of the "FFT module", respectively. The direct output consists of positive frequency components and the negative frequency components. So does the quadrature output. Thus the amplitude of a frequency has two parts, one from the direct output and the other from the quadrature output. Consequently, the positive and negative frequency components of this FFT are usually not mirror images. In the right half of the full spectrum plot, the amplitudes of the positive frequency components (also known as forward frequency components) are shown. In the left half, the amplitudes of the negative frequency components (also known as backward frequency components) are shown. The term "forward" means that the rotation of this frequency component is in the direction of the planar vibration motion. The term "backward" means that the rotation of this frequency component is in the opposite direction of the planar vibration motion.

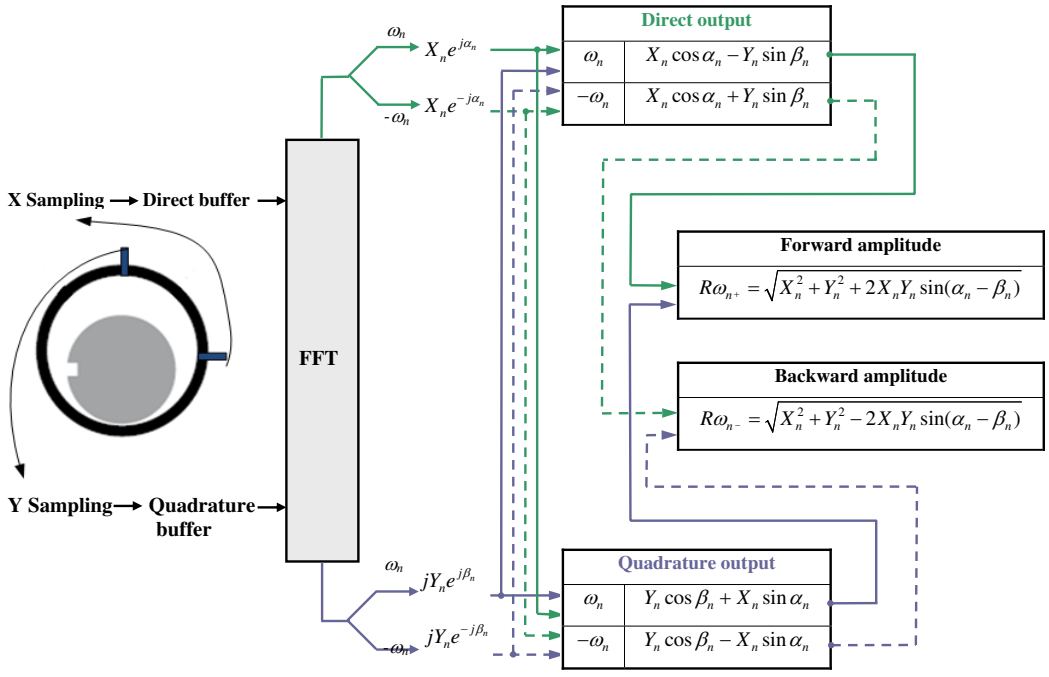


Figure 2.2: Mathematical procedure of obtaining a full spectrum [8]

An example is used to illustrate the characteristic of full spectrum. Let $x(k) = 0.5\cos(40\pi k)+$

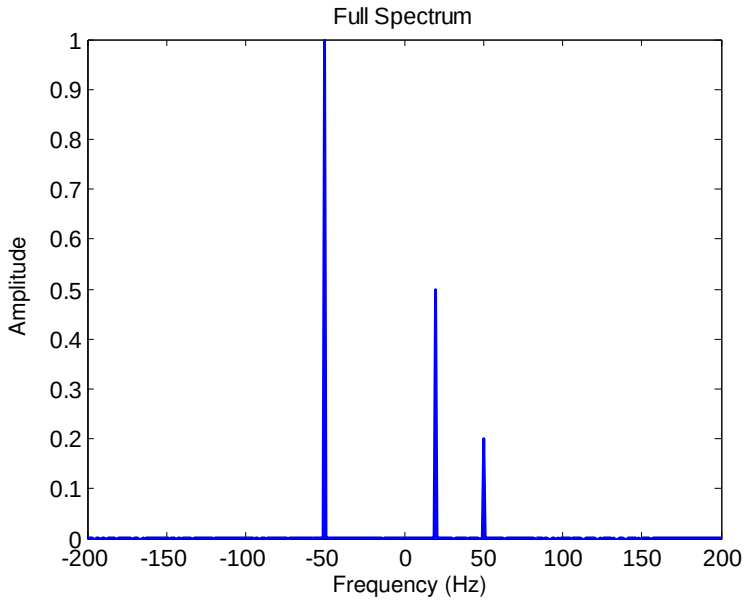


Figure 2.3: The amplitude of a full spectrum

$0.2\cos(100\pi k) + \cos(-100\pi k)$ and $y(k) = 0.5\sin(40\pi k) + 0.2\sin(100\pi k) + \sin(-100\pi k)$ be the two signals measured from horizontal direction and vertical direction, respectively. The two signals describe a planar motion consisting of three simple motions: (1) a counter-clockwise rotation at a frequency of 20 Hz with amplitude of 0.5; (2) a counter-clockwise rotation at 50 Hz with amplitude of 0.2; and (3) a clockwise rotation at 50 Hz with amplitude of 1. $x(k)$ and $y(k)$ are used as the direct buffer and the quadrature buffer, respectively. So the positive (respectively negative) frequency corresponds to counter-clockwise (respectively clockwise) direction. Figure 2.3 shows the amplitude of the full spectrum. In Figure 2.3, there are three peaks, at -50 Hz, 20 Hz and 50 Hz with amplitude of 1, 0.5 and 0.2, each of which corresponds to a simple motion. For example, the frequency component at -50 Hz having amplitude of 1 corresponds to the clockwise rotation at 50 Hz with amplitude of 1. If conventional Fourier transform is conducted on $x(k)$ and $y(k)$ individually, the amplitudes of their Fourier spectra are the same as Figure 2.1. In Figure 2.1, the planar rotation's directivity (i.e. clockwise or counter-clockwise) and amplitude cannot be revealed.

It is worth noting that the direction which the forward frequency components correspond to depend on the setting of the direct part and the quadrature part. In the above example, the forward frequency component corresponds to counter-clockwise direction. To make the clockwise direction associate with forward frequency components, $x(k)$ and $y(k)$ must be used as the quadrature buffer and the direct buffer respectively in the calculation of full spectrum using Figure 2.2.

In this section, Fourier transform, which is a signal processing technique to obtain the frequency spectrum, is introduced. In the next section, empirical mode decomposition, a

time-frequency domain analysis technique, is introduced.

2.2 Empirical Mode Decomposition

Empirical Mode Decomposition (EMD) decomposes a raw signal into a set of complete and almost orthogonal components called intrinsic mode functions (IMFs). IMFs represent the natural oscillatory modes embedded in the raw signal. Each IMF covers a certain frequency range. Materials and notations in this section follow [9].

There are two types of EMD: standard EMD and multivariate EMD. Standard EMD deals with a real-valued signal and is introduced in Section 2.2.1. Multivariate EMD deals with a multivariate signal and is introduced in Section 2.2.2.

2.2.1 Standard Empirical Mode Decomposition

For a real-valued signal, $x(k)$, standard EMD finds a set of IMFs, $c_i(k)$, and a residual signal, $r(k)$, so that

$$x(k) = \sum_{i=1}^{N} c_i(k) + r(k). \tag{2.2}$$

The IMFs are defined so as to have symmetric upper and lower envelopes with the number of zero crossings and the number of extrema differing at most by one. To extract IMFs, a sifting algorithm is employed, which is described in Table 2.1.

Table 2.1: Algorithm for standard EMD

Step 1. Find the locations of all the extrema of $x(k)$.
Step 2. Interpolate between all the minima to obtain the lower signal envelope, $e_{\min}(k)$. Interpolate between all the maxima to obtain the upper signal envelope, $e_{\max}(k)$.
Step 3. Compute the local mean, $$m(k) = [e_{\min}(k) + e_{\max}(k)]/2. \tag{2.3}$$
Step 4. Subtract the mean from $x(k)$ to obtain the "oscillatory mode", i.e. $s(k) = x(k) - m(k)$.
Step 5. If $s(k)$ satisfies the stopping criterion, then define $c_1(k) = s(k)$ as the first IMF; otherwise, set new $x(k) = s(k)$ and repeat the process from Step 1.

The same procedure in Table 2.1 is applied iteratively to the residue, $r(k) = x(k) - c_1(k)$, to extract other IMFs. The standard stopping criterion terminates the shifting process when the defined conditions for an IMF is met for a certain number of times [88].

Standard EMD considers only real-valued signals (i.e. signals measured from one unidirectional sensor only). When dealing with signals from multiple sensors or sensors from multiple directions, standard EMD needs to decompose signals from each sensor individually. This may result in unmatched decomposition results in terms of either the number

or the frequency content [89], because of the local and self-adaptive nature of the standard EMD. The above statement can be explained by a bivariate signal $x(k)$ which is defined as follows:

$$x(k) = x_1(k) + j \cdot x_2(k), \tag{2.4}$$

where

$$x_1(k) = 0.5\cos(40\pi k) + \cos(200\pi k), \tag{2.5}$$

$$x_2(k) = 1.2\cos(100\pi k) + 0.5\cos(40\pi k) + 0.9\cos(200\pi k). \tag{2.6}$$

Figure 2.4 shows the decomposition results by applying standard EMD to $x_1(k)$ and $x_2(k)$, separately. The first row shows the two raw signals. The last row shows the two residuals. The middle rows are the IMFs. There are two IMFs obtained for $x_1(k)$, and three IMFs obtained for $x_2(k)$. This is an unmatched problem in terms of the number of IMFs. The second IMF of $x_1(k)$, $c2_{x1}$, is at the frequency of 20 Hz, whereas the second IMF of $x_2(k)$, $c2_{x2}$, is at the frequency of 50 Hz. This is an unmatched problem in terms of the frequency content of the same IMF number. To overcome the unmatched problems, multivariate EMD [9] needs to be employed which is explained in Section 2.2.2.

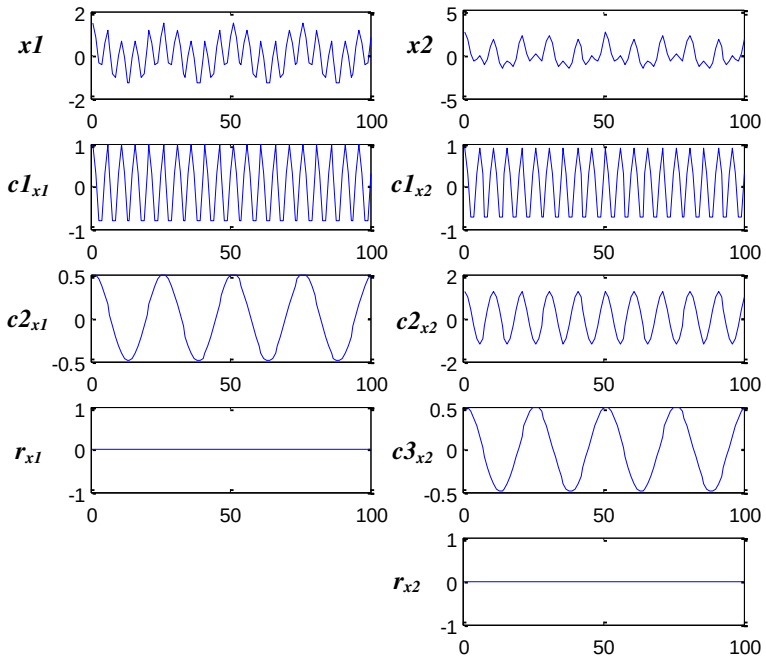


Figure 2.4: Decomposition results using standard EMD

2.2.2 Multivariate Empirical Mode Decomposition

Standard EMD considers only one-dimensional signals and the local mean is calculated by averaging the upper and lower envelopes. For multivariate signals, however, the local maxima and the local minima cannot be defined directly and the notion of "oscillatory modes" defining an IMF is rather confusing. To deal with these problems, multidimensional envelopes are firstly generated by taking a signal's projections along different directions, and then the average of these multidimensional envelopes are taken as the local mean [9,90, 91]. This calculation of local mean can be considered an approximation of the integral of all envelopes along the multiple projection directions in an m-dimensional space. The accuracy of this approximation depends on the uniformity of the chosen direction vectors. Thus, how to choose a suitable set of direction vectors for the projection becomes the main issue. For bivariate signals, points can be uniformly selected along a circle (also called 1-sphere) with the radius of 1, and each point represents a direction vector [90]. For trivariate signals, points need to be uniformly selected on a sphere (also called 2-sphere) with the radius of 1 [91]. For a general case (m-dimension), points that are uniformly distributed on a (m-1)-sphere with the radius of 1 needs to be selected. When m is large, the selection of such points is a problem. Rehman and Mandic [9] proposed to utilize low-discrepancy sequences for the generation of such points, and generalized the standard EMD to the multivariate EMD.

Before introducing the multivariate EMD, let us first see what are low-discrepancy sequences and how to generate them. Low-discrepancy sequences are also called quasi-random or sub-random sequences, due to their common use as a replacement of uniformly distributed random numbers. The following materials are from [9].

A convenient way of generating a low-discrepancy sequence involves the family of Halton and Hammersley sequences [9]. Let $b_1, b_2, ..., b_m$ be the first m prime numbers (i.e. 2, 3, 5, 7, 11,...), then the i^{th} sample of a one-dimensional Halton sequence, denoted by r_i^b, is given by

$$r_i^b = \frac{a_0}{b} + \frac{a_1}{b^2} + \frac{a_2}{b^3} + ... + \frac{a_s}{b^{s+1}}, \tag{2.7}$$

where a_i is an integer in [0, b-1], and is determined by

$$i = a_0 + a_1 b + a_2 b^2 + ... + a_s b^s. \tag{2.8}$$

As an example, a 2-dimensional Halton sequence is generated; one dimension of the Halton sequence is based on 2 (i.e. $b_1 = 2$) and the other on 3 (i.e. $b_2 = 3$). To generate the sequence for $b_1 = 2$, we start by let i=1. According to Equation (2.8), we have $1 = 1 + 0 \times 2 + 0 \times 2^2 + \ldots$, thus $a_0 = 1, a_1 = a_2 = \cdots = 0$. Substitute the values of a_i into Equation (2.7), we have $r_1^2 = \frac{1}{2}$.

Similarly, when i=2, according to Equation (2.8), $2 = 0 + 1 \times 2 + 0 \times 2^2 + \ldots$, thus $a_0 = 0, a_1 = 1, a_2 = a_3 = \cdots = 0$. Substitute the values of a_i into Equation (2.7), we have $r_2^2 = \frac{0}{2} + \frac{1}{2^2} = \frac{1}{4}$.

When i=3, according to Equation (2.8), we have $3 = 1 + 1 \times 2 + 0 \times 2^2 + \ldots$, thus $a_0 = 1, a_1 = 1, a_2 = a_3 = \cdots = 0$. Substitute the values of a_i into Equation (2.7), we have $r_2^2 = \frac{1}{2} + \frac{1}{2^2} = \frac{3}{4}$. Following the same way, the one-dimensional Halton sequence based on 2 is generated: $\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8}, \ldots$.

In the same way, the one-dimensional Halton sequence for $b_2 = 3$ is $\frac{1}{3}, \frac{2}{3}, \frac{1}{9}, \frac{4}{9}, \frac{7}{9}, \ldots$. When we pair them up, we get a 2-dimensional Halton sequence: $(\frac{1}{2}, \frac{1}{3})$, $(\frac{1}{4}, \frac{2}{3})$, $(\frac{3}{4}, \frac{1}{9})$, $(\frac{1}{8}, \frac{4}{9})$, $(\frac{5}{8}, \frac{7}{9})$, ...

The Hammersley sequence is used when the total number of sequence points, n, is known a priori; in this case, the i^{th} sample within the m-dimensional Hammersley sequence is calculated as $(\frac{i}{n}, r_i^{b_1}, r_i^{b_2}, \ldots, r_i^{b_{m-1}})$.

The Hammersley sequence can be used to generate direction vectors, or (equivalently) points on a $(m-1)$-dimensional sphere. Take a 3-dimensional Hammersley sequence as an example. It is used to find direction vectors on a 2-sphere. First, each dimension is scaled to be in the range [-1, 1], i.e. $(\frac{i}{n}, r_i^{b_1}, r_i^{b_2}) \longrightarrow (p_1, p_2, p_3) \in [-1, 1] \times [-1, 1] \times [-1, 1]$. Second, each sample is normalized to have the total length of 1, i.e. $p_i = \frac{p_i}{\sqrt{p_1^2 + p_2^2 + p_3^2}}$, (i=1, 2, 3). Figure 2.5 shows the points generated by Hammersley sequence on the 2-sphere. It can be seen that the points are uniformly distributed.

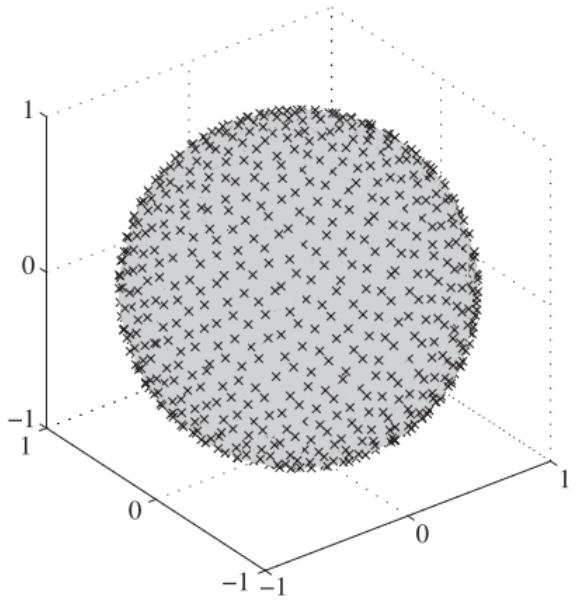


Figure 2.5: A Hammersley sequence on a 2-sphere [9]

The Halton and Hammersley sequence-based points correspond to a set of direction vectors, along which projections of an input multidimensional signal are calculated. The extrema of such projected signals are interpolated component-wise to yield multidimensional envelopes. The multidimensional envelopes, each of which corresponds to a particular direction vector, are then averaged to obtain the mean of the multi-dimensional signal.

Let $X(k) = \left(x_1(k), x_2(k), \ldots, x_m(k)\right)$ be an m-dimensional signal and $P^i = \left(p_1^i, p_2^i, \ldots, p_m^i\right)$ denote the i^{th} direction vector in a direction set, P. The procedure for multivariate EMD is outlined as follows [9].

Table 2.2: Algorithm for multivariate EMD

Step 1. Choose a suitable set of direction vectors, P.
Step 2. Calculate the i^{th} projection, $o^k(t)$, of the input signal $X(k)$ along the i^{th} direction vector, P^i, for each i (i.e. i=1, 2,..., l where l is the total number of direction vectors in P).
Step 3. Find the time instants, k_j^i, corresponding to the maxima of the projected signal, $o^i(k)$, for each i.
Step 4. Interpolate $[t_j^i, X(t_j^i)]$ to obtain multivariate envelopes, $E^i(k)$, for each i.
Step 5. The mean is estimated by

$$M(k) = \frac{1}{l}\sum_{i=1}^{l} E^i(k). \tag{2.9}$$

Step 6. Calculate $D(k) = X(k) - M(k)$. If $D(k)$ fulfills the stopping criterion for a multivariate IMF, then assign $D(k)$ as an IMF and apply the above procedure from Step 2 to $X(k) - D(k)$ to extract the next IMF; otherwise, apply it to $D(k)$.

The stopping criterion for multivariate IMFs is similar to that for the standard EMD. The difference is that the condition for the equality of the number of extrema and the number of zero crossings is not imposed.

Multivariate EMD is applied to $X(k)$ defined in Equation (2.4). The decomposition result is shown in Figure 2.6. The first row is the raw signal. The last row is the residual. Three middle rows represent three IMFs. It can be seen that the numbers of IMFs for the two components of $X(k)$ are the same. The unmatched problem in terms of the number of IMF doesn't exist any more. Furthermore, the signals of the two columns at each row share the same frequency content. For example, the second IMFs, $c2_{x1}$ and $c2_{x2}$, both have the frequency of 50 Hz; because there is no component in $x_1(k)$ having the frequency of 50 Hz, so the amplitude of $c2_{x1}$ is really small. This means that the unmatched problem in terms of the frequency content doesn't exist any more using multivariate EMD.

In this section, two signal processing techniques are introduced. Before the techniques on feature selection and machine learning are introduced, different measurement scales are reviewed in Section 2.3 because the selections of feature selection algorithm and machine learning algorithm need the information on the measurement scales of features and fault conditions.

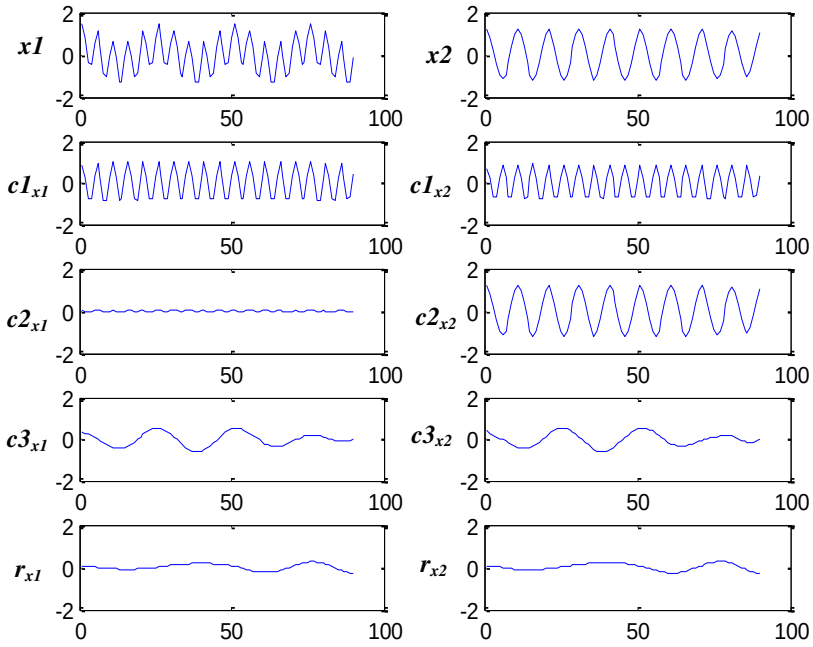


Figure 2.6: Decomposition results using multivariate EMD

2.3 Measurement Scales and Types of Variables

Steven [92] divided the scales of measurement into four types: nominal, ordinal, interval, and ratio. The nominal scale is used to describe variables that have two or more categories but do not have an intrinsic order, e.g. fruits (D={apple, pear, orange}). The categories in this variable are parallel to each other, that is they cannot be compared using the word "better" or "worse". When there are only two categories in this variable, the variable is called binary (or dichotomous), e.g. gender (D={male,female}). If a variable is measured in nominal scale, it is called a **nominal variable**.

The ordinal scale is rank-ordered but does not necessarily have metric information, e.g. grades of students (D={A+, A, A-, . . . , F}). The variable measured by an ordinal scale has discrete values (called **"ranks"** in this thesis). These ranks can be compared qualitatively using the word "better" or "worse". That is, there is **ordinal information** (also called monotonic relation, preference relation [93]) among the ranks. However a quantitative comparison between different ranks is impossible. For example, we can say "grade A+ is better than A-"; but it is improper to say "Grade A+ is 10 times better than A-". If a variable is measured in an ordinal scale, it is called an **ordinal variable**.

The interval scale and the ratio scale are for variables which have metric information. The difference between the two scales lies in the "zero point". The "zero point" on an ratio scale is not arbitrary, and negative values can not be used on this scale. For example, the Kelvin temperature scale has an absolute zero, which is denoted 0 K. This zero point is not arbitrary because the particles that compose matter at this temperature have zero kinetic

energy. On the internal scale, there is no true "zero point" and negative values can be used. Ratios between numbers on this scale are not meaningful. A typical example is Celsius temperature scale. We cannot say the 40 °C water is twice hotter than the 20 °C water. Under the interval or ratio scales, the quantitative comparison is achieved. If a variable is measured on the interval/or ratio scale, it is called a **continuous variable (or numerical variable)**. In this thesis, it is called, a continuous variable.

Let us see the measurement scales for features and labels (fault conditions) in fault diagnosis. Features extracted from vibration signals are usually continuous variables (e.g. kurtosis) [73, 94]. Some features, such as the status of a valve (on/off)), are nominal variables [95, 96].

Fault conditions in fault detection and isolation describes different fault types (e.g. pitting, crack). They are nominal variables and can be regarded as different classes. So fault detection and isolation can be achieved through classification algorithms.

Fault conditions in fault identification describes different fault levels. They are ordinal variables and can be regarded as different ranks. Fault identification can be achieved through ordinal ranking algorithms.

The above information is helpful in the description of techniques on feature selection and machine learning. Next the commonly used tools in feature selection, correlation coefficient, is introduced.

2.4 Correlation Coefficients

Correlation coefficients evaluate the correlation (or dependence) between two variables. Depending upon the types of variables, different correlation coefficients are defined, some of which are listed in Table 2.3 [97–99]. The Phi correlation coefficient, the Rank-biserial correlation coefficient and the Point-biserial correlation coefficient are defined for correlations between two binary variables, between one binary variable and one ordinal variable, and between one binary variable and one continuous variable, respectively. The Polychoric, Spearman rank and Kendal rank correlation coefficients are defined for correlations between two ordinal variables. The Polyserial correlation coefficient is used for the correlation between one continuous variable and one ordinal variable. The Pearson correlation coefficient, the most popular one, deals with two continuous variables, and is introduced in Section 2.4.1.

2.4.1 Pearson Correlation Coefficient

The Pearson correlation coefficient between two continuous variables is defined as the covariance of the two variables divided by the product of their standard deviations. It evaluates the linear correlation between two variables. The absolute value of Pearson correlation coefficient ($|\rho_{xy}|$) is the most commonly used in feature selection, which is defined in Equation

Table 2.3: List of correlation coefficients

Types of variables	Nominal (binary)	Ordinal	Continuous
Nominal (binary)	Phi	Rank-biserial	Point-biserial
Ordinal	Rank-biserial	Polychoric, Spearman rank, Kendal rank	Polyserial
Continuous	Point-biserial	Polyserial	Pearson

(2.10) where x and y are two continuous variables, $\overline{x}$ and $\overline{y}$ are their mean values.

$$\rho_{xy} = \left| \frac{\text{cov}(x, y)}{\sigma_x \cdot \sigma_y} \right| = \left| \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \right| \tag{2.10}$$

$|\rho_{xy}|$ takes values between 0 and 1. The value of 1 means the two variables are perfectly correlated; the value of 0 means the two variables are not correlated at all. Figure 2.7 shows two cases with $|\rho_{xy} = 1|$ (left) and $|\rho_{xy} = 0|$ (right), respectively. It can be seen that y randomly changes with x when $|\rho_{xy} = 0|$. When $|\rho_{xy} = 1|$, y changes linearly with x. Note that in this case, y can increase or decrease linearly with x, even though in Figure 2.7, y increases linearly with x. The above materials are from [97].

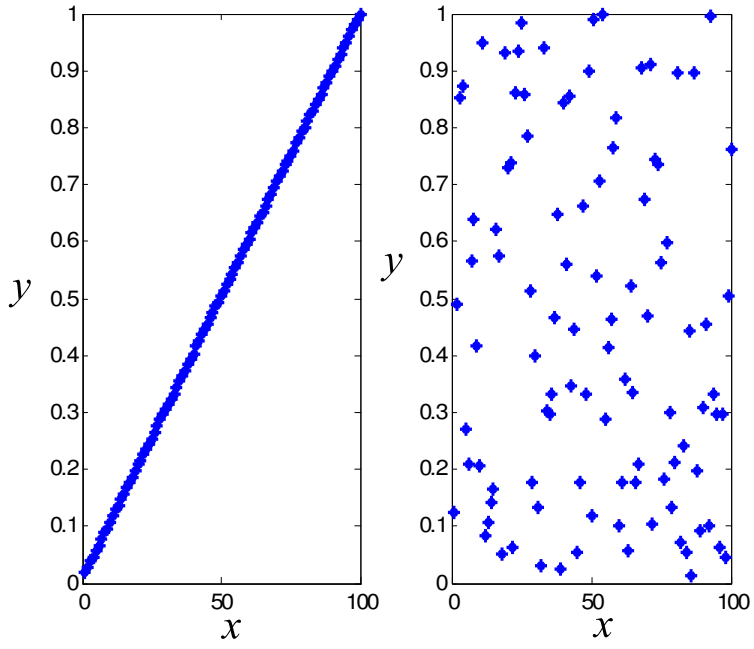


Figure 2.7: Absolute value of Pearson correlation coefficient ($|\rho_{xy}|$): 1 (left) and 0 (right)

Equation (2.10) can be used to evaluate the relevance not only between two continuous variables, but also between a continuous variable and a nominal variable, or between two nominal variables. The reason is explained as follows.

The Point-biserial and the Phi correlation coefficients mathematically equal to the Pearson correlation coefficient. According to Table 2.3, Equation (2.10) is also applicable to a binary and a continuous variable or two binary variables. Furthermore, by using a complete disjunctive coding [94], Pearson correlation coefficient can evaluate the relevance between a nominal variable and a continuous variable, and the correlation between two nominal variables. Disjunctive coding is explained as follows by a nominal variable that has 4 values. Using the disjunctive coding, this nominal variable is represented by 4 binary variables C_1([1 0 0 0]), C_2([0 1 0 0]), C_3([0 0 1 0]) and C_4 ([0 0 0 1]), respectively. In order to evaluate the correlation between this nominal variable and a continuous variable x, the following steps are conducted. First, the Point-biserial correlation (mathematically equals to the Pearson correlation) coefficient between x and each of the four variables C_i ($i = 1, 2, 3, 4$) are calculated. Then the average of the Point-biserial correlation coefficient is taken as the correlation coefficient between the nominal variable and the continuous variable. Similar procedure can be applied to two nominal variables.

In the maximum relevance and minimum redundancy scheme of feature selection, the feature-label relevance and feature-feature redundancy need to be evaluated. In fault diagnosis, features are often continuous and/or nominal as stated in Section 2.3. Thus Equation (2.10) can be used for feature-label relevance and feature-feature redundancy [94,100]. The Pearson correlation coefficient will be employed for feature selection in Chapters 6 and 7 in this thesis.

2.4.2 Polyserial Correlation Coefficient

Polyserial correlation coefficient evaluates the correlation between a continuous variable and an ordinal variable. Materials in this section are from [101]. Let x and y be two continuous variables. In some cases, the continuous variable (e.g. y) could only be measured in rough categories like low, medium, high, etc, using an ordinal variable z (as shown in Equation (2.11)). Here $z_1, z_2, \ldots, z_r$ are known increasing ranks (i.e. $z_i < z_{i+1}$) and $\mathbf{b} = (b_0, b_1, \ldots, b_r)$ is a vector of known thresholds with $b_0 = -\infty$ and $b_r = +\infty$.

$$z = z_i, \quad if \quad b_{i-1} < y < b_i, \quad i = 1, 2, \ldots, r \tag{2.11}$$

The Polyserial correlation coefficient between observed data (x and z) is actually an estimation of the Pearson correlation between x and y. Without the loss of generality, it is assumed that $x \sim N(\mu_x, \sigma_x)$, $y \sim N(\mu_y, \sigma_y)$, and the joint distribution of x and y follows the bivariate normal distribution.

$$p(x, y) = \frac{1}{2\pi\sigma_x\sqrt{1-\widetilde{\rho_{xy}}}} \exp\left(-\frac{\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\widetilde{\rho_{xy}}xy - \mu_{xy}}{\sigma_x} + y^2}{2(1-\widetilde{\rho_{xy}}^2)}\right), \tag{2.12}$$

where $\widetilde{\rho_{xy}}$ is called the Polyserial correlation coefficient between x and z. There are mainly two ways to estimate $\widetilde{\rho_{xy}}$: the maximum-likelihood estimator and the two-step estimator

[101]. The latter is computationally simpler than the former, and is used in this thesis. In the two-step estimator [101], μ_x and σ_x in (in Equation (2.12)) are estimated by the sample mean and the sample covariance of x, and the thresholds $\mathbf{b}$ (Equation (2.11)) are estimated by the normal inverse cumulative distribution function evaluated at the cumulative marginal proportions of z. For example, if the total number of samples is 200 and $z = 1$ is observed for 40 times, then $b_1 = \Phi^{-1}(40/200) = -0.84$ where $\Phi^{-1}(p)$ is the inverse cumulative distribution function and p is the probability. After that, $\widetilde{\rho_{xy}}$ can be estimated by maximizing the likelihood of n observations of (x, z) with respect to $\widetilde{\rho_{xy}}$ only, i.e.

$$L = \prod_{i=1}^{n} p(x_i, z_i) = \prod_{i=1}^{n} p(x_i) p(z_i|x_i), \tag{2.13}$$

where $p(x_i)$ is the probability density function of x, and $p(z_i|x_i)$ is the conditional probability density function which is a function of $\widetilde{\rho_{xy}}$. For further details, refer to [101]. In this thesis, the *R* software is used for the calculation of $\widetilde{\rho_{xy}}$.

$|\widetilde{\rho_{xy}}|$ takes values between 0 and 1. The value of 1 means that two variables are perfectly monotonically correlated; the value of 0 means that two variables are not correlated at all. Figure 2.8 shows two cases with $|\widetilde{\rho_{xy}}| = 1$ (left) and $|\widetilde{\rho_{xy}}| = 0$ (right), respectively. It can be seen that the rank z randomly changes with x when $|\widetilde{\rho_{xy}}| = 0$. When $|\widetilde{\rho_{xy}}| = 1$, z changes monotonically with x. Note that in this case, z can monotonically increase or decrease with x, even though in Figure 2.8, z increases with x.

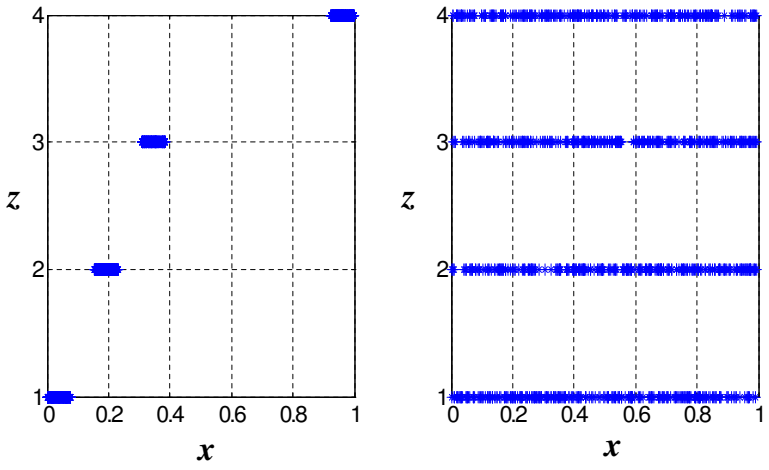


Figure 2.8: Absolute value of Polyserial correlation coefficient ($|\widetilde{\rho_{xy}}|$): 1 (left) and 0 (right)

It is worth mentioning that the Pearson correlation coefficient is not suitable for evaluating the relevance between a continuous variable and an ordinal variable. Figure 2.8 shows a clear monotonic trend between x and z based on Polyserial correlation coefficient. However, if the Pearson correlation coefficient is used, a value of 0.93 is generated instead of 1.

The Polyserial correlation coefficient will be employed for feature selection in Chapters 6 and 7 in this thesis.

2.5 Rough Sets

A set (U) is a collection of samples. In a crisp set, the degree of membership of any sample in the set is either 0 or 1. A rough set is a formal approximation of a crisp set in terms of a pair of sets which are the lower and the upper approximation of the crisp set. The concept of rough sets is first proposed by Pawlak [64] in 1991. The rough set model he proposed is called Pawlak rough set model, which is described as follows [64].

2.5.1 Pawlak Rough Set

An information system is a pair $IS = (U, A)$, where $U = \{u_1, ..., u_n\}$ is a nonempty finite set of (n number of) samples and $A = \{a_1, ..., a_m\}$ is a finite set of descriptors to describe the samples. In this rough set, the relation between a pair of samples are described by the equivalence relation (IND). That is if the descriptors' values for two samples are the same, then the two samples are in equivalence relation. Mathematically, for a subset $B \subseteq A$, there is an associated **equivalence relation**, IND(B):

$$\mathrm{IND}(B) = \Big\{(u_i, u_j) \in U \times U : f(u_i, a) = f(u_j, a), \forall a \in B\Big\}, \tag{2.14}$$

where $f(u_i, a)$ is the value of descriptor a for sample u_i. The relation IND(B) is called a B-indiscernibility relation. The portion of U is a family of all equivalence classes of IND(B). With the equivalence relation defined, the following set ($[u_i]_B$) is associated with u_i. All the samples in $[u_i]_B$ have the same value as u_i.

$$[u_i]_B = \Big\{u_j : f(u_i, a) = f(u_j, a), \forall a \in B\Big\}, \tag{2.15}$$

Let $X \subseteq U$ be a target set that is used to be represented using B. The target set X can be approximated using the information contained in B by constructing the $B-$lower and $B-$upper approximations of X as follows:

$$\underline{B}X = \bigcup \Big\{u_i : [u_i]_B \subseteq X\Big\}, \tag{2.16}$$

$$\overline{B}X = \bigcup \Big\{u_i : [u_i]_B \cap X \neq \emptyset\Big\}. \tag{2.17}$$

X is said to be definable if $\underline{B}X = \overline{B}X$; otherwise, X is a rough set. $B_N(X) = \overline{B}X - \underline{B}X$ is the boundary of X.

Example 1 *The concept of the Pawlak rough set is illustrated in Figure 2.9. The big rectangle represents the universe (U). The samples in U are granulated into a number of mutually exclusive equivalence information granules shown as the lattices. To describe a subset $X \in U$ (indicated by the circle) with these granules, two subsets of granules: the*

lower approximation (marked in hatch area) and the upper approximation (marked in grey and hatch area) are to be found. The lower approximation is a minimal subset of granules which are included in X. The upper approximation is a maximal subset of granules which includes X.

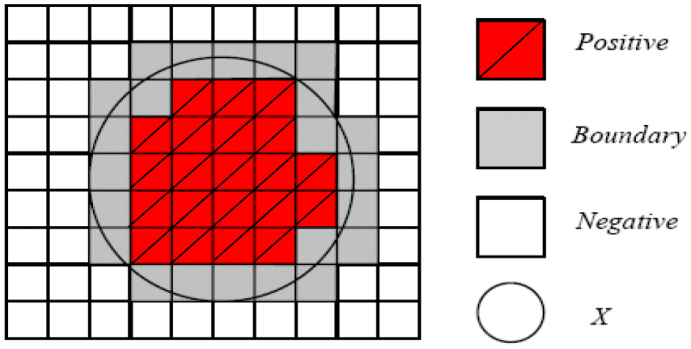


Figure 2.9: Pawlak rough set [10]

If the descriptors A can be expressed in the form of $A = C \bigcup D$, where C is a set of features describing the samples' characteristics, and D is the label which specifies the samples' classes or ranks, the information system (U, A) is called a **decision table**, $DT = (U, C \bigcup D)$. For the convenience of description, C is called features and D is called labels in this thesis.

Let D_l denotes the set of samples having the same label value l. The lower approximation and the upper approximation of D_l using B are defined as

$$\underline{B}D_l = \bigcup \left\{ u_i : [u_i]_B \subseteq D_l \right\}, \tag{2.18}$$

$$\overline{B}D_l = \bigcup \left\{ u_i : [u_i]_B \cap D_l \neq \emptyset \right\}, \tag{2.19}$$

respectively. Assume there are L values in label set D. The lower approximation and the upper approximation of D using B are defined as

$$\underline{B}D = \bigcup_{l=1}^{L} \underline{B}D_l, \tag{2.20}$$

$$\overline{B}D = \bigcup_{l=1}^{L} \overline{B}D_l. \tag{2.21}$$

The lower approximation ($\underline{B}D$) is a modest estimation of D using B. Samples in the set $\underline{B}D$ have the same feature values and the same label. The approximation quality (also called dependency) is defined as

$$\gamma_B(D) = \frac{|\underline{B}D|}{|U|} \tag{2.22}$$

where $|\cdot|$ is the cardinality of a set (i.e. the number of elements in a set). $\gamma_B(D)$ can be interpreted as the portion of samples in a decision table for which it suffices to know the

features in B to determine the label D. In another word, $\gamma_B(D)$ reflects the ability of B to approximate D. The value of approximation quality ranges from 0 to 1. The feature subset whose approximation ability is high plays an important role in determining the label.

Pawlak rough set model utilizes the equivalence relation (Equation (2.14)), and thus is suited for nominal features only. In fault diagnosis, continuous features are often extracted via vibration analysis. In order to handle continuous features, Hu et al. [10] introduced neighborhood relation and extended the Pawlak rough set to a neighborhood rough set, which is described in Section 2.5.2.

2.5.2 Neighborhood Rough Set

In order to handle continuous features, neighborhood rough set replaces the equivalence relation in Pawlak rough set with neighborhood relation. Before neighborhood relation is introduced, let us see what is neighborhood.

Given $u_i \in U$, and u_i is described by a feature subset B, the **neighborhood** of u_i in terms of B is a set denoted by $\delta_B(u_i)$, which is mathematically defined by

$$\delta_B(u_i) = \Big\{u_j : u_j \in U, \Delta_B(u_i, u_j) \leq \delta\Big\}, \tag{2.23}$$

where δ is the **neighborhood size** and $\Delta_B(u_i, u_j)$ is a distance function that evaluates the distance between two samples u_i and u_j in the feature space expanded by B. For nominal features, the distance function is defined in Equation (2.24). In such a case, the neighborhood rough set is the same as the Pawlak rough set.

$$N_B(u_i, u_j) = \begin{cases} 1, & f(u_i, a) \neq f(u_j, a) \quad \forall a \in B \\ 0, & \text{otherwise} \end{cases} \tag{2.24}$$

For continuous features, Minkowsky distance (Equation (2.25)) can be used as the distance function:

$$\Delta_B(u_i, u_j) = \left[\sum_{k=1}^{m} |f(u_i, a_k) - f(u_j, a_k))^p|^{\frac{1}{p}}\right], \tag{2.25}$$

where m is the total number of features in B, $f(u_i, a_k)$ and $f(u_j, a_k)$ are the values of feature a_k for samples u_i and u_j respectively, and p is a real number. (1) If $p = 1$, it is called the Manhattan distance; (2) if $p = 2$, it is called the Euclidean distance; and (3) if $p = \infty$, it is the Chebychev distance.

A **neighborhood relation** (N) on the universe can be written as a relation matrix, $N = (r_{ij})_{n\times n}$, where r_{ij} is the neighborhood relation between two samples u_i and u_j.

$$r_{ij} = \begin{cases} 1, & u_j \in \delta_B(u_i) \\ 0, & \text{otherwise} \end{cases} \tag{2.26}$$

With the neighborhood relation N defined, the decision table $(U, C \bigcup D)$ is called the neighborhood decision table denoted by $NIS = (U, C \bigcup D, N)$. Suppose $D_1, D_2, ..., D_L$ are

the sets that have labels 1, 2, ..., L, respectively. The neighborhood of u_i, $\delta_B(u_i)$, is a set that contains samples within the neighborhood of sample u_i generated by the feature subset $B \subseteq C$. The lower and upper approximations of decision D with respect to B are defined as

$$\underline{N_B D} = \bigcup_{l=1}^{L} \underline{N_B} D_l, \tag{2.27}$$

$$\overline{N_B D} = \bigcup_{l=1}^{L} \overline{N_B} D_l, \tag{2.28}$$

where,

$$\underline{N_B} D_l = \left\{ u_i | \delta_B(u_i) \subseteq D_l, u_i \in U \right\}, \tag{2.29}$$

$$\overline{N_B} D_l = \left\{ u_i | \delta_B(u_i) \cap D_l \neq \emptyset, u_i \in U \right\}. \tag{2.30}$$

With the lower approximation defined, the approximation ability of B in determining D is defined by Equation (2.31).

$$\gamma_B(D) = \frac{|\underline{N_B} D|}{|U|} \tag{2.31}$$

Example 2 *Now, an example is used to illustrate the concept of the neighborhood rough set. Table 2.4 shows nine samples described by two features a_1 and a_2. The label for each sample is given in D. A 2-dimensional representation of the samples are given in Figure 2.10. The three values (1,2,3) in D represents three categories of a nominal variable, i.e. three classes.*

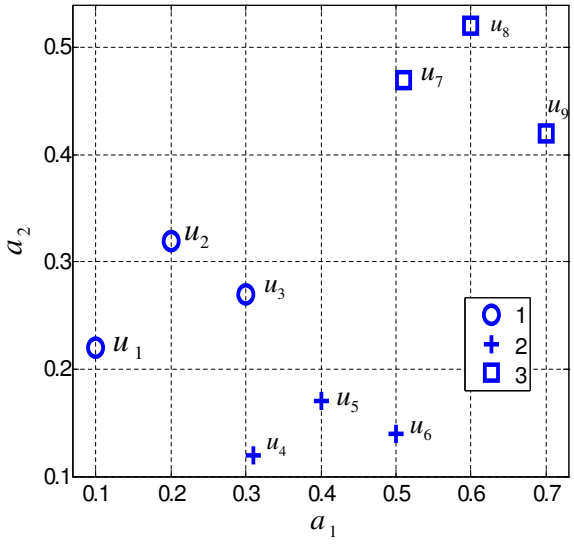


Figure 2.10: Plot of samples

Table 2.4: Values of features (a_1 and a_2) and labels (D) for samples

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9
a_1	0.10	0.20	0.30	0.31	0.40	0.50	0.51	0.60	0.70
a_2	0.20	0.30	0.25	0.10	0.15	0.12	0.45	0.50	0.40
D	1	1	1	2	2	2	3	3	3

The approximation quality defined in neighborhood rough set is calculated to evaluate the significance of the two features in classifying D. In the following, a_1 *is taken as an example. A neighborhood size (i.e.* δ *in Equation (2.23)) needs to be chosen first. Here* $\delta = 0.05$ *is used. The samples whose distance to sample* u_i *is within 0.05 are associated with* $\delta_{a_1}(u_i)$*. Specifically,*

$$\begin{aligned}
\delta_{a_1}(u_1) &= \{u_1\},\\
\delta_{a_1}(u_2) &= \{u_2\},\\
\delta_{a_1}(u_3) &= \{u_3, u_4\},\\
\delta_{a_1}(u_4) &= \{u_3, u_4\},\\
\delta_{a_1}(u_5) &= \{u_5\},\\
\delta_{a_1}(u_6) &= \{u_6, u_7\},\\
\delta_{a_1}(u_7) &= \{u_6, u_7\},\\
\delta_{a_1}(u_8) &= \{u_8\},\\
\delta_{a_1}(u_9) &= \{u_9\}.
\end{aligned}$$

Let D_l *denote the set of samples having label l, thus,*

$$\begin{aligned}
D_1 &= \{u_1, u_2, u_3\},\\
D_2 &= \{u_4, u_5, u_6\},\\
D_3 &= \{u_7, u_8, u_9\}.
\end{aligned}$$

Then, each $\delta_{a_1}(u_i)$ $(i = 1, 2, ..., 9)$ *are compared with* D_l*. According to Equation (2.29), if* $\delta_{a_1}(u_i) \subseteq D_l$*, then* u_i *is put into set* $\underline{N_{a_1}} D_l$*. The following sets are associated.*

$$\begin{aligned}
\underline{N_{a_1}} D_1 &= \left\{u_1, u_2\right\}\\
\underline{N_{a_1}} D_2 &= \left\{u_5\right\}\\
\underline{N_{a_1}} D_3 &= \left\{u_8, u_9\right\}
\end{aligned}$$

Therefore,

$$\underline{N_{a_1}} D = \bigcup_{i=1}^{3} \underline{N_{a_1}} D_l = \left\{u_1, u_2, u_5, u_8, u_9\right\}.$$

The approximation quality of a_1 is

$$\gamma_{a_1}(D) = \frac{|\underline{N_{a_1} D}|}{|U|} = \frac{5}{9} = 0.56.$$

The same procedure is applied for a_2, and an approximation quality of $\gamma_{a_2}(D) = 1.00$ is obtained. Because $\gamma_{a_2}(D) > \gamma_{a_1}(D)$, a_2 is more important than a_1 in classifying D. It is consistent with our intuition. As observed from Figure 2.10, samples are more scattered in the vertical projection (i.e. feature a_2) than in the horizontal projection (i.e. feature a_1).

The neighborhood size (δ) largely affects the evaluation results. In the above example, if $\delta = 0.0005$, then $\gamma_{a_2}(D) = \gamma_{a_1}(D) = 1$ meaning that a_1 and a_2 are equally important. The choice of the neighborhood size will be discussed in Chapter 4.

Neighborhood rough set deals with classification problems whose labels are nominal variables. If the labels are ordinal variables, dominance rough set and fuzzy preference based rough set need to be employed in order to extract the preference structure on the labels. This is presented in the next two sections.

2.5.3 Dominance Rough Set

In ordinal ranking, the elements in the label (D) has a preference structure. The elements in D are called ranks. Without loss of generality, $d_1 \leq d_2 \leq, \ldots, \leq d_L$ are assumed, where d_l is the l^{th} element in D and L is the total number of elements (ranks). Let u_i, u_j be two samples in U, $a \in C$ be a feature describing the samples, and $f(u_i, a)$, $f(u_j, a)$ be their values in terms of a. If a reflects the ordinal information in D clearly, then u_i's rank should not be worse than u_j's rank when $f(u_i, a) \geq f(u_j, a)$. This is different from classification. In classification, the elements in D are called classes. u_i and u_j are classified into the same class, if $f(u_i) - f(u_j) = 0$ (in Pawlak rough set) or $f(u_i) - f(u_j) \leq \delta$ (in neighborhood rough set).

In order to consider the preference structure, dominance rough set replaces the equivalence relation in Pawlak rough set with the **dominance relation**. The features describing the samples are either ordinal or continuous. For $\forall u_i, u_j \in U$, u_i **dominates** u_j with respect to feature a, if u_i is better than u_j in terms of the value of feature a, i.e. $f(u_i, a) \geq f(u_j, a)$. The upward dominance relation and downward dominance relation between u_i and u_j are expressed as the two equations below, respectively.

$$r_{ij}^{\geq} = \begin{cases} 1, & f(u_i, a) \geq f(u_j, a), \\ 0, & \text{otherwise.} \end{cases} \tag{2.32}$$

$$r_{ij}^{\leq} = \begin{cases} 1, & f(u_i, a) \leq f(u_j, a), \\ 0, & \text{otherwise.} \end{cases} \tag{2.33}$$

With the dominance relation defined, the following two sets are associated. The first set, $[u_i]_a^{\geq}$, is called a-dominating set. It consists of samples that are not worse than sample

u_i with respect to a. The second set, $[u_i]_a^{\leq}$ called a-dominated set, consists of samples that are not better than sample u_i with respect to a.

$$[u_i]_a^{\geq} = \left\{u_j : f(u_j, a) \geq f(u_i, a)\right\} \tag{2.34}$$

$$[u_i]_a^{\leq} = \left\{u_j : f(u_i, a) \leq f(u_j, a)\right\} \tag{2.35}$$

Let $D_l^{\geq} = \bigcup_{p \geq l} D_p$ be the set of samples whose ranks are not worse than rank d_l, and $D_l^{\leq} = \bigcup_{p \leq l} D_p$ be the set of samples whose ranks are not better than rank d_l. The lower approximations of $D_l^{\geq}$ and $D_l^{\leq}$ with respect to feature a are defined with the upward lower approximation (Equation (2.36)) and the downward lower approximation (Equation (2.37)), respectively.

$$\underline{a}_{D_l^{\geq}}^{\geq} = \left\{u_i : [u_i]_a^{\geq} \subseteq D_l^{\geq}\right\} \tag{2.36}$$

$$\underline{a}_{D_l^{\leq}}^{\leq} = \left\{u_i : [u_i]_a^{\leq} \subseteq D_l^{\leq}\right\} \tag{2.37}$$

Furthermore, the upward lower approximation of D using feature a is defined with Equation (2.38). The downward lower approximation of D using feature a is defined with Equation (2.39).

$$\underline{a}_{D^{\geq}}^{\geq} = \bigcup_{l=1}^{L} \underline{a}_{D_l^{\geq}}^{\geq} \tag{2.38}$$

$$\underline{a}_{D^{\leq}}^{\leq} = \bigcup_{l=1}^{L} \underline{a}_{D_l^{\leq}}^{\leq} \tag{2.39}$$

There are many ways to estimate the approximation quality of a. In this thesis, the one (Equation (2.40)) reported in [93] is adopted.

$$\gamma_a(D) = \frac{|\underline{a}_{D^{\geq}}^{\geq}| + |\underline{a}_{D^{\leq}}^{\leq}|}{2|U|} \tag{2.40}$$

Example 3 *Now the same data as in Example 2 are used to illustrate how the dominance rough set works. The data are listed in Table 2.4 and also plotted in Figure 2.10. The nine samples are interpreted as nine manuscripts submitted to a journal. Two Features describe the originality (a_1) and writing quality (a_2) of each manuscript. Decisions of the manuscripts are "*reject*" (rank 1), "*revise*" (rank 2), and "*accept*" (rank 3). The importance of a_1 and a_2 in the decision making of a manuscript using the dominance rough set are compared.*

Let $[u_i]_{a_j}^{\geq}$ (respectively $[u_i]_{a_j}^{\leq}$) be the sets of samples whose value of feature a_i are higher (respectively smaller) than or equal to the value for sample u_i. Take a_1 as an example. Using

Equation (2.34), the following sets are associated.

$$\begin{aligned}
[u_1]^{\geq}_{a_1} &= \{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9\} \\
[u_2]^{\geq}_{a_1} &= \{u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9\} \\
[u_3]^{\geq}_{a_1} &= \{u_3, u_4, u_5, u_6, u_7, u_8, u_9\} \\
[u_4]^{\geq}_{a_1} &= \{u_4, u_5, u_6, u_7, u_8, u_9\} \\
[u_5]^{\geq}_{a_1} &= \{u_5, u_6, u_7, u_8, u_9\} \\
[u_6]^{\geq}_{a_1} &= \{u_6, u_7, u_8, u_9\} \\
[u_7]^{\geq}_{a_1} &= \{u_7, u_8, u_9\} \\
[u_8]^{\geq}_{a_1} &= \{u_8, u_9\} \\
[u_9]^{\geq}_{a_1} &= \{u_9\}
\end{aligned}$$

Let $D^{\geq}_l$ be the set of samples whose ranks are not less than rank r_l. The following sets are associated.

$$D^{\geq}_1 = \{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9\} \tag{2.41}$$

$$D^{\geq}_2 = \{u_4, u_5, u_6, u_7, u_8, u_9\} \tag{2.42}$$

$$D^{\geq}_3 = \{u_7, u_8, u_9\} \tag{2.43}$$

Now Equation (2.36) is used to compare the set $[u_j]^{\geq}_{a_1}$ with the set $D^{\geq}_l$, and get the upward lower approximation of $D^{\geq}_l$. For instance, set $[u_1]^{\geq}_{a_1}$ is compared with set $\underline{a_1}^{\geq}_{D^{\geq}_1}$ (because the rank of sample u_1 is rank 1). The two sets meet the condition, $[x]^{\geq}_{a_1} \subseteq D^{\geq}_1$. Thus u_1 is put into the set $\underline{a_1}^{\geq}_{D^{\geq}_1}$. The comparison is done for each sample, and the set $\underline{a_1}^{\geq}_{D^{\geq}_1}$ is found.

$$\underline{a_1}^{\geq}_{D^{\geq}_l} = \left\{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9\right\}$$

Similarly, the downward lower approximation can be obtained using Equations (2.35), (2.37) and (2.39).

$$\underline{a_1}^{\leq}_{D^{\leq}_l} = \left\{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9\right\}$$

Using Equation (2.40), the approximation quality of a_1 in approximating D is thus,

$$\gamma_{a_1}(D) = \frac{|\underline{a_1}^{\geq}_{D^{\geq}}| + |\underline{a_1}^{\leq}_{D^{\leq}}|}{|U|} = \frac{9+9}{2\times 9} = 1.00. \tag{2.44}$$

When the above procedure is applied to a_2, the approximation quality of 0.61 is obtained. Because $\gamma_{a_1}(D) > \gamma_{a_2}(D)$, a_1 is more important in determining the rank D than a_2. It can be seen from Figure 2.10 that even though different ranks are more separately scattered using a_2, the values of a_2 don't show a monotonic trend with the increase of ranks. On the other hand, a_1 can keep a better monotonic trend with the increase of ranks. Thus a_1 is more significant in expressing the ordinal information in D.

In dominance rough set, the preference relation is expressed qualitatively. That is, sample u_i is preferred to u_j as long as the feature value of the former is larger than that of the latter (i.e. $f(u_i, a) \geq f(u_j, a)$). It doesn't matter how much larger u_i is than u_j. If u_i and u_j belong to different ranks but $f(u_i, a)$ and $f(u_j, a)$ are close to each other, the approximation quality is sensitive to the their values. For example in the above example, if the value of a_1 for u_7 is changed from 0.51 to 0.49, the approximation quality will change from 1.00 to 0.89. To overcome this, fuzzy preference based rough set is proposed in [93] considering preference relation quantitatively.

2.5.4 Fuzzy Preference Based Rough Set

In dominance rough set, the a-dominating set ($[u_i]_a^{\geq}$ defined by Equation (2.34)) and the a-dominated set ($[u_i]_a^{\leq}$ defined by Equation (2.35)) of sample u_i in terms of feature a are two crisp sets. A sample $u_j \in U$ either belongs to a crisp set (i.e. 1) or not (i.e. 0), as defined by the characteristic functions (i.e. dominance relations shown in Equations (2.32) and (2.33)).

Non-crisp sets are called fuzzy sets, for which also a characteristic function is defined. This function is called a membership function. The membership function associates each sample a grade of membership. In contrast to classical set theory, a membership function of a fuzzy set can have in the normalized closed interval [0, 1]. Therefore, the membership function maps samples into real numbers in [0, 1].

In fuzzy preference based rough set, the fuzzy concept is brought in. The dominance relations (Equation (2.32) and Equation (2.33)) in dominance rough set is replaced by the **fuzzy preference relation**. A fuzzy preference relation, R, is expressed as an $n \times n$ matrix $R = (r_{ij})_{n \times n}$, where r_{ij} is the preference degree of sample u_i over sample u_j. The value of r_{ij} is calculated by a membership function, which will be introduced in the next paragraph. r_{ij}=1 is used to represent that u_i is absolutely preferred to u_j, r_{ij}=0.5 to represent that there is no difference between u_i and u_j, $r_{ij} > 0.5$ to represent that u_i is more likely to be preferred to u_j, and $r_{ij} < 0.5$ to represent that u_i is less likely to be preferred to u_j. When $r_{ij} = 0$, u_i is certainly not preferred to u_j.

Let $f(u, a)$ be the value of feature a for sample u. The upward and downward fuzzy preference relations between samples u_i and u_j can be computed by

$$r_{ij}^{>} = \frac{1}{1 + e^{-s(f(u_i,a) - f(u_j,a))}}, \text{ and } r_{ij}^{<} = \frac{1}{1 + e^{-s(f(u_j,a) - f(u_i,a))}}, \tag{2.45}$$

respectively. In the above two equations, s is a positive parameter which can be used to adjust the shape of the membership function and it is determined by specific applications. $r_{ij}^{>}$ represents how much u_i is larger than u_j, and $r_{ij}^{<}$ represents how much u_i is smaller than u_j. Thus fuzzy preference relations reflect not only whether sample u_i is larger/smaller than u_j (qualitatively), but also how much u_i is larger/smaller than u_j (quantitatively). Let $f(u_i, a)$ and $f(u_j, a)$ be the values of feature a for samples u_i and u_j, respectively. Figure

2.11 shows r_{ij} for different values of s. Only when the distance between u_i and u_j reaches to a certain value, u_i is certainly preferred to u_j (r_{ij}=1) or certainly not preferred to u_j (r_{ij}=0).

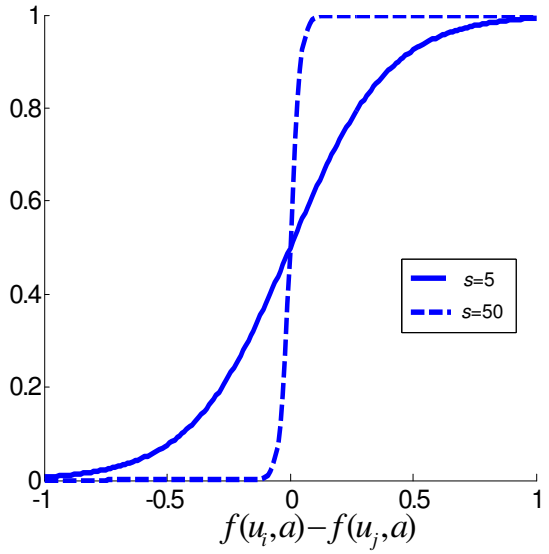


Figure 2.11: Fuzzy upward preference function with different s values

Let $R^{>}$ and $R^{<}$ be the upward fuzzy preference relation and the downward fuzzy preference relation induced by feature a, respectively. The memberships of a sample u_i to the lower approximations of $D_l^{\geq}$ and $D_l^{\leq}$ can be defined with: the upward fuzzy lower approximation (Equation (2.46)) and the downward fuzzy lower approximation (Equation (2.47)).

$$\underline{R}_{D_l^{\geq}}(u_i)^{>} = \inf_{u_j \in U} \max\left\{1 - R^{>}(u_j, u_i), D_l^{\geq}(u_j)\right\} \tag{2.46}$$

$$\underline{R}_{D_l^{\leq}}(u_i)^{<} = \inf_{u_j \in U} \max\left\{1 - R^{<}(u_j, u_i), D_l^{\leq}(u_j)\right\} \tag{2.47}$$

Hu et al [93] proved that the above equations are equivalent to the following two equations.

$$\underline{R}_{D_l^{\geq}}(u_i)^{>} = \inf_{u_j \notin D_l^{\geq}} 1 - R^{>}(u_j, u_i) \tag{2.48}$$

$$\underline{R}_{D_l^{\leq}}(u_i)^{>} = \inf_{u_j \in D_l^{\leq}} 1 - R^{<}(u_j, u_i) \tag{2.49}$$

Equation (2.48) indicates that the membership of u_i to the lower approximation of $D_l^{\geq}$ depends on the samples that do not belong to $D_l^{\geq}$ and produces the greatest preference over u_i. Equation (2.49) indicates that the membership of u_i to the lower approximation of $D_l^{\leq}$ depends on the samples that belong to $D_l^{\geq}$ and produces the greatest preference over u_i.

The fuzzy preference approximation qualities (FPAQ) of D with respect to a are then defined with:

$$\gamma_a(D) = \frac{\sum_l \sum_{u_i \in D_l} \underline{R}^{<}_{D_l^{\leq}}(u_i) + \sum_l \sum_{u_i \in D_l} \underline{R}^{>}_{D_l^{\geq}}(u_i)}{\sum_l |D_l^{\leq}| + \sum_l |D_l^{\geq}|} \tag{2.50}$$

where $|D_l^{\geq}|$ and $|D_l^{\leq}|$ are the numbers of samples with ranks dominating and dominated by D_l, respectively. FPAQ measures the significance of a feature in approximating ranks. The higher the value is, the more significant the feature is.

Example 4 *Now the same data as used in Example 3 are used to illustrate how fuzzy preference based rough set works. Data are listed in Table 2.4 and also plotted in Figure 2.10. The importance of two features a_1 and a_2 are compared in the decision making of the rank D using fuzzy preference based rough set.*

The upward and downward fuzzy preference relations induced by a_1 are presented in Table 2.5 and Table 2.6, respectively, where $s = 25$.

Table 2.5: Upward fuzzy preference relation ($R^{>}$) induced by a_1

u_i \ u_j	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9
u_1	0.50	0.92	0.99	0.99	1.00	1.00	1.00	1.00	1.00
u_2	0.08	0.50	0.92	0.94	0.99	1.00	1.00	1.00	1.00
u_3	0.01	0.08	0.50	0.56	0.92	0.99	1.00	0.99	1.00
u_4	0.01	0.06	0.44	0.50	0.90	0.99	1.00	0.99	1.00
u_5	0.00	0.01	0.08	0.10	0.50	0.92	0.99	0.94	1.00
u_6	0.00	0.00	0.01	0.01	0.08	0.50	0.92	0.56	0.99
u_8	0.00	0.00	0.01	0.01	0.06	0.44	0.90	0.50	0.99
u_7	0.00	0.00	0.00	0.00	0.01	0.08	0.50	0.10	0.92
u_9	0.00	0.00	0.00	0.00	0.00	0.01	0.08	0.01	0.50

Table 2.6: Downward fuzzy preference relation ($R^{<}$) induced by a_1

u_i \ u_j	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9
u_1	0.50	0.08	0.01	0.01	0.00	0.00	0.00	0.00	0.00
u_2	0.92	0.50	0.08	0.06	0.01	0.00	0.00	0.00	0.00
u_3	0.99	0.92	0.50	0.44	0.08	0.01	0.00	0.01	0.00
u_4	0.99	0.94	0.56	0.50	0.10	0.01	0.00	0.01	0.00
u_5	1.00	0.99	0.92	0.90	0.50	0.08	0.01	0.06	0.00
u_6	1.00	1.00	0.99	0.99	0.92	0.50	0.08	0.44	0.01
u_7	1.00	1.00	0.99	0.99	0.94	0.56	0.10	0.50	0.01
u_8	1.00	1.00	1.00	1.00	0.99	0.92	0.50	0.90	0.08
u_9	1.00	1.00	1.00	1.00	1.00	0.99	0.92	0.99	0.50

Sets $D_l^{\geq}$ and $D_l^{\leq}$ are associated as follows.

$$
\begin{aligned}
D_1^{\geq} &= \{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9\} && (2.51)\\
D_2^{\geq} &= \{u_4, u_5, u_6, u_7, u_8, u_9\} && (2.52)\\
D_3^{\geq} &= \{u_7, u_8, u_9\} && (2.53)\\
D_1^{\leq} &= \{u_1, u_2, u_3\} && (2.54)\\
D_2^{\leq} &= \{u_1, u_2, u_3, u_4, u_5, u_6\} && (2.55)\\
D_3^{\leq} &= \{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9\} && (2.56)
\end{aligned}
$$

The memberships of object u_i belonging to the lower approximations of $D_l^{\geq}$ and $D_l^{\leq}$ can be obtained by Equations (2.46) and (2.47). Tables 2.7 and 2.8 show the results. Consider u_1 as an example. The values of $\underline{R}^{>}_{D_1^{\geq}}(u_1)$, $\underline{R}^{>}_{D_2^{\geq}}(u_1)$, and $\underline{R}^{>}_{D_3^{\geq}}(u_1)$ are the memberships that u_1 belongs to sets $D_1^{\geq}$, $D_2^{\geq}$ and $D_3^{\geq}$, respectively. And the values of $\underline{R}^{<}_{D_1^{\leq}}(u_1)$, $\underline{R}^{<}_{D_2^{\leq}}(u_1)$, and $\underline{R}^{<}_{D_3^{\leq}}(u_1)$ are the memberships that u_1 belongs to sets $D_1^{\leq}$, $D_2^{\leq}$ and $D_3^{\leq}$, respectively.

$$
\begin{aligned}
\underline{R}^{>}_{D_1^{\geq}}(u_1) &= 1\\
\underline{R}^{>}_{D_2^{\geq}}(u_1) &= inf_{u \notin D_2} 1 - R^{>}(u, u_1)\\
&= inf\Big\{1 - R^{>}(u_1, u_1), 1 - R^{>}(u_{2,u}\, 1), 1 - R^{>}(u_3, u_1)\Big\}\\
&= inf\Big\{1 - 0.50, 1 - 0.92, 1 - 0.99\Big\} = inf\Big\{0.50, 0.08, 0.01\Big\} = 0.01\\
\underline{R}^{>}_{D_3^{\geq}}(u_1) &= inf_{u \notin d_3} 1 - R^{>}(u_{,u}\, 1)\\
&= inf\Big\{1 - R^{>}(u_1, u_1), 1 - R^{>}(u_2, u_1), 1 - R^{>}(u_3, u_1), 1 - R^{>}(u_4, u_1),\\
&\quad 1 - R^{>}(u_5, u_1), 1 - R^{>}(u_6, u_1)\Big\}\\
&= inf\Big\{1 - 0.5, 1 - 0.92, 1 - 0.99, 1 - 1.001 - 1.00, 1 - 1.00\Big\}\\
&= inf\Big\{0.08, 0.50, 0.92, 0, 0, 0\Big\} = 0
\end{aligned}
$$

$$
\begin{aligned}
\underline{R}^{<}_{D_1^{\leq}}(u_1) &= inf_{u \notin D_1} 1 - R^{<}(u, u_1)\\
&= inf\Big\{1 - R^{<}({}_u 4, u_1), 1 - R^{<}(u_5, u_1), 1 - R^{<}(u_6, u_1), 1 - R^{<}(u_7, u_1),\\
&\quad 1 - R^{<}(u_8, u_1), 1 - R^{<}(u_9, u_1)\Big\}\\
&= inf\Big\{1 - 0.01, 1 - 0.00, 1 - 0.00, 1 - 0.00, 1 - 0.00, 1 - 0.00\Big\}\\
&= inf\Big\{0.99, 1, 1, 1, 1, 1\Big\} = 0.99\\
\underline{R}^{<}_{D_2^{\leq}}(u_1) &= inf\Big\{1 - R^{<}(u_7, u_1), 1 - R^{<}(u_8, u_1), 1 - R^{<}(u_9, u_1)\Big\}\\
&= inf\Big\{1 - 0.00, 1 - 0.00, 1 - 0.00\Big\}\\
&= inf\Big\{1, 1, 1\Big\} = 1\\
\underline{R}^{<}_{D_3^{\leq}}(u_1) &= 1
\end{aligned}
$$

Table 2.7: Downward fuzzy lower approximation

u_i \ $\underline{R}^{<}_{D_l^{\leq}}$	$D_1^{\leq}$	$D_2^{\leq}$	$D_3^{\leq}$
u_1	0.99	1.00	1.00
u_2	0.94	1.00	1.00
u_3	0.56	0.99	1.00
u_4	0.00	0.99	1.00
u_5	0.00	0.94	1.00
u_6	0.00	0.56	1.00
u_7	0.00	0.00	1.00
u_8	0.00	0.00	1.00
u_9	0.00	0.00	1.00

Table 2.8: Upward fuzzy lower approximation

u_i \ $\underline{R}^{>}_{D_l^{\geq}}$	$D_1^{\geq}$	$D_2^{\geq}$	$D_3^{\geq}$
u_1	1.00	0.00	0.00
u_2	1.00	0.00	0.00
u_3	1.00	0.00	0.00
u_4	1.00	0.56	0.00
u_5	1.00	0.92	0.00
u_6	1.00	0.99	0.00
u_7	1.00	0.99	0.56
u_8	1.00	1.00	0.92
u_9	1.00	1.00	0.99

Using Table 2.7, Table 2.8 and Equations (2.51)-(2.56), the fuzzy preference approximation quality of D with respect to a_1 is obtained by Equation (2.50).

$$\gamma_{a_1}(D) = \frac{\sum_l \sum_{u_i \in D_l} \underline{R}^{<}_{D_l^{\leq}}(u_i) + \sum_l \sum_{u_i \in D_l} \underline{R}^{>}_{D_l^{\geq}}(u_i)}{\sum_l |D_l^{\leq}| + \sum_l |D_l^{\geq}|} = \frac{16.97 + 16.93}{18 + 18} = 0.94$$

Similarly, the $\gamma_{a_2}(D)$ is calculated and a value of 0.83 is obtained. Because $\gamma_{a_1}(D) > \gamma_{a_2}(D)$, a_1 is more important in determining the rank D than a_2. This is consistent with the result of dominance rough set (Example 3).

Compared with the dominance rough set, the fuzzy preference approximation quality is more robust to the change of feature values. For example if the value of a_1 for u_7 is slightly changed from 0.51 to 0.49, the approximation quality changes slightly from 0.94 to 0.93.

It is worth mentioning that the approximation qualities defined by the dominance rough set (Equation (2.40)) and fuzzy preference rough set (Equation (2.50)) can evaluate the monotonic relevance between a signal feature and the rank. However, they cannot be applied directly to a set of features, because the monotonic relation is defined between two

variables (i.e. the feature and the rank) only. When dealing with a set of features, the features need to be combined into one variable using some mathematical operations (e.g. picking up the minimum value of these features [93] or choosing the first principle component of these features [102]).

In the section, techniques that can be used for feature selection are introduced. In the next section, the machine learning algorithms are introduced.

2.6 Machine Learning Based on Support Vector Machine

As stated in Section 1.2.2.3, support vector machine (SVM) is a powerful tool for machine learning. In this section, two machine learning algorithms based on support vector machine are introduced: one for classification (Section 2.6.1), and the other for ordinal ranking (Section 2.6.2).

2.6.1 Support Vector Classification

A Support Vector Machine performs classification by constructing a hyperplane that optimally separates the data into two classes. In this section, the concept of support vector classification is introduced. First a technique for constructing an optimal separating hyperplane between two perfectly separated classes is discussed, and then it is generalized to the nonseparable case where the classes may not be separable by a linear boundary. Materials in this section are from [103].

Given a training data set

$$T = \left\{(\mathbf{X}, y_i) \ : \ \mathbf{x}_i \in R^m, y_i \in \{-1, 1\}\right\}_{i=1}^{n} \tag{2.57}$$

where $\mathbf{x}_i$ is an m-dimensional input vector, y_i is either -1 or 1, indicating two classes, and n is the total number of data points (samples). In the case of linearly data, it is possible to determine a hyperplane that separates the given data

$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0, \tag{2.58}$$

where $\mathbf{w}$ is the normal vector to the hyperplane and b is a scalar. The value $\frac{|b|}{\|\mathbf{w}\|}$ determines the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$. Figure 2.12 illustrates a linearly separable classification problem in a two-dimensional space. All squares are labeled "-1" while all circles are labeled "1". The solid line represents the separating plane. The two dash lines represent two parallel planes ($\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 1$ and $\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = -1$), and are called boundaries. The sold squares and the solid circles that are on the boundaries are called support vectors. The distance between the boundaries is called the margin.

The separating hyperplane that divides the samples having $y_i = 1$ from those having $y_i = -1$ and creates the maximum margin is called the optimal separating hyperplane. The

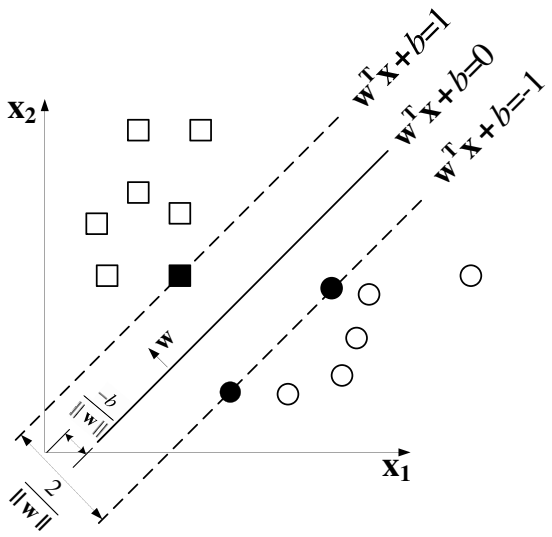


Figure 2.12: A 2-dimensional linearly separable classification problem (SVM Principal explanation)

margin, that is distance between these two boundaries, is calculated as

$$D = \left| \frac{(\mathbf{w}^{\mathrm{T}}\mathbf{x} + b - 1) - (\mathbf{w}^{\mathrm{T}}\mathbf{x} + b + 1)}{\|\mathbf{w}\|} \right| = \frac{2}{\|\mathbf{w}\|}. \tag{2.59}$$

To maximize the margin, $\|\mathbf{w}\|$ needs to be minimized. The norm of $\mathbf{w}$ involves a square root, which is difficult to solve in an optimization problem. So researchers alter the minimization of $\|\mathbf{w}\|$ with the minimization of $\mathbf{w}^{\mathrm{T}}\mathbf{w}$. Also, data points need to be prevented from falling into the margin, so the following constraints are added: $y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1$. Put this together, an optimization problem is obtained:

$$\begin{aligned} &\underset{\mathbf{w},b}{\text{minimize}} \quad \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \\ &\text{subject to} \quad y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \ldots, n. \end{aligned} \tag{2.60}$$

In the nonseparable case, there exist no hyperplanes that can split the "-1" and "1" samples. In this case, the so called **soft margin method** that allows for mislabeled samples needs to be applied [103]. Soft margin method chooses a hyperplane that splits the samples as clearly as possible, while still maximizing the margin. A slack variable, ξ_i, which measures the degree of misclassification of points failing on the wrong side of the margin is introduced.

$$y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i) + b \geq 1 - \xi_i, \; i = 1, 2, ..., n. \tag{2.61}$$

The optimal separating hyperplane is the one that maximizes the margin and minimizes the

classification error. It can be obtained by solving the following optimization problem:

$$\begin{aligned} \underset{\mathbf{w},b,\boldsymbol{\xi}}{\text{minimize}} \quad & \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \qquad\qquad i = 1, 2, \ldots, n. \end{aligned} \tag{2.62}$$

where ξ_i represents the distance between a data point lying on the wrong side of the margin and the boundary in its virtual class side, and C is a non-negative constant, called the error penalty. The separable case (described as the optimization problem (2.60)) is actually a special nonseparable case described by (2.62) with $C = \infty$.

The optimization problem (2.62) is quadratic with linear inequality constraints, hence it is a convex optimization problem. A quadratic programming solution is described by introducing Lagrangian multipliers, α_i and β_i. The Lagrangian primal function is

$$L = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1 + \xi_i) - \sum_{i=1}^{n}\beta_i\xi_i, \tag{2.63}$$

which is minimized with respect to $\mathbf{w}$, b, and ξ_i. Setting the respective derivatives to zero, we get

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Longrightarrow \mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i, \tag{2.64}$$

$$\frac{\partial L}{\partial b} = 0 \Longrightarrow \sum_{i=1}^{n}\alpha_i y_i = 0, \tag{2.65}$$

$$\frac{\partial L}{\partial \xi} = 0 \Longrightarrow \alpha_i + \beta_i = C,\ i = 1, 2, \ldots, n. \tag{2.66}$$

Substituting Equations (2.64) and (2.66) into Equation (2.63), the Lagrangian (Wolfe) dual objective function is obtained

$$\begin{aligned} \text{maximize inf} \quad & L(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j. \\ \text{subject to} \quad & \sum_{i=1}^{n}\alpha_i y_i = 0, \\ & C \geq \alpha_i \geq 0,\ i = 1, 2, \ldots, n. \end{aligned} \tag{2.67}$$

By solving the dual optimization problem (2.67), the coefficients α_i can be obtained. Then according to Equation (2.64), $\mathbf{w}$ is expressed as

$$\mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i = \sum_{j=1}^{p}\alpha_j y_j \mathbf{x}_j, \tag{2.68}$$

where p is the number of support vectors. Because a_i is only nonzero for support vectors, so the second half part of Equation (2.68) holds. According to Karush-Kuhn-Tucker (KKT)

conditions, the products of the dual variables and the constraints should be equal to zero for the support vectors, i.e.

$$\alpha_j(y_j(\mathbf{w}^{\mathrm{T}}\mathbf{x}_j + b) - 1) = 0, \;\; j = 1, 2, \ldots, p. \tag{2.69}$$

Any of these margin points can be used to solve for b, and typically an average of all the solutions for numerical stability can be used. Thus b can be expressed as:

$$b = \frac{1}{p}\sum_{j=1}^{p}(y_j - \mathbf{w}^{\mathrm{T}}\mathbf{x}_j). \tag{2.70}$$

Once $\mathbf{w}$ and b are available, the linear decision function can be given by:

$$G_f(\mathbf{x}) = \mathrm{sign}[\sum_{i=1}^{p}\alpha_i y_i(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}) + b]. \tag{2.71}$$

Given a new input data point ($\mathbf{x}$), if $G_f(\mathbf{x})$ is positive, then this data point is classified to class 1 ($y = 1$); if $G_f(\mathbf{x})$ is negative, then this data point belongs to class 2 ($y = -1$).

The support vector classifier described so far finds linear boundaries in the original input feature space. By introducing the kernel concept, the support vector classifier can be extended to nonlinear classifiers [103]. The idea is introducing a mapping function to project the original input data onto a high dimensional feature space in which the input data can be linearly separated. The linear boundaries in the high dimensional space translate to nonlinear boundaries in the original space. Once the mapping function is selected, the procedure of generating linear boundaries is the same as before. As a result, data points that can not be separated by a linear function in the input space can be linearly separated in the high dimensional feature space. The above argument is illustrated in Figure 1.7.

After adopting the mapping, the decision function, Equation (2.71), is modified to be:

$$G_f(\mathbf{x}) = \mathrm{sign}[\sum_{i=1}^{p}\alpha_i y_i < \phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}) > + b], \tag{2.72}$$

$$= \mathrm{sign}[\sum_{i=1}^{p}\alpha_i y_i \mathbf{K}(\mathbf{x}_i, \mathbf{x}) + b]. \tag{2.73}$$

where $\phi : \mathbb{R}^m \longrightarrow \chi$ is a mapping function which transforms the original input space into a high dimensional feature space. The kernel function is defined as the inner product of $\phi(\cdot)$, that is $\mathbf{K}(\mathbf{x_i}, \mathbf{x}) =< \phi(\mathbf{x_i}), \phi(\mathbf{x}) >$. It is worth mentioning that the mapping function does not explicitly affect the decision function, it is the kernel function that matters. Any function that satisfies Mercer's theorem [104] can be used as a kernel function. There are different kernel functions used in SVMs, such as linear, polynomial and Gaussian RBF. They are defined as follows.

$$\text{Linear kernel :} \qquad K(\mathbf{x}_i, \mathbf{x}) = \mathbf{x}_i^{\mathrm{T}}\mathbf{x}, \tag{2.74}$$

$$\text{Gaussian kernel :} \qquad K(\mathbf{x}_i, \mathbf{x}) = \exp(\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{2\sigma^2}), \tag{2.75}$$

$$\text{Polynomial kernel :} \qquad K(\mathbf{x}_i, \mathbf{x}) = (\mathbf{x}^{\mathrm{T}}\mathbf{x}_i)^d, \tag{2.76}$$

where σ represents the width parameter of the Gaussian kernel and d is the polynomial degree. Gaussian kernel and polynomial kernel are often adequate for most applications [105, 106]. In this thesis, the second degree polynomial kernel function ($d = 2$) is used.

The above content are for binary classification. Multi-class classification can be achieved by building binary classifiers which distinguish (i) between one of the classes to the rest (**one-versus-all**) or (ii) between every pair of classes (**one-versus-one**). Classification of new samples for one-versus-all case is done by a winner-takes-all strategy, in which the classifier with the highest output function assigns the class. For the one-versus-one approach, classification is done by a max-wins voting strategy, in which every classifier assigns the sample to one of the two classes, then the vote for the assigned class is increased by one vote, and finally the class with most votes determines the sample classification.

2.6.2 Support Vector Ordinal Ranking

As stated in section 1.2.2.3, ordinal ranking is a special kind of machine learning problem whose label is an ordinal variable. Ordinal ranking is similar to classification in the sense that the rank is a finite set. Nevertheless, besides representing the nominal variables as classification labels, ranks of ordinal ranking also carry ordinal information. That is, two ranks can be compared by the "<" (better) or ">" (worse) operation. Ordinal ranking is also similar to regression, in the sense that ordinal information is similarly contained in the label. However, unlike the real-valued regression labels; the discrete ranks do not carry metric information. That is, it is reasonable to say "rank A > rank B", but it is hard to say quantitatively how much larger rank A is.

Ordinal ranking has not been studied as much as in classification. In this section, a review on ordinal ranking is introduced first. Then a ordinal ranking algorithm is given.

2.6.2.1 Review on Ordinal Ranking

Ordinal ranking was studied from the statistic perspective two decades ago [107]. However, there is not much study on this problem from the perspective of machine learning until recently. A commonly used idea to conduct ordinal ranking is to transform the ranking problem to a set of binary classification problems, or to add additional constraints to traditional classification formulations. Herbrich et al. [108] proposed a loss function between pairs of ranks which gave a distribution independent bound, and then applied the principle of maximum margin to solve the ordinal ranking problem. The idea of pairwise comparison was also adopted in ordinal ranking algorithms proposed by Freund et al. [109] and Lin and Li [110]. However, because there are $O(N^2)$ pairwise comparisons out of N training samples, the computation complexity will be high when N is large. Crammer and Singer [111] generalized an algorithm using multiple thresholds to predict r ranks. The feature space was divided into r parallel equally-ranked regions, where each region stood for

one rank. With this approach, the loss function was calculated pointwisely and the quadratic expansion problem could be avoided. Following the same idea, Shashua and Levin [112] generalized support vector machine (SVM) into ordinal ranking by finding $r-1$ thresholds that divided the real line into r consecutive intervals for the r ranks. Chu and Keerthi [11] improved the approach in [112] and proposed two new approaches by imposing the ordinal inequality constraints on the thresholds explicitly in the first approach and implicitly in the second one. Li and Lin [113] presented a reduction framework from ordinal ranking to binary classification. Cardoso and Costa [114] transformed the ordinal ranking to a single binary classifier and also implemented it using SVM and neural networks, respectively. Sun et al. [115] expanded the kernel discriminant analysis by a rank constraint to solve ordinal ranking problems.

Among the above methods, SVM-based methods have shown great promise in ordinal ranking. The algorithm which implicitly adds constraints to SVM proposed in [11] (called support vector ordinal regression (SVOR)) is straightforward and easy to interpret, and therefore is adopted in this thesis. The idea of this algorithm is stated next.

2.6.2.2 A Reported Support Vector Ordinal Ranking Algorithm

In an ordinal ranking problem, a certain number of ranks (e.g. L) need to be identified. The number of ranks in an ordinal ranking problem are not smaller than three (i.e. $L \geq 3$); otherwise, the ordinal information can not be expressed. Chu et al. [11] proposed an ordinal ranking algorithm (named "SVOR") based on SVM. The idea is briefly stated as follows. The support vector formulation attempts to find an optimal mapping direction $\mathbf{w}$, and $r-1$ thresholds which define $r-1$ parallel discriminant hyperplanes for the r ranks correspondingly, as shown in Figure 1.8. The point ($\mathbf{x}$) satisfying $b_{j-1} < \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) < b_j$ are assigned the rank j. The ranking model is thus

$$z = (\text{rank })j, \quad \text{if} \quad b_{j-1} < \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) < b_j \tag{2.77}$$

Data points that are located outside the margin of a threshold will be penalized. For a threshold b_j, the function values ($\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i^k)$) of all points from all the lower ranks should be less than the lower margin b_{j-1}. If data point violates this requirement, then

$$\xi_{ki}^{j} = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i^k) \tag{2.78}$$

is taken as the error associated with the point $\mathbf{x}_i^k$ for b_j, where k is the true rank of $\mathbf{x}_i^k$ and $k \leq j$. Similarly, the function values of all points for the upper rank should be greater than the boundary b_{j+1}, otherwise,

$$\xi_{ki}^{*j} = (b_j + 1) - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i^k) \tag{2.79}$$

is the error associated with the point $\mathbf{x}_i^k$ for b_j where $k > j$. Figure 2.13 gives an illustration of this idea in a 2-dimensional space. For the threshold b_1, points in Figure 2.13 with rank

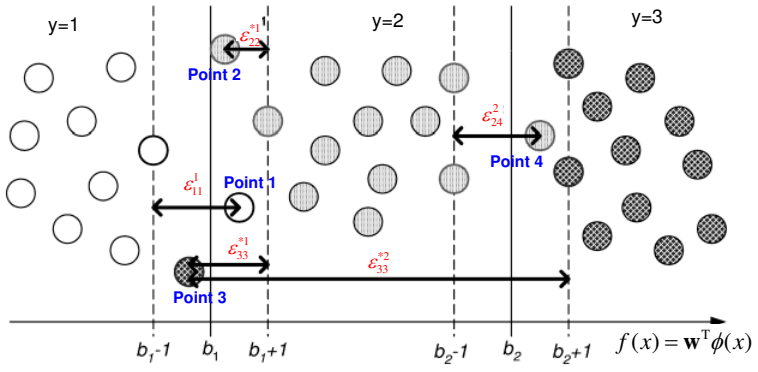


Figure 2.13: Illustration of calculation of slack variables in SVOR algorithm [11]

$z = 1$ are supposed to locate on the left-hand side of $b_1 - 1$. Point 1 violates this constraint, so ξ_{11}^1 is associated with this point. Points with ranks higher than 1 (i.e. $z = 2$ or $z = 3$) are supposed to be on the right-hand side of $b_1 + 1$. Points 2 and Points 3 are on the left-hand side of $b_1 + 1$, so ξ_{22}^{*1} and ξ_{33}^{*1} are associated with Point 2 and Point 3, respectively. Similarly, for the threshold b_2, ξ_{24}^2 is associated with Point 4 because it is not on the left-hand side of $b_2 - 1$; is associated with Point 3 because it is not on the right-hand side of $b_2 + 1$. Considering all the error terms associated with all $r - 1$ thresholds, the primal problem to find the optimal $\mathbf{w}$ and thresholds $\boldsymbol{b}$ are defined as follows:

$$
\begin{aligned}
\underset{\mathbf{w},\boldsymbol{b},\xi}{\text{minimize}} \quad & \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{j=1}^{L-1}\left(\sum_{k=1}^{j}\sum_{i=1}^{n^k}\xi_{ki}^{j} + \sum_{k=j+1}^{j}\sum_{i=1}^{n^k}\xi_{ki}^{*j}\right) \\
\text{subject to} \quad & \mathbf{w}\cdot\phi(\mathbf{x}_{\mathbf{i}}^{\mathbf{k}}) - b_j \leq -1 + \xi_{ki}^{j}, \\
& k = 1, 2, \ldots, j \text{ and } i = 1, 2, \ldots, n^k \\
& \mathbf{w}\cdot\phi(\mathbf{x}_{\mathbf{i}}^{\mathbf{k}}) - b_j \leq +1 + \xi_{ki}^{*j} \\
& k = j+1, j+2, \ldots, L \text{ and } i = 1, 2, \ldots, n^k.
\end{aligned}
\tag{2.80}
$$

where j runs over $1, 2, \ldots, L-1$ and n^k is the number of data points with the rank of k. By solving this optimization problem, the optimal $\mathbf{w}$ and $\boldsymbol{b}$ are found, and thus the ranking model (Equation (2.77)) is built.

2.7 Summary

This chapter provides the fundamentals on techniques that will be used later in this thesis. The conventional Fourier spectrum and the full spectrum are introduced in Section 2.1. Empirical mode decomposition is introduced in Section 2.2. Definitions of different types of variables are given in Section 2.3. Correlation coefficients, including Pearson correlation coefficient and Polyserial correlation coefficient, are introduced in Section 2.4. Different rough sets models, including Pawlak rough set, neighborhood rough set, dominance rough

set and fuzzy preference based rough set, are introduced in Section 2.5. Two machine learning algorithms, support vector classification and support vector ordinal ranking, are introduced in Section 2.6.

In machine-learning-based fault diagnosis, fault detection and isolation can be regarded as a classification problem. Support vector classification is employed for this purpose in Chapter 4. Neighborhood rough set model is applied for feature selection in classification problems in this chapter.

Fault identification is studied from signal-based diagnosis point of view in Chapter 5 by employing Fourier spectrum, empirical mode decomposition and fuzzy preference rough set.

Fault identification is studied from machine-learning-based diagnosis point of view using support vector ordinal ranking in Chapter 6. A feature selection method is proposed based on correlation coefficients in this chapter. In the next chapter, the experimental systems and the lab experimental data acquisition are described.

Chapter 3

Experimental Data

This thesis aims to develop effective methods for fault diagnosis of rotating machinery. In order to examine the effectiveness of the proposed methods in later chapters for industrial applications, the data collected from two test rigs will be used. These test rigs are: a slurry pump test rig and a planetary gearbox test rig. The two test rigs were designed and established for collaborative research projects between Reliability Research Lab (Department of Mechanical Engineering, University of Alberta) and Syncrude Research Center. The experiments completed as described in this chapter were conducted by researchers in Reliability Research Lab of the University of Alberta and engineers of Syncrude Canada Ltd. Technical reports [2, 3, 13, 15, 16] were prepared by the research team members documenting these experiments and data collected. The author of this thesis participated in some of these experiments. However, the author did not design or perform these experiments specifically for this thesis research. The data collected will simply be used to examine the methods to be reported later in this thesis. Materials in this chapter are based on [2, 3, 13, 15, 16].

The two test rigs are used, because they are representative rotating machines. The slurry pump is structurally simple, but the flow inside the pump is complicated and affects the vibration signals measured on the casing. This causes difficulties in pump fault diagnosis. The planetary gearbox is structurally complicated , which makes the fault diagnosis of planetary gearbox a challenging problem. This chapter introduces the two test rigs and experimental data collected that will be utilized in later chapters of this thesis.

3.1 Slurry Pump Test Rig

Centrifugal pumps are widely used for moving fluids in many applications, such as oil sands and mining. Figure 3.1 shows the structure of a centrifugal pump. The two main components of a centrifugal pump are the impeller and the volute. The impeller produces fluid velocity. The volute converts velocity to pressure. The distance between the impeller and the volute is the smallest at the cutwater.

As slurries contain abrasive and erosive solid particles, the impeller in a centrifugal slurry pump is subjected to severe wear. This is a main cause of reduced pump performance

and eventual failure. Monitoring the wear condition of impellers in pumps provides useful information for effective pump operation and maintenance [13].

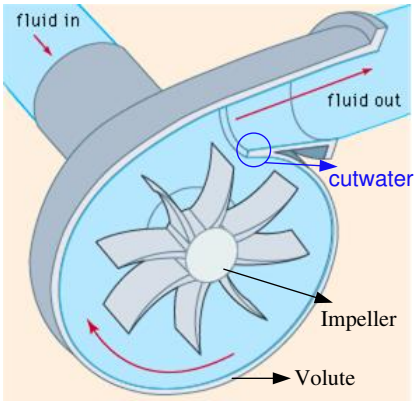


Figure 3.1: Structure of a pump [12]

Slurry impellers are more commonly of the closed type as shown in Figure 3.2. The impeller has side walls on both sides, called hub and shroud, respectively. The impeller also has several vanes to impart the centrifugal force to the slurries. The center of the impeller is called the impeller eye. The region near the impeller eye is called vane leading edge. The region at the rim of the vane is called vane trailing edge. Khalid and Sapuan [116] analyzed wears on slurry pump impellers and found that wear may occur at both the vane leading edge and the vane trailing edge. Experience of Syncrude engineers confirmed these findings.

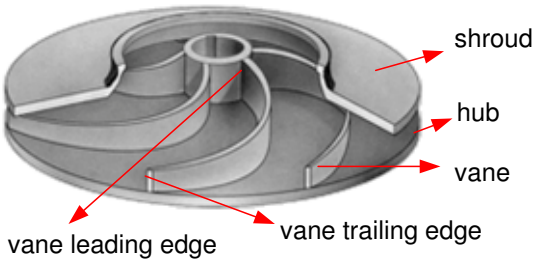


Figure 3.2: Structure of an impeller [12]

3.1.1 System Description

The schematic diagram of the experimental test rig at Reliability Research Lab is shown in Figure 3.3. This rig consists of the following key components [2, 13]:

- Slurry pump: Warman 3/2 CAH slurry pump with a closed impeller having five vanes. Side view of the impeller is given in Figure 3.4.

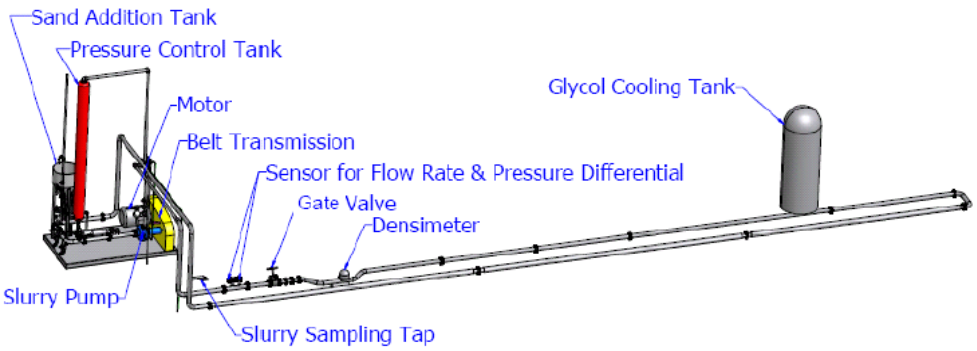


Figure 3.3: Schematic of the laboratory test-rig pump loop [13]

- Motor: 40 HP drive motor complete with variable frequency drive.
- Data acquisition system: 12-channel National Instruments SCXI system.
- PLC control panel: designed to control and monitor the operation of the system.
- Computer: Dell Inspiron 9200 laptop computer for data collection via LabView.
- Others: sensors, seal water pump, inlet pressure control tank, sand addition tank, safety rupture disk, various valves, pipes, and the glycol cooling system.

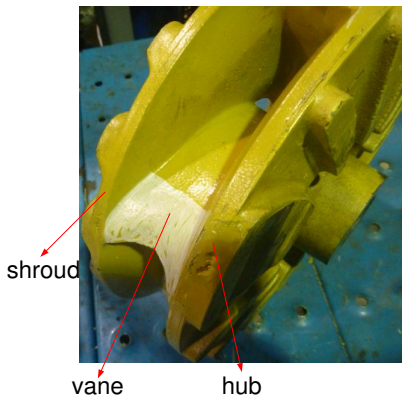
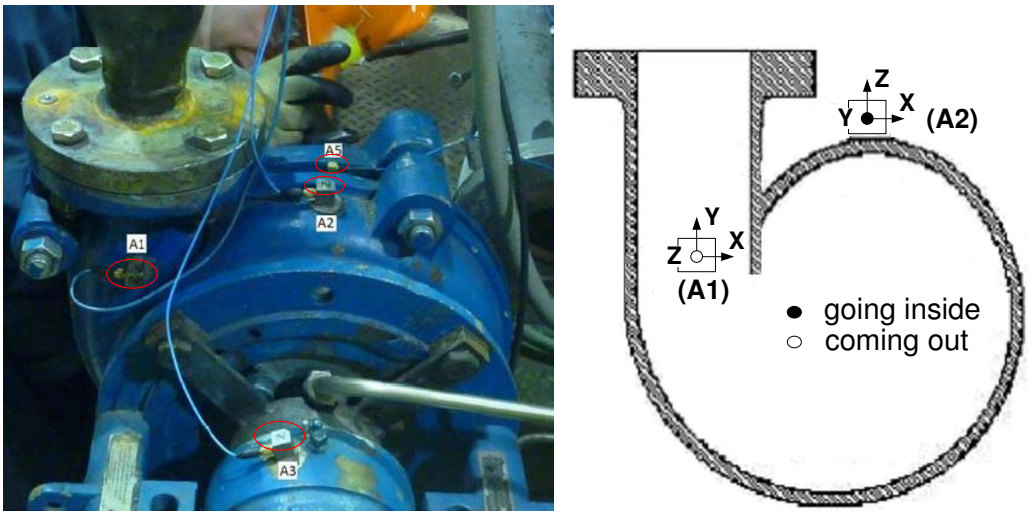


Figure 3.4: Side view of the impeller [2]

With this test rig, three types of experimental data were collected: (1) vibration, (2) pressure pulsation, and (3) process data. Vibration data were measured by accelerometers. Four accelerometers were installed, as shown in Figure 3.5 (a). These four accelerometers were labeled A1, A2, A3, and A5. Accelerometers A1, A2 and A3 are three-axis sensors which measure vibrations in three co-ordinate directions X, Y and Z. Accelerometer A5 is a uni-directional sensor. Directions of A1, A2 and A3 are given in Figure 3.5 (b). For the simplicity of description, the channels of these accelerometers will be referred to as A1-X, A1-Y, A1-Z, A2-X, A2-Y, A2-Z, A3-X, A3-Y, and A3-Z. Pressure pulsations were

measured by two dynamic pressure transducers located on the pump outlet pipe. They are labeled PL and PH in Figure 3.6. Process data includes flow rate, motor power, motor speed, inlet pressure (static), outlet pressure (static), density of the pumping medium and temperature. The locations of two pressure gauges for the inlet/outlet static pressure are also shown in Figure 3.6.



(a) Locations of accelerometers

(b) Directions of accelerometers (A3 is the same as A2)

Figure 3.5: Locations and directions of three accelerometers [2]

Note that this test rig was designed for other purposes. The data collected were for other purposes initially. This chapter introduces only the experiments that will be used in later chapters. Moreover, vibration data measured by three tri-axial accelerometers (A1, A2 and A3) only will be used in this thesis.

3.1.2 Impeller Vane Leading Edge Damage Experiments

With the vane leading edge damage (LED), the vane length of an impeller is reduced. According to [116], the length reductions of different vanes of an impeller are quite similar. Field observations also confirmed that the damage was usually uniform on all vanes of an impeller [2]. The leading edge wear of a vane may reduce the vane length by 40% before the impeller gets thrown out [2]. Thus in our lab experiments, the severe leading edge damage was designed to be 40% vane length reduction from the leading edge side. The total vane length of an undamaged impeller is 123 mm. The 40% vane length reduction corresponds to 49.20 mm and the volume reduction of fluid passage in impeller is calculated to be a certain value, say V. The moderate and slight damage levels were set to the volume reductions of 1/2 V and 3/4 V, which correspond to 29.3% (36 mm) and 20.3% (25 mm) reduction in the vane length, respectively. Table 3.1 lists the quantitative values for each damage level in LED experiments. Figure 3.7 shows an impeller with severe vane leading

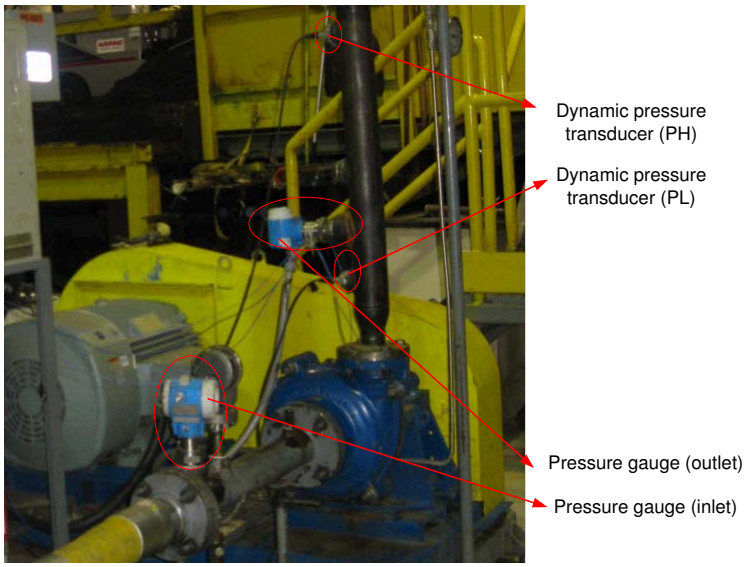


Figure 3.6: Locations of two dynamic pressure transducers and two pressure gauges [2]

edge damage.

Table 3.1: Damage levels in terms of vane length for LED [2]

Damage level	Reduction in vane length	Remaining vane length
Baseline (No Damage)	0	123 mm
Slight Damage	20.3% (25 mm)	123−25=98 mm
Moderate Damage	29.3% (36 mm)	123−36=87 mm
Severe Damage	40% (49.20 mm)	123−49.2=73.8 mm

According to Table 3.1, each of the five vanes of the impeller were artificially shortened using Electro Discharge Machining (EDM) to get slight damaged, moderate damaged and severe damaged impellers for LED experiments. The four damage levels mimic the life cycle of an impeller from normal to failure because of vane leading edge damage. Experiments were conducted using water as the pumping medium. First, the undamaged impeller is installed. Then three experiments were conducted at each of the three flow rate (70% BEPQ, 85% BEPQ, and 100% BEPQ where BEPQ is the flow rate corresponding to the best efficiency point of the pump) when the pump was running 2400 Revolution Per Minute (RPM). At each flow rate, six-minute vibration data were collected from each of the three accelerometers (A1, A2 and A3). The sampling frequency is 5 KHz. The same procedures above were repeated for each of the other three damaged impellers.

3.1.3 Impeller Vane Trailing Edge Damage Experiments

With the trailing edge damage (TED), the vane length of an impeller is reduced. Similar to LED experiments (Section 3.1.2), design of TED levels are based on the reduction of the

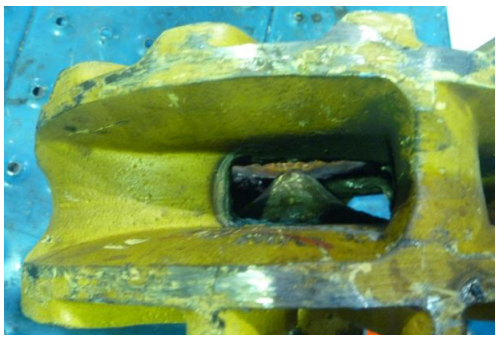


Figure 3.7: An impeller with severe vane leading edge damage [2]

fluid passage volume. The severe TED level is set to have the same fluid passage volume reduction as the severe LED, i.e. V, which corresponds to the vane length reduction of 29.3% (i.e 36 mm) from the trailing edge side of the impeller. Slight and moderate damage levels are set to have the volume reduction of 1/2 V and 3/4 V, respectively. They correspond to the vane length reduction of 22% (27 mm) and 26% (32 mm), respectively. Table 3.2 lists the quantitative values for each damage level in TED experiments. Figure 3.8 shows an impeller with severe vane trailing edge damage.

Table 3.2: Damage levels in terms of vane length for TED [2]

Damage level	Reduction in vane length	Remaining vane length
Baseline (No damage)	0	123mm
Slight damage	22% (27 mm)	123−27=96 mm
Moderate damage	26% (32 mm)	123−32=91 mm
Severe damage	29.3% (36 mm)	123−36=87 mm

Figure 3.8: An impeller with severe vane trailing edge damage [2]

According to Table 3.2, each of the five vanes of the impeller were artificially fabricated using electro discharge machining (EDM) to get slight damage, moderate damage and severe damage for TED experiments. The four damage levels mimic the life cycle of an impeller from normal to failure because of TED. Experiments were conducted using water as the pumping medium. First, the undamaged impeller is installed. Then three experiments

were conducted at each of the three flow rate (70% BEPQ, 85% BEPQ, and 100% BEPQ) when the pump was running 2400 Revolution Per Minute (RPM). At each flow rate, six-minute vibration data were collected from each of three accelerometers (A1, A2 and A3). The sampling frequency is 5 KHz. The same procedures above were repeated for each of the three damaged impellers.

In the above experiments, the author participated in the design of different damage levels and the collection of experimental data. The TED and LED vibration data will be used in Chapters 4, 5 and 7. In Chapter 4, diagnosis of fault types (i.e. no damage, trailing edge damage and leading edge damage) is conducted. In Chapter 5, the fault levels (i.e. damage levels of TED and LED) are identified. In Chapter 7, both fault types and fault levels are diagnosed. The vibration data collected at three flow rates are used in these chapters, unless otherwise specified.

3.2 Planetary Gearbox Test Rig

Planetary gearbox is a type of gearbox consisting of one or more planet gears, a sun gear and a ring gear. Figure 3.9 shows the structure of a planetary gearbox. The four planet gears are mounted on a carrier which itself may rotate relative to the sun gear. The four planet gears also mesh with a ring gear. Planetary gearboxes have many advantages, e.g. high power output, small volume, multiple kinematic combinations, pure torsional reactions and coaxial shafting [117]. They are widely used in oil sands, helicopters, trucks and other large-scale machinery.

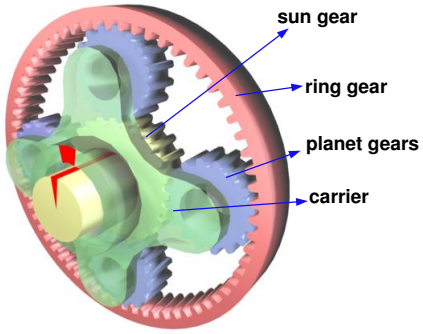


Figure 3.9: Structure of a planetary gearbox [14]

Planetary gearboxes are structurally more complicated than fixed-shaft gearboxes, and possess several unique behaviors [118]. For instance, gear mesh frequencies of planetary gearboxes are often completely suppressed, and sidebands are not as symmetric as those of fixed-shaft gearboxes [117]. Therefore, there is a need to develop effective fault diagnosis methods for planetary gearboxes.

3.2.1 System Description

The planetary gearbox test rig shown in Figure 3.10 was designed to perform controlled experiments for developing a reliable diagnosis system for the planetary gearbox. The test rig includes a 20 HP drive motor, a bevel gearbox, two planetary gearboxes, two speed-up gearboxes and a 40 HP load motor. The load was applied through the drive motor. A torque sensor was installed at the output shaft of the second stage planetary gearbox. The transmission ratios of each gearbox are listed in Table 3.3.

Table 3.3: Specification of the planetary gearbox test rig [3]

Gearbox		Number of teeth	Ratio
Bevel	input gear output gear	18 72	4
The first-stage planetary	sun gear three planet gear ring gear	28 62 152	6.429
The second-stage planetary	sun gear four planet gear ring gear	19 31 81	5.263
The first-stage speed-up	input gear middle gear 1 middle gear 2 output gear	72 32 80 24	0.133
The second-stage speed-up	input gear middle gear 1 middle gear 2 output gear	48 18 64 24	0.141

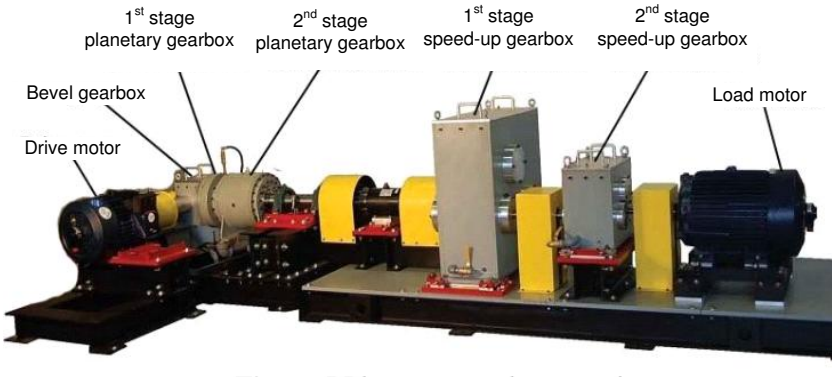


Figure 3.10: The planetary gearbox test rig [3]

Figure 3.11 shows the schematic diagram of the two planetary gearboxes. There are three planet gears in the 1st stage planetary gearbox and four planet gears in the 2nd stage planetary gearbox. The 1st stage sun gear is connected to the bevel gear by shaft #1. The

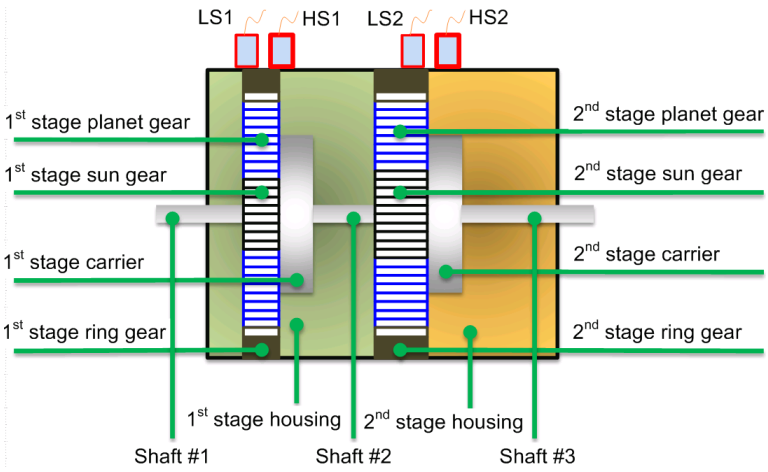


Figure 3.11: Schematic of the planetary gearbox [3]

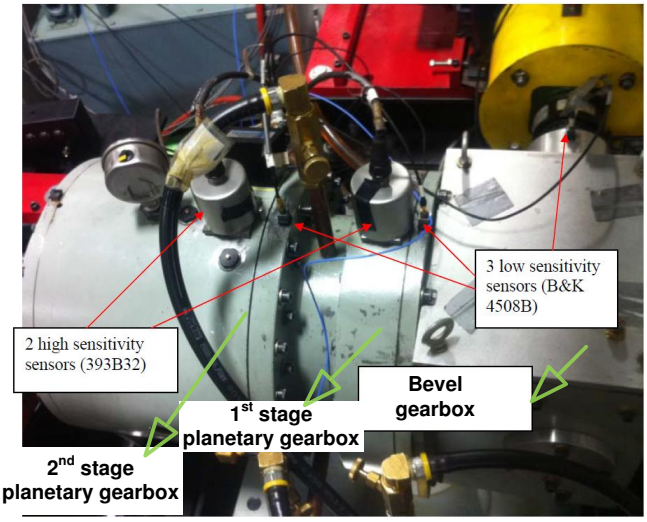


Figure 3.12: Locations of accelerometers [15]

1st stage planet gears are mounted on the 1st stage carrier which is connected to the 2nd stage sun gear by shaft #2. The 2nd stage carrier is located on shaft #3. Ring gears of the 1st stage and the 2nd stage planetary gearboxes are mounted on the housing of their stages, respectively.

In this test rig, five accelerometers and two acoustic sensors are installed. Two acoustic sensors are installed on the casing of the 2nd stage planetary gearbox. One low-sensitivity accelerometer is installed on the casings of the bevel gearbox. One low-sensitivity accelerometer and one high-sensitivity accelerometer are located on the casing of the 1st planetary gearbox; they are called LS1 and HS1, respectively. One low-sensitivity accelerometer and one high-sensitivity accelerometer are located on the casing of the 2nd planetary gearbox; they are called LS2 and HS2, respectively. For LS1 and LS2, the sensitivity is 99.4 mV/g, and the measurement frequency range is 0.3 Hz - 8 KHz. For HS1 and HS2,

the sensitivity is 5.0 V/g, and the measurement frequency range is 0.1 Hz - 300 Hz. The locations of the accelerometers are shown in Figure 3.12.

Note that this test rig was designed for other purposes, not specifically for this thesis. This chapter introduces only the experiments that will be used in later chapters. Moreover, vibration data measured by four tri-axial accelerometers (LS1, LS2, HS1 and HS2) only will be used in this thesis.

3.2.2 Pitting Experiments

Pitting is a form of extremely localized corrosion that leads to the creation of small holes in the gear tooth. It occurs due to the repeated loading of tooth surface and the contact stress exceeding the surface fatigue strength of the material. The pit itself causes stress concentration and soon the pitting spreads to adjacent region till the whole surface is covered. Subsequently, higher impact load resulting from pitting may cause fracture of already weakened tooth [16]. Based on stress calculations [16], the 2nd stage planet gears are highly stressed and are more likely to suffer from pitting. So pitting was artificially created using Electro Discharge Machining (EDM) on one of the four planet gears on the 2nd stage planetary gearbox. This section describes the experimental design and data collection for pitting experiments. Materials in this Section are from [16] and [15].

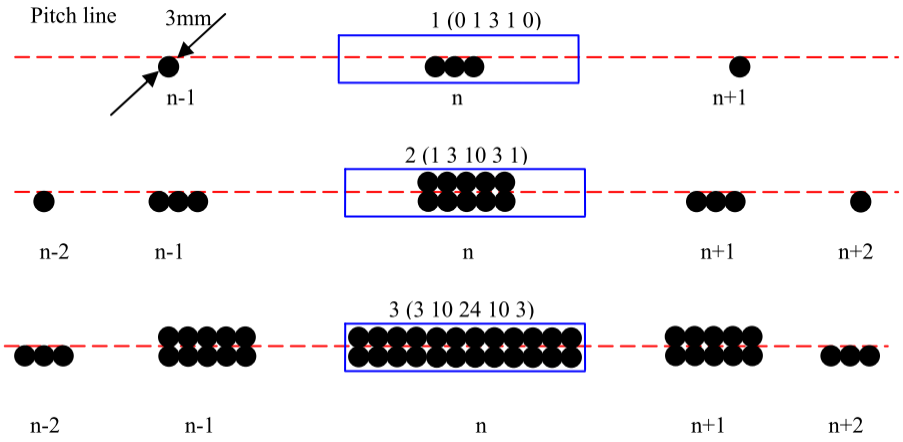


Figure 3.13: Schematic of pitting damage (slight, moderate and severe - from top to bottom) on the n^{th} tooth and its neighboring teeth [16]

Based on the literature review [119], the size and number of pits are defined. Circular holes with the diameter of 3mm and the depth of 0.1mm are created on the planet gear teeth. The number of pits are varied to mimic the slight, moderate and severe pitting. Figure 3.13 provides a schematic of the pitting levels on gear tooth. Figure 3.14 shows the pits created on planet gears. The rationales on number of pits for each pitting level are listed below.

1) Slight pitting: three teeth having totally five pits in a gear (three pits on one tooth and one pit on each of the two neighboring teeth). The percentages of simulated pitted area

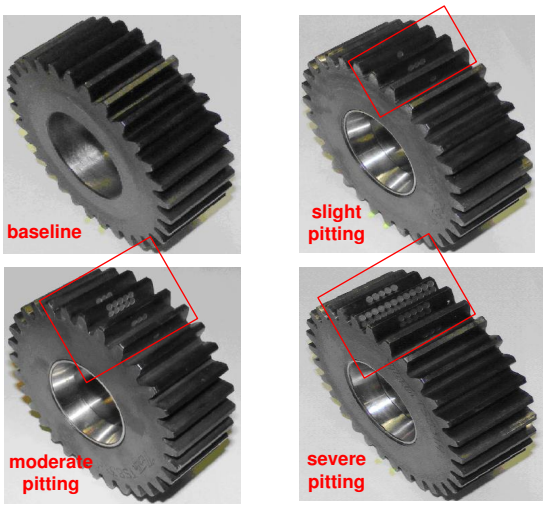


Figure 3.14: Planet gears with artificially created pitting damage of different levels [16]

are 2.65%, 7.95%, and 2.65% for the three teeth. This design corresponds to ASM level 2 pitting according to ASM handbook [120] (3%-10% of the tooth surface area).

2) Moderate pitting: five teeth having totally 18 pits in a gear (10 pits on one tooth, three pits on each of the two immediate neighboring teeth, and one pit on each of the next neighboring teeth on symmetric sides). The pitting areas of the five teeth are 2.65%, 7.95%, 26.5%, 7.95% and 2.65%, respectively. The most pitted tooth corresponds to ASM level 3 pitting (15%-40% of tooth surface area) [120].

3) Severe pitting: five teeth having totally 50 pits in a gear (24 holes on one tooth, 10 pits on each of the two immediate neighboring teeth and three pits on each of the next neighboring teeth on symmetric sides). The pitting areas of the five teeth are 7.95%, 26.5%, 63.6%, 26.5% and 7.95%, respectively. The most pitted tooth corresponds to ASM level 4 pitting (50% - 100% of tooth surface area) [120].

There are four planet gears in the 2nd stage planetary gearbox (See Table 3.3). During the pitting experiments, three normal planet gears and one pitting damaged planet gear are installed in the test rig. For each pitting level, experiments were conducted on two separate days. On each day, the experiment was run at four drive motor speeds (300, 600, 900, and 1200 RPM) and two loading conditions (namely, low load and high load). At the low load condition, the load motor was off. This does not mean that the planetary gearboxes encountered a zero load, as there were friction in the two speedup gearboxes and the rotor in the load motor was also rotating. According to the readings of the torque sensor, at this low load condition, the load that was applied at the output shaft of the planetary gearboxes ranged from 191.9 [N-m] to 643.6 [N-m]. The high load condition was selected based on the gear materials and stress calculation [16]. The loading applied by the load motor was adjusted to reach an average of 1130 [N-m] as displayed by the torque sensor, so that the system would run with a comfortable safe margin. The actual readings of the torque sensor

fluctuated from 812.9 [N-m] to 1455.2 [N-m]. At each combination of the running speed and the loading condition, five-minute vibration data were collected from each of the five accelerometers at the sampling frequency of 10 KHz.

3.3 Summary

The slurry pump test rig and the planetary gearbox test rig designed for other purposes were described in this chapter. The experimental data were collected for other purpose initially, some of which will be used to examine the effectiveness of fault diagnosis methods to be investigated in later chapters.

In slurry pump experiments, vibration data for two fault types (i.e. vane trailing edge damage and vane leading edge damage) and four fault levels (i.e baseline, slight, moderate and severe) for each fault type will be analyzed. In planetary gearbox experiments, vibration data for one fault types (i.e. pitting) with four fault levels (i.e. baseline, slight, moderate and severe) will be analyzed. In Chapter 4, slurry pump data will be employed to validate a proposed feature selection method for a machine-learning-based diagnosis of faut types. In Chapter 5, slurry pump data are used to validate two signal-based methods for diagnosis of fault levels. In Chapter 6, gear pitting data are used to validate a machine-learning based method for diagnosis of fault levels. In Chapter 7, slurry pump data are used to validate a machine-learning-based method for diagnosis of both fault types and fault levels.

Chapter 4

A Feature Selection Method Based on Neighborhood Rough Set for Machine-Learning-Based Fault Detection and Isolation

Fault detection and isolation (FDI) refers to the detection of the presence of a fault and the determination of the fault type. FDI is the first step in fault diagnosis, and thus is studied in this chapter. Signal-based methods and machine-learning-based methods are two ways for fault diagnosis. Based on literature review in Chapter 1, signal-based methods are often used for detection of the presence of a fault. When there are many fault types involved, machine-learning-based methods are often used. In this chapter, the machine-learning based methods are studied for FDI. For the convenience of description, the terminology of diagnosis of fault types is used in the place of FDI.

Machine-learning-based methods regard different fault types as different classes, and use classifiers to determine the fault type for an input data (i.e. a feature vector). A feature vector is a set of features which are extracted from measured signals (e.g. vibration signals). Features reflect health information of a monitored system. Classifiers map the information provided by input features to the fault type. Some features, however, might be irrelevant or redundant to the fault type. The irrelevant and redundant information would decrease the performance of a classifier [53]. Thus, selecting features that are relevant to fault types and are not redundant to each other is an important step for successful diagnosis. This step is called feature selection.

As discussed in Chapter 1, rough set [64] has been shown to be an effective tool for feature selection. Its main advantage is that it requires no additional parameters (e.g. the number of features to be selected) to operate other than the supplied data. The fundamentals of rough set have been provided in Section 2.5.

The classical rough set model (Pawlak rough set) is best suited to nominal features. In fault diagnosis of rotating machinery, vibration signals are often measured. Features ex-

tracted from vibration signals are often continuous [73, 94–96]. For definitions on nominal and continuous features, please refer to Section 2.3. Hu et al. [10] proposed a neighborhood rough set model for continuous feature selection. As shown in Section 2.5.2, neighborhood size is an important factor that affects feature evaluation. However, a common neighborhood size that works for all features is usually hard to obtain. Thus, determination of the neighborhood size is a problem to be solved.

In this chapter, the effect of neighborhood size on feature selection is discussed. The neighborhood rough set model reported in [10] is modified, based on which a feature selection method is proposed. The rest of the chapter is organized as follows. Section 4.1 describes the background on the neighborhood rough set. Section 4.2 proposes the modified neighborhood rough set. Section 4.3 presents a feature selection method based on the modified neighborhood rough set. Sections 4.4 applies the proposed method to the diagnosis of fault types for impellers in slurry pumps. Summary comes in Section 4.5. The major contribution of this chapter has been published in [95]. [1]

4.1 Background

The neighborhood rough set proposed by Hu et al. [10] is described in Section 2.5.2. Its concept is illustrated with an example in a two-dimensional feature space shown in Figure 4.1. The set including all samples is denoted by U. Samples labeled with "*" are from one class (class 1) and samples labeled with "+" are from the other class (class 2). The neighborhood of u_i, denoted by $\delta(u_i)$, is the set including samples whose Euclidian distances from u_i are less than the neighborhood size, δ. Let D_j be the set of samples that belong to class j and $\underline{N}D_j$ be the lower approximation of D_j. If all samples in $\delta(u_i)$ have the same label as u_i, then u_i belongs to $\underline{N}D_j$. Taking samples u_i $(i = 1, 2, 3)$ as examples. By checking Figure 4.1, it is found that $\delta(u_1) \subseteq D_1$ and $\delta(u_3) \subseteq D_2$, whereas $\delta(u_2) \cap D_1 \neq \emptyset$ and $\delta(u_2) \cap D_2 \neq \emptyset$, thus $u_1 \in \underline{N}D_1$ and $u_3 \in \underline{N}D_2$. The above comparisons are conducted for each samples, and it is found that samples in Region A_1 and A_3 are lower approximation of D_1 and D_2, respectively. That is $\underline{N}D_1 = A_1$ and $\underline{N}D_2 = A_3$. Let D be the set including labels for all sample. The lower approximation of D is $\underline{N}D = \underline{N}D_1 \cup \underline{N}D_2$. The approximation quality, as defined with Equation (2.31), is the ratio between the number of samples in $\underline{N}D$ and the total number of samples in U.

The approximation quality can be used in feature selection as a measure to evaluate the performance of a feature (or a set of features). The higher the approximation quality, the more significant a feature (or a set of features). The neighborhood rough set model reported by Hu et al. [10] uses a common neighborhood size (δ) for all features. The neighborhood size (δ) largely influences the approximation ability of a feature (or a feature set), as shown

[1] A version of this chapter has been published in "Xiaomin Zhao, Qinghua Hu, Yaguo Lei and Ming J Zuo, 2010. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 224: 995-1006."

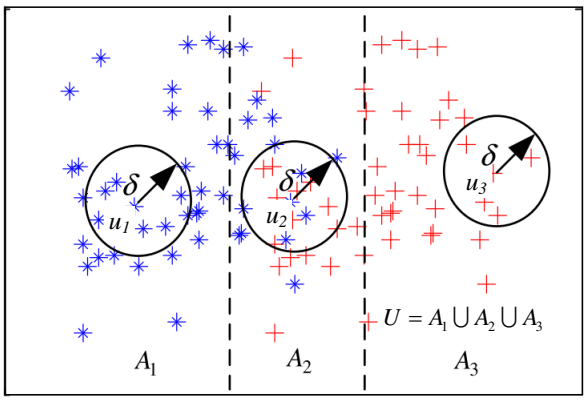


Figure 4.1: An example illustrating the concept of neighborhood rough set [10]

in Example 2 in Section 2.5.2.

We now check the physical meaning of neighborhood size (δ). Neighborhood size (δ) can be interpreted as the level of noise which inevitably contains in features [121]. Take a feature, that is the amplitude at shaft rotating frequency calculated from a vibration signal of a pump, for example. It is expected that the feature value represents that health condition of the pump. However, a precise value can hardly be obtained in practice. In another word, even the health condition is the same, the value of the feature still inevitably fluctuates, mainly because of the following two reasons. (1) The ambient environment is not completely clean during the measurement. That is, the health condition is not the only factor that affects the vibration signal. There are uncertain factors that influence and cause the fluctuation of vibration amplitudes, e.g. how components were installed in the pump system. Even clear procedures are followed during installation, it is hard to ensure that the installation of each component is exactly the same as designed. (2) The resolution of the vibration sensor also affects the precision of the vibration measurement. The resolution of the sensor is the smallest change the sensor can detect in the quantity that it measures. Because the resolution can not reach infinitesimal, the real measurement is not exactly the true physical quantity; thus noise is in the collected vibration signals. Here all the factors that cause the fluctuation of vibration measurement are referred to as noise.

So, the measured vibration signal, $x(t)$, is actually the true signal ($s(t)$) added by noise, $e(t)$, as shown in Figure 4.2. In a 2-D feature space, assume u_1 is the feature value obtained from the measured signal ($x(t)$), the true feature value is located in the region near u_1. The size of this region is determined by the level of noise. In neighborhood rough set, we call this region, the neighborhood of u_1 and its size is controlled by the neighborhood size (δ). The above is illustrated in Figure 4.3.

From the above discussion, it can be seen that the neighborhood size depends on the level of noise. The noise is contributed by the external factors and the sensor itself. In fault diagnosis, many sensors are often installed for vibration measurement. Each sensor measures vibration in different locations and directions. The measured vibration signal by

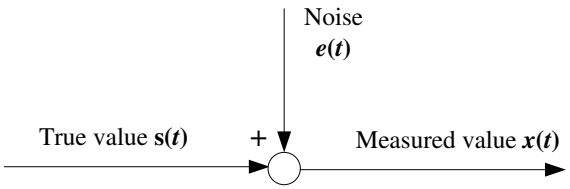


Figure 4.2: Relationship of a measured signal ($x(t)$) and noise ($e(t)$) [10]

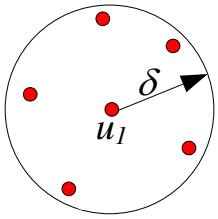


Figure 4.3: Neighborhood of a sample (u_1) [10]

each sensor is contaminated by different levels of noise. Thus the features obtained from different sensors are usually subjected to different levels of noise. If the same neighborhood size is used for two features with different noise levels as did in [10], the significance of each feature may be wrongly evaluated. Therefore, based on the above understanding, the modified neighborhood rough set is proposed.

4.2 Modified Neighborhood Rough Set

In this section, the effect of neighborhood size is analyzed, and then a modified neighborhood rough set model is proposed.

4.2.1 Effect of Neighborhood Size

The neighborhood size affects the value of approximation quality which is the measure of significance of a feature. Here we take two features as an example, say two features a_1, a_2 with the levels of noise δ_1 and δ_2, respectively. Assume a_2 is noisier than a_1, i.e. $\delta_1 < \delta_2$. The neighborhoods of u_i in terms of feature a_1 and feature a_2 are defined as:

$$\delta_{a_1}(u_i) = \{u_j | u_j \in U, \Delta_{a_1}(u_i, u_j) \leq \delta_1\}, \tag{4.1}$$

$$\delta_{a_2}(u_i) = \{u_j | u_j \in U, \Delta_{a_2}(u_i, u_j) \leq \delta_2\}. \tag{4.2}$$

If a common neighborhood size, δ, is used to compare the significance of a_1 and a_2, when a smaller value is chosen, say $\delta = \delta_1$, then $\delta_2 > \delta$ thus

$$\delta_{a_1}(u_i)|_\delta = \delta_{a_1}(u_i)|_{\delta_1}, \tag{4.3}$$

$$\delta_{a_2}(u_i)|_\delta = \{u_j | u_j \in U, \Delta_{a_2}(u_i, u_j) \leq \delta\} \tag{4.4}$$

$$\subseteq \{u_j | u_j \in U, \Delta_{a_2}(u_i, u_j) \leq \delta_2\} = \delta_{a_2}(u_i)|_{\delta_2}. \tag{4.5}$$

According to Equation (2.29), the lower approximations of D_i have the following relation:

$$\underline{N_{a_1}} D_i|_\delta = \underline{N_{a_1}} D_i|_{\delta_1}, \tag{4.6}$$

$$\underline{N_{a_2}} D_i|_\delta \supseteq \underline{N_{a_2}} D_i|_{\delta_2}. \tag{4.7}$$

According to Equation (2.31), the approximation qualities calculated by different neighborhood sizes have the following relation:

$$\gamma_{a_1}|_\delta = \gamma_{a_1}|_{\delta_1}, \tag{4.8}$$

$$\gamma_{a_2}|_\delta > \gamma_{a_2}|_{\delta_2}. \tag{4.9}$$

It means that the significance of a_2 is overestimated. Otherwise, if $\delta = \delta_2$ is used as the common neighborhood size, then $\gamma_{a_1}|_\delta < \gamma_{a_1}|_{\delta_1}$. This means that the significance of a_1 is underestimated.

Figure 4.4 further illustrates this effect of neighborhood size using a 2-class example. Circles and squares represent class 1 and class 2, respectively. Each class has three samples (u_i), and each sample is described by the value of feature a_1. Let D_j be the set of samples that belong to class j ($j = 1, 2$), $\underline{ND}_j$ be the lower approximation of D_j, and $\underline{ND} = \underline{ND}_1 \bigcup \underline{ND}_2$. The three subplots in Figure 4.4 shows the cases with different neighborhood sizes.

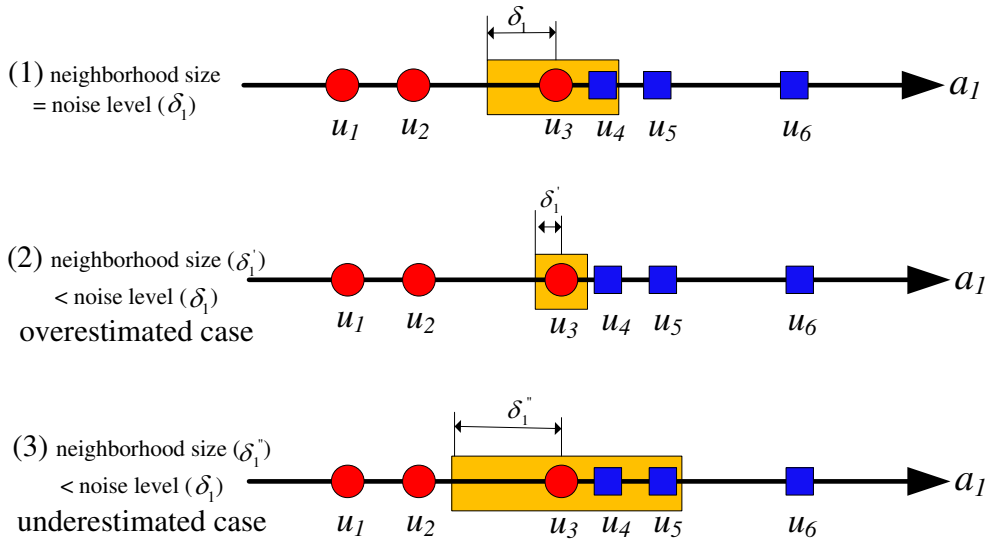


Figure 4.4: Effect of neighborhood size

- Figure 4.4 (1) marks the level of noise for a_1, that is δ_1. The neighborhood of u_1 includes two samples (i.e. u_1 and u_2), and the two samples are from the same class. According to Equation (2.29), we have $u_1 \in \underline{ND}_1$. For u_3, its neighborhood have two samples u_3 and u_4 and they are from different classes, so $u_3 \notin \underline{ND}_1$. The same

procedure is applied to other samples. Finally we have the lower approximation of the two classes, i.e. $\underline{ND}_1 = \{u_1, u_2\}$, $\underline{ND}_2 = \{u_5, u_6\}$. Thus $\underline{ND} = \{u_1, u_2, u_5, u_6\}$ and according to Equation (2.31), $\gamma_{a_1}|_{\delta_1} = 4/6 = 0.67$.

- In Figure 4.4 (2), a smaller neighborhood size, $\delta_1^{'}$, is used ($\delta_1^{'} < \delta_1$). The lower approximation of the two classes is found to be $\underline{ND} = \{u_1, u_2, u_3, u_4, u_5, u_6\}$. Thus the approximation quality $\gamma_{a_1}|_{\delta_1^{'}} = 6/6 = 1.00$. Because 1.00>0.67, that means a_1 is overestimated when the smaller δ is used.

- In Figure 4.4 (3), a larger neighborhood size, $\delta_1^{''}$, is used ($\delta_1^{''} > \delta_1$). The lower approximation of the two classes is found to be $\underline{ND} = \{u_1, u_2, u_6\}$. Thus the approximation quality $\gamma_{a_1}|_{\delta_1^{''}} = 3/6 = 0.50$. Because 0.50<0.67, that means a_1 is underestimated when the larger δ is used.

Feature a_i can be properly estimated only if its neighborhood size equals to the noise level it suffers. Therefore, it is necessary to apply different neighborhood sizes for different features based their noise levels. Considering the effect of the neighborhood size as discussed above, the original neighborhood rough set reported by Hu et al. [10] is modified by assigning different neighborhood sizes for different features, which is introduced next.

4.2.2 Modified Neighborhood Rough Set Using Multiple Neighborhood Sizes

Given a neighborhood decision system, $NIS = (U, C \bigcup D, N)$, where $U = \{u_1, u_2, \ldots u_n\}$ is a nonempty finite set of samples, $C = \{a_1, a_2, \ldots, a_m\}$ is a set of features describing each sample, $D = \{d_1, d_2, \ldots, d_L\}$ is the label specifying the class of each sample, and $N = [r_{ij}]$ is a matrix describing the neighborhood relation between each pair of samples, u_i and u_j. r_{ij} is given by Equation (2.26). In the calculation of r_{ij}, the neighborhood of u_i needs to be defined first. The neighborhood of u_i is defined with

$$\delta_C(u_i) = \left\{u_j : u_j \in U, \left|f(u_i, a_k) - f(u_j, a_k)\right| \leq \delta_k \text{ for } \forall a_k \in C\right\} \tag{4.10}$$

where $f(u_i, a_k)$ is the value of feature a_k for sample u_i, and δ_k is the neighborhood size for a_k.

The difference between the present work and the original neighborhood rough set [10] lies in the calculation of neighborhood for continuous features. In the modified neighborhood rough set, Equation (4.10) is used to calculate the neighborhood for sample u_i. In the original neighborhood rough set, Equations (2.23) and (2.25) are used. Take a two-dimensional feature space for example, i.e. $C = \{a_1, a_2\}$. If p in Equation (2.25) is set as 2 (that is Euclidian distance is calculated), then the neighborhood of u_i is calculated by Equation (4.11) in the original neighborhood rough set; and the neighborhood of u_i is calculated

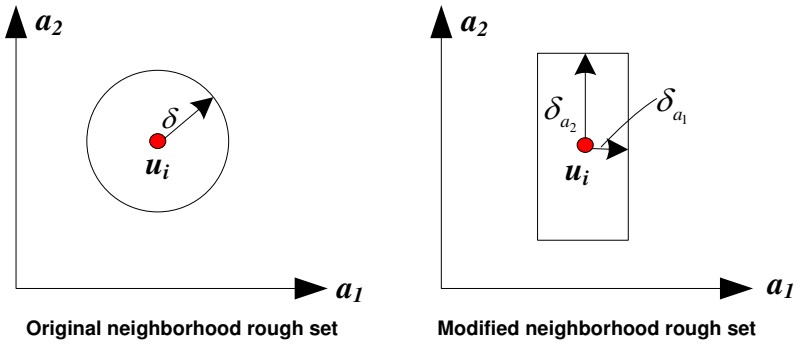


Figure 4.5: Neighborhood of a sample (u_i) in the original and modified neighborhood rough sets

by Equation (4.12) in the modified one.

$$\delta_{a_1 \cup a_2}(u_i) = \Big\{u_j : u_j \in U, \Big(\sum_{k=1}^{2}(f(u_i, a_k) - f(u_j, a_k))^2\Big)^{\frac{1}{2}} \leq \delta\Big\} \tag{4.11}$$

$$\delta_{a_1 \cup a_2}(u_i) = \tag{4.12}$$
$$\Big\{u_j : u_j \in U, \Big|f(u_i, a_1) - f(u_j, a_1)\Big| \leq \delta_1 \bigcap \Big|f(u_i, a_2) - f(u_j, a_2)\Big| \leq \delta_2\Big\}$$

Figure 4.5 illustrates the two neighborhoods defined by Equation (4.11) and (4.12). In the original neighborhood rough set, the neighborhood of u_i is a circle. The neighborhood is rectangle in the modified neighborhood rough set.

Furthermore, Equation (4.12) is mathematically equivalent to

$$\delta_{a_1, \cup, a_2}(u_i) = \delta_{a_1}(u_i) \cap \delta_{a_2}(u_i). \tag{4.13}$$

This means that the neighborhood of u_i with respect to a_1 and a_2 are the intersection of the neighborhood u_i with respect to a_1 and the neighborhood of u_i with respect to a_2. This characteristic can be extended to more than two features (say m), that is

$$\delta_{a_1 \cup a_2 \cdots \cup a_m}(u_i) = \delta_{a_1}(u_i) \cap \delta_{a_2}(u_i) \cdots \cap \delta_{a_m}(u_i). \tag{4.14}$$

Equation (4.14) enables us to calculate the neighborhood of each feature independently and the intersection will be the neighborhood of the set of all features. With the neighborhood defined, for a feature subset $B \subseteq C$, the approximation quality of B can be calculated following the same equation (Equation (2.31)), and is summarized below.

$$\gamma_{_B}(D) = \frac{|\underline{N_B} D|}{|U|} \tag{4.15}$$

where,

$$\underline{N_B} D \quad = \quad \bigcup_{l=1}^{L} \underline{N_B} D_l, \tag{4.16}$$

$$\underline{N_B}D_l \quad = \quad \left\{u_j : \Delta_B(u_j) \subseteq D_l, u_j \in U\right\}. \tag{4.17}$$

As introduced in Section 2.5, the approximation quality reflects the ability of a feature (or a feature subset) in approximating label D. The approximation quality is often used to evaluate the performance of a feature (or a feature subset) in feature selection. As stated in the beginning of this chapter, feature selection is to find a subset of features that are most relevant to the label and contain less redundant information. A feature subset that has the minimum number of features and has the higher approximation quality provides most relevant but less redundant information for machine learning. Rough set achieves feature selection by finding such a feature subset. A feature selection algorithm based on the modified neighborhood rough set is discussed below.

4.3 Feature Selection Based on Modified Neighborhood Rough Set

The approximation quality defined in rough set evaluates the performance of a feature subset in approximating labels (D). Feature selection based on rough set is achieved by searching a feature subset V that produces the highest approximation quality and has the least number of features.

Hu et al. [10] proposed a feature selection algorithm based on the original neighborhood rough set and sequential forward search strategy. The determination of the neighborhood size is a problem for the original neighborhood rough set. Hu et al. [10] suggested to find the optimum neighborhood size by enumeration. Enumeration has the disadvantage of being less efficient and time-consuming. Moreover, a common neighborhood size is used in the original neighborhood rough set, which causes poor estimation of the approximation quality of a feature (or a feature subset), as illustrated in Section 4.2.1.

In this section, the modified neighborhood rough set is used for feature selection. As introduced in Section 4.2.2, multiple neighborhood sizes are adopted in the modified neighborhood rough set. First, a neighborhood size should be determined for each feature.

4.3.1 Determination of Neighborhood Size

To determine the neighborhood size for a feature, the noise level that the feature encounters needs to be estimated. Depending upon how the feature is obtained, its noise level can be estimated in different ways. Here the vibration signal is taken as an example.

One commonly used type of features in vibration-based fault diagnosis is the amplitudes at some characteristic frequencies from the Fourier spectrum of a vibration signal. For such features, their noise levels can be approximated by checking the Fourier spectrum and estimating the noise amplitude in the frequency range enclosing the characteristic frequencies. Here the slurry pump data are taken as an example (the pump data will be

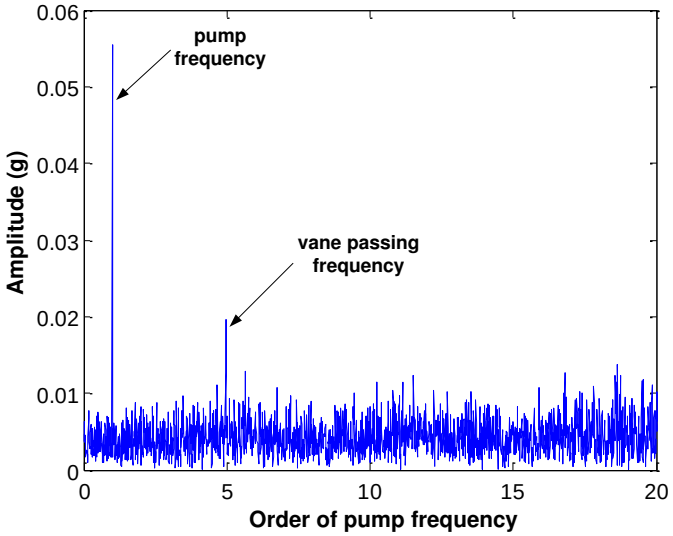


Figure 4.6: Fourier spectrum of a vibration signal from a pump

analyzed in Section 4.4 in detail). The vibration signal measured from the pump at the perfectly healthy condition is shown in Figure 4.6. Two amplitudes at the pump frequency and the vane passing frequency are respectively extracted as two features. It can be seen that there is white noise contained in the vibration signal because the amplitude values throughout the frequency range (except the pump frequency and the vane passing frequency) are equally likely. Thus the noise energy (σ) is estimated by Equation (4.18), where A_j is the amplitude at frequency j, and j is a frequency which is not pump frequency or the vane passing frequency and the range of j covers the frequencies whose amplitudes are extracted as features.

$$\sigma = \sqrt{\frac{1}{N}\sum_{j=1}^{N} A_j^2} \tag{4.18}$$

For other features, such as kurtosis and skewness, their noise levels can not be directly obtained as in Equation (4.18). In this case, it is suggested to collect N records of vibration signals, calculate N feature values and use the standard deviation of the N feature values to estimated the noise level [122].

4.3.2 A Feature Selection Method

In this section, the feature selection algorithm based on the modified neighborhood rough set is introduced. The same feature selection algorithm proposed by Hu et al. [10] is followed, replacing the approximation quality defined in the original neighborhood rough set in that algorithm with the approximation quality defined in the modified neighborhood rough set. The sequential forward search strategy which is also introduced in Section 1.2.2.2 is adopted.

The feature selection algorithm is presented in Table 4.1. The selection starts with an empty set, $S = \emptyset$. The significance of a feature a added to the feature subset S is evaluated using Equation (4.19). If $S = \emptyset$, then $\gamma_V(D) = 0$. Equation (4.19) is used to evaluate the significance of each feature that hasn't been included in S. The feature whose significance is the highest is added to S. If the significance value is zero for all features that are not included in S, then the selection stops and S is output.

$$R_a = \gamma_{V \cup a}(D) - \gamma_V(D) \tag{4.19}$$

Table 4.1: A feature selection method based on the modified neighborhood rough set

Input: C- the raw feature set; D- the set of decisions. Output: S- the set of selected features Step 1. S=$\emptyset$. //Initialization Step 2. for each $a_i \in C - S$, its significance (R_{a_i}) is calculated using Equation (4.19). Step 3. select a_k that $R_{a_k} = \max_i R_{a_i}$, if $R_{a_k} > 0$, then $S = S \bigcup a_k$, go to Step 4; otherwise, go to Step 5. Step 4. if $C = S$, then go to Step 5; otherwise, go to Step 2. Step 5. return S.

In next section, the feature selection algorithm are applied to diagnosis of pump fault types.

4.4 Application to Fault Detection and Isolation of Impellers in Slurry Pumps

In this section, fault detection and isolation (diagnosis of fault types) of impellers in slurry pumps is studied. As stated in Chapter 3, two fault types, impeller trailing edge damage (TED) and impeller leading edge damage (LED) are considered in this thesis. Detection of early fault is the most challenging problem [123]. However it is often desirable to detect faults as early as possible, so that the fault propagation can be monitored and preventive maintenance can be scheduled before the fault becomes severe. Therefore, in this chapter, the initial (slight) damage only is focused. Specifically, there are three fault conditions considered in this chapter, that is no damage (ND), slight TED and slight LED. Other advanced damage levels will be studied in Chapters 5 and 7. As described in Section 3.1, vibration data were collected for each fault type. The vibration data were saved as 54 samples, so totally there are 162 samples (i.e. 54 samples per fault type $\times$ 3 faut types) available.

As stated in Chapter 1, machine-learning-based fault diagnosis contains three steps: feature extraction, feature selection and fault classification. They are presented, for this specific application, in the following subsections. The focus here is feature selection.

4.4.1 Feature Extraction

Conventional Fourier spectrum (half spectrum) is a commonly used tool for feature extraction of pumps [17]. The mathematical description of conventional Fourier spectrum is given in Section 2.1. Amplitude at a certain frequency f represents the energy at f, and contains information on pump health condition [17, 95].

According to the technical bulletin of Warman slurry pump [124], the informative frequencies of a pump are: the pump rotating frequency (1X), its 2^{nd} harmonic (2X), 3^{rd} harmonic (3X), 4^{th} harmonic (4X) and the vane passing frequency (5X). Zhang et al. [125] showed that a flow pattern called (jet-wake) which may occur at the impeller outlet results in the second harmonic of the vane passing frequency (10X). Therefore, the six frequencies 1X, 2X, 3X, 4X, 5X and 10X are chosen as valuable frequencies. The amplitude values at the six frequencies are extracted as features. The slurry pump test rig has three tri-axial accelerometers as shown in Figure 3.5. The three tri-axial accelerometers produce nine channels. Each channel outputs a vibration signal. So the total number of features is 54 (6 features per channel $\times$ 9 channels).

4.4.2 Feature Selection

Now features are selected using the algorithm presented in Section 4.3. First, the neighborhood size needs to be determined for each feature. As discussed in Section 4.1, the neighborhood size reflects the noise level that a feature suffers. As stated in Section 4.3.1, the noise level for the pump feature (amplitude) obtained from a Fourier spectrum can be estimated by Equation (4.18) defined in the same Fourier spectrum. Each channel produces a vibration signal, and therefore a Fourier spectrum. Features from the same channel have the same neighborhood size. But features from different channels have different neighborhood sizes, because each channel has its own noise level depending on its location and the sensor characteristic. Samples from perfectly healthy condition are used to estimate the noise level. The noise levels of these samples are averaged to get the noise level for this channel. Table 4.2 lists the neighborhood size for each feature (noise level for each channel). It can be seen that features from different channels have different neighborhood sizes.

With the neighborhood size determined, the feature selection algorithm (Table 4.1) is applied, and three features are selected as shown in Table 4.3. In the next step, the selected features are imported into classifiers for diagnosis of impeller fault types.

Table 4.2: Neighborhood sizes for different features

Channel	A1X	A1Y	A1Z	A2X	A2Y
Feature No.	1~6	7~12	13~18	19~24	25~30
Neighborhood Size	0.0015	0.0013	0.0021	0.0011	0.0021
Channel	A2Z	A3X	A3Y	A3Z	
Feature No.	31~36	37~42	43~48	49~54	
Neighborhood Size	0.0009	0.0013	0.0010	0.0007	

Table 4.3: Feature name and source of each selected feature

Feature No.	Feature Name	Channel (Sensor-direction)
37	Amplitude at 5X	A1-Y
11	Amplitude at 2X	A2-X
20	Amplitude at 1X	A3-X

4.4.3 Fault Classification

In this step, the selected features are fed into classifiers to diagnose impeller fault types. The diagnosis result is used to evaluate the performance of selected features. Here the diagnosis result is expressed as classification error defined in Equation (4.20), where d_i and d_i' are the true fault type and the predicted fault type of the i^{th} sample respectively.

$$\frac{1}{n}\sum_{i=1}^{n} t_i \quad \text{where } t_i = \begin{cases} 1, & d_i = d_i' \\ 0, & \text{otherwise} \end{cases} \tag{4.20}$$

Three-fold cross validation is adopted to generate diagnosis results. Specifically, the whole data set was evenly split into three subsets, two of them are used as the training set and the remaining one as the test set. Using the training set, the parameter of a classifier is tuned to get a classification model. The classification model is then tested with samples in the test set and a classification error is obtained. The process is repeated three times, with each of the three subsets used exactly once as the test data. The three classification results are then averaged to produce a mean classification error. A standard deviation is also calculated based on the three classification results.

To show that the feature selection method is not classifier-sensitive, three commonly used classifiers are adopted for fault classification. They are probabilistic neural network (PNN), K-nearest neighbor (KNN), and support vector machine (SVM). Mechanisms of these classifiers are given in Section 1.2.2.3. Diagnosis results using the three classifiers are given next.

4.4.4 Results and Discussions

To illustrate the effect of the features selected by the modified neighborhood rough set, we compare it with (i) the case that feature selection is not performed (i.e. all 54 features are

used), and (ii) the case that feature selection is conducted by the original neighborhood rough set.

When selecting features using the original neighborhood rough set, one common neighborhood size (δ in Equation (2.23)) is used for all features needs to be determined. Hu et al. [10] stated that the classification error varies with the neighborhood size; however, the determination of a proper neighborhood size was not reported. They suggested to try different neighborhood size and choose the one that produce the smallest classification errors. Since the proper value of the neighborhood size for all features is not known, three different neighborhood sizes are tested: 0.0007, 0.0013 and 0.0021. They are the minimum, mean and maximum values of the nine noise levels for the nine channels listed in Table 4.2. The minimum value represents the case where the neighborhood size is smaller than its noise level for most features; and in this case, the features are likely to be overestimated. The maximum value represents the case in which the neighborhood size is larger than the noise level for most features; in this case, the features are likely to be underestimated. The mean value represents the case where the neighborhood size is closer to the noise level for most features. Table 4.4 shows the features selected under these three neighborhood sizes (δ).

Table 4.4: Features selected under different neighborhood sizes

Neighborhood size (δ)	Selected feature subset
0.0007	No. 37 (Amplitude at 1X from channel A3-X
0.0013	No. 37 No. 20 (Amplitude at 2X from channel A2-X)
0.0021	No. 17 (Amplitude at 5X from channel A1-Z) No. 35 (Amplitude at 5X from channel A2-Z) No. 3 (Amplitude at 3X from channel A1-X)

Table 4.5: Classification errors (mean ± standard deviation) generated by different feature subsets

Classifier / Feature subset		SVM	KNN	PNN	Average	
All features		0.111 ±0.064	0.045 ±0.034	0.101 ±0.083	0.086	
Features selected by modified neighborhood rough set		0.015 ±0.011	0.043 ±0.033	0.083 ±0.060	0.047	
Features selected by original neighborhood rough set	δ=0.0007	0.102 ±0.083	0.111 ±0.064	0.137 ±0.061	0.117	0.109
	δ=0.0013	0.028 ±0.016	0.059 ±0.048	0.099 ±0.060	0.062	
	δ=0.0021	0.150 ±0.043	0.124 ±0.101	0.168 ±0.061	0.147	

1) The effect of feature selection Table 4.5 shows the mean and standard deviation of three-fold cross validation for each classifier when different feature subsets were used. If all 54 features are used, the mean values of classification errors produced by SVM, KNN and PNN are 0.111, 0.045 and 0.101, respectively. Using the three features selected by the modified neighborhood rough set (Table 4.4), the mean classification errors produced by SVM, KNN and PNN are reduced to 0.015, 0.043 and 0.083, respectively. The corresponding standard deviations are also reduced. This shows the effectiveness of feature selection.

2) The effect of the neighborhood size The classification error for each feature subset selected with a certain neighborhood size is also provided in Table 4.5. It can be seen that different neighborhood sizes generate different feature subsets, and correspondingly different classification results. Overall, the classification errors are larger than the results of modified neighborhood rough set.

It can be seen that among the three δ values, $\delta = 0.0013$ gives smallest errors (both mean and standard deviation) no matter which classifier is used. This is because $\delta = 0.0013$ is the mean noise level of 54 features, so it represents the noise level better than the minimum ($\delta = 0.0007$) and the maximum ($\delta = 0.0021$) ones. This supports the idea that the closer the neighborhood size is to the actual noise level, the higher the classification accuracy. This, on the other hand, proves the feasibility of Equation (4.18) as a neighborhood size approximation calculation formula. Even though $\delta = 0.0013$ generates the lowest mean errors among all three δ values (that is 0.028, 0.059 and 0.099 for SVM, KNN and PNN, respectively), it is worse than the results generated by the modified neighborhood rough set (that is 0.015, 0.083 and 0.043 for SVM, KNN and PNN). This is because the modified neighborhood rough set uses different neighborhood sizes for different features according to the noise levels these features suffer.

As discussed in Section 4.2.1, when the neighborhood size is smaller than its noise level, the feature's significance may be overestimated. Thus the feature selection process may stop early and miss some good features. For example in Table 4.4, when $\delta = 0.0007$, only feature No. 37 is selected, meaning that its approximation quality reaches the highest value (i.e. 1) and therefore feature selection process stops and no other features are selected. However, Table 4.2 shows that the neighborhood size should be 0.0013, so feature No. 37 is overestimated when $\delta = 0.0007$. When $\delta = 0.0013$, the approximation quality of feature No. 37 is not 1 anymore and additional features are added to increase the approximation quality, as can also be seen in Table 4.4. On the other hand, if the neighborhood size used is larger than the value it should be, the feature's significance may be underestimated. This, in turn, results in missing good features or adding bad features. For example, when $\delta = 0.0021$, most features are underestimated, the approximation quality of feature No. 37 is not the highest anymore so it is not selected. Instead, three features No. 17, No. 35 and No. 3 are selected, which generates poor classification results as shown in Table 4.5.

From the above discussion, it can be concluded that the modified neighborhood rough

set performs better than the original neighborhood rough set, because it considers the physical generation of noise levels and includes this information in the choice of neighborhood size.

4.5 Summary

This chapter studies feature selection based on neighborhood rough set. Neighborhood size is a key factor in neighborhood rough set. First the effect of neighborhood size is analyzed, based on which the problem of using the original neighborhood rough for feature selection is discussed. That is, features are likely to be wrongly estimated if a constant neighborhood size is used for all features, when features are obtained from different sources. To overcome this problem, the original neighborhood rough set model is modified. The modified neighborhood rough sets considers the physical meaning of neighborhood size, i.e. neighborhood size is the noise level a feature encounters. Thus in the modified neighborhood rough set model, each features is associated with a neighborhood size that stands for its noise level.

A feature selection method based on the modified neighborhood rough set is proposed and applied to the diagnosis of pumps. Fourier spectrum are used for feature extraction. Neighborhood sizes for these features are calculated by checking the Fourier spectrum from which these features are obtained. Results show that features selected by the modified neighborhood rough set achieves lower classification errors than do the raw features and features selected by the original neighborhood rough set. The above statement is supported by the fault classification results of all three classifiers (i.e. probabilistic neural network, K-nearest neighbor and support vector machine).

The disadvantage of the proposed method is that in the proposed method, the noise level of each feature needs to be estimated, which requires the collection of a large set of samples.

Chapter 5

Signal-Based Fault Identification

As discussed in Section 1.1.1, fault detection and isolation (FDI) is the first step in fault diagnosis, through which the fault type is determined. After that, fault identification is needed. Fault identification refers to the determination of the severity of a fault type. It plays an important role in maintenance scheduling. Different from FDI, fault identification has a unique characteristic; that is, the fault severity level has inherent ordinal information among different levels while fault type doesn't. For example, "a severe fault" is worse than "a moderate fault", and even worse than "a slight fault". So the severity level is an ordinal variable (see Section 2.3). This makes the fault identification more complicated than FDI [73].

Having ordinal information is an important characteristic of fault severity levels. Thus keeping the ordinal information is a necessary requirement for fault identification. In this chapter, signal-based fault identification is studied. Machine-learning-based fault identification will be studied in Chapter 6.

Signal-based fault identification aims to generate an indicator that monotonically varies with the fault progression. With such an indicator, the fault severity can be tracked by monitoring the value of this indicator. Such an indicator is, however, usually not easy to extract, especially for complex systems. In this thesis, fault diagnosis is studied using vibration signals. Several vibration sensors at different directions/locations are often used to collect vibration signals. Different sensors may provide health information from different perspectives. The indicator representing fault progression needs to make use of the information from different sensors. Thus, how to efficiently extract an indicator from two or more sensors becomes a key research issue.

In this chapter, two methods for generating such an indicator are proposed. The rest of the chapter is organized as follows. Section 5.1 summarizes two ways of integrating information from multiple sensors. Following the two ways, two indicator generation methods are proposed in Section 5.2 and Section 5.3, respectively. The two methods are applied to the identification of damage levels of impeller vane tailing edge and impeller vane leading edge, respectively. Finally, a summary is given in Section 5.4. Results in this chapter have

been published in [44], [102] and [126]. [1]

5.1 Background

In order to extract information from more than one sensors, researchers have developed several techniques. For example, full spectrum [8, 36] overcomes the limitation of conventional Fourier spectrum which deals with one-dimensional signals only. Full spectrum handles two-dimensional signals measured from two orthogonal sensors. In another word, conventional Fourier spectrum describes a one-dimensional vibration motion and full spectrum describes a two-dimensional (i.e. planar) vibration motion. Detailed descriptions on conventional Fourier spectrum and full spectrum are given in Section 2.1. Full spectrum reveals, not only the amplitude of a frequency component (as does the half spectrum), but also the directivity of a frequency component with respect to the planar rotational direction. This gives full spectrum great potential in condition monitoring of rotating machinery [37, 127]. Patel and Darpe [128] used full spectrum to detect the crack depth of a rotor and found that the positive 2X (twice the rotor rotation speed) frequency component becomes stronger with the increase of crack depth.

Besides full spectrum, statistical-based methods such as principal component analysis (PCA) and independent component analysis (ICA) are also widely used for multi-dimensional signals. Zhang et al. [129] utilized principal component (PC) representations of features from two sensors to monitor a double-suction pump. Cempel [130] applied singular value decomposition (SVD) to a set of features, and then proposed an indicator based on the singular values to track the health condition of a diesel engine.

The reported techniques of integrating information from different sensors fall into two categories. The first category (e.g. full spectrum and Holospectrum [34]) regards the signals from different sensors as a multi-dimensional signal, and uses signal processing techniques that are capable of handling multi-dimensional signals directly to analyze this multi-dimensional signal. The work by Patel and Darpe [128] using full spectrum directly on a two-dimensional signal (i.e. two signals measured from two orthogonal sensors, respectively) belongs to this category. This way of integrating information from different sensors is often used when the physical pattern due to a fault is known, and then a proper signal processing technique is applied to capture this pattern.

The second category regards signals from different sensors as a set of one-dimensional signals, applies signal processing techniques to the signal from each individual sensor or some sensors together for feature extraction and then combines features from all sensors. The work by Cempel [130] applying SVD to features calculated from each individual sen-

[1] Versions of this chapter have been published in "Xiaomin Zhao, Tejas H Patel and Ming J Zuo, 2011. Mechanical Systems and Signal Processing, 27:712-728.", "Xiaomin Zhao, Ming J Zuo and Tejas H Patel, 2012. Measurement Science and Technology. 23:1-11.", and "Xiaomin Zhao, Ming J. Zuo and R. Moghaddass, 2012. Book Chapter, Diagnostics and Prognostics of Engineering Systems: Methods and Techniques, IGI Global."

sor belongs to this group. However, when using this way to generate an indicator for fault levels, the selection of sensitive features is an issue to be addressed. Because not all features have positive contributions to the indicator generation, especially considering the requirement that the indicator needs to show a monotonic trend with the fault severity level.

Following the two ways of utilizing information from two or more sensors, two indicator generation methods are proposed in the two sections next, addressing the issues listed above.

5.2 Method I: Fault Identification Using Multivariate EMD and Full Spectrum (for two sensors only)

The idea of this method is to integrate information from two sensors by processing signals from two sensors together using full spectrum. Full spectrum is capable of revealing the directivity and amplitude of each spectral component in a planar vibration motion. However, not all the spectral components are sensitive to faults. Selecting the sensitive (fault-affected) spectral components can help diagnose fault levels more efficiently. Fixed band pass filtering can be used to choose spectral components in a certain frequency range. However, prior knowledge on the sensitive frequency range is required before processing the vibration data. Moreover, the noise contamination inevitably occurred during vibration data measurements makes the fixed band pass filtering less efficient. If noise is mixed with the interested frequency range of a true signal, the fixed band pass filtering process will be ineffective. To overcome this, empirical mode decomposition (EMD) can be used as an adaptive filter [131].

EMD decomposes a raw signal into a set of complete and almost orthogonal components called intrinsic mode functions (IMFs). IMFs represent the natural oscillatory modes embedded in the raw signal. Each IMF covers a certain frequency range. The IMFs work as the basis functions which are determined by the raw signal rather than by pre-determined functions. EMD has been widely used for fault diagnosis of rotating machinery [79, 132]. However, standard EMD has the limitation in that it works only for single real-valued signals. When dealing with data from multiple sensors, standard EMD needs to decompose signals measured from each individual sensor separately. However, because of the local and self-adaptive nature of the standard EMD, the decomposition results of signals from multiple sources may not match in either the number of IMFs or the frequency content of an IMF number [89], as described in Section 2.2.1.

To ensure proper decomposition of signals from multiple sources, standard EMD has recently been extended to the multivariate versions of EMD including those suitable for the bivariate signals [90], trivariate signals [91] and multivariate signals [9]. The mathematical description of multivariate EMD is given in Section 2.2.2.

It is worth mentioning that when comparing signals from different health conditions, the signals from not only all sensors but also all different health conditions need to be com-

bined together into a multi-dimensional signal before being decomposed by multivariate EMD. Otherwise, the unmatched problem would occur among signals from different health conditions. Therefore, even when only two sensors are used, a multi-dimensional signal needs to be generated and decomposed by multivariate EMD. Using full spectrum and multivariate EMD, an indicator generation method (method I) is proposed below.

5.2.1 Method Description

For rotating machinery, full spectrum provides information on both energy (i.e. vibration amplitude) and the directivity of vibration. Depending on the input signal, full spectrum could reveal whether the vibration is in the direction of rotation (i.e. forward) or in the opposite direction of rotation (i.e. backward). When a fault occurs, the energy and the directivity of vibration might change. As not all spectral components (frequency components) are sensitive to the fault, picking up the most sensitive (fault-affected) spectral component is important.

Empirical mode decomposition can be used as an adaptive filter to select the most sensitive frequency component [131]. An IMF represents a simple oscillatory mode and serves as a filter. Moreover, multivariate EMD, the multivariate extension of EMD, is able to find common oscillatory modes within signals from multiple sensors. Thus a complicated rotation represented by two signals (e.g. x and y) can be decomposed into two sets of IMFs (i.e. imf_x and imf_y) using multivariate EMD. Each IMF in set imf_x (respectively set imf_y) represents a projection of a simpler rotation in the X direction (respectively Y direction). That is, a complicated rotation is decomposed into a set of simple rotations. In this way, the selection of the most sensitive spectral component (i.e. a simple rotation) can be achieved by selecting the most sensitive IMF.

To select the sensitive IMF, criteria based on the correlation coefficient between an IMF and its raw signal was used in [133, 134]. Note that the correlation coefficient reflects only linear relationship between two variables. To account for nonlinear relationship as well, mutual information [135] is employed and a selection criterion is proposed. The criterion takes into account two kinds of mutual information: (1) that between the n^{th} IMF and its raw signal, and (2) that between the n^{th} IMF of a signal with certain health condition and the n^{th} IMFs of signals with different health conditions. The following example explains how the criterion works.

Suppose that two signals, $x_0(t)$ and $y_0(t)$, collected from the X and Y directions under the normal operation, are denoted by a two-dimensional signal, $x_0(t)\overrightarrow{i} + y_0(t)\overrightarrow{j}$. Two other signals, $x_1(t)$ and $y_1(t)$, collected from the X and Y directions at a fault condition, are denoted by a two-dimensional signal, $x_1(t)\overrightarrow{i} + y_1(t)\overrightarrow{j}$. Let $cn_{x0}(t)$, $cn_{y0}(t)$, $cn_{x1}(t)$, and cn_{y1} be the n^{th} IMF of $x_0(t)$, $y_0(t)$, $x_1(t)$ and $y_1(t)$, respectively, obtained by multivariate EMD. The proposed criterion (Equation (5.2)) for selecting a sensitive IMF is described in Table 5.1.

Table 5.1: Proposed method for the selection of sensitive IMF

Step 1. Calculate the mutual information, a_n, between the n^{th} IMF of a two-dimensional normal signal, $cn_{x0}(t)\vec{i} + cn_{y0}(t)\vec{j}$, and the signal itself, $x_0(t)\vec{i} + y_0(t)\vec{j}$.
Step 2. Calculate the mutual information, b_n, between the n^{th} IMF of a two-dimensional fault signal, $cn_{x_1}(t)\vec{i} + cn_{y_1}(t)\vec{j}$, and the signal itself, $x_1(t)\vec{i} + y_1(t)\vec{j}$.
Step 3. Calculate the mutual information, e_n, between $cn_{x1}(t)\vec{i} + cn_{y1}(t)\vec{j}$ and $cn_{x0}(t)\vec{i} + cn_{y0}(t)\vec{j}$.
Step 4. Calculate the sensitivity factor, λ_n, for the n^{th} IMF using

$$\lambda_n = \frac{a_n + b_n}{2} - e_n. \tag{5.1}$$

In this equation, the first part, $(a_n + b_n)/2$, represents the average mutual information between the n^{th} IMFs and their raw signals; the second part, e_n, represents the mutual information between the n^{th} IMF of the normal signal and the n^{th} IMF of the signal under different health conditions (i.e. the fault signal in this example). To ensure that an IMF is informative enough to represent the original signal, the first part is expected to be high; to enable the easy detection of the fault, the second part is expected to be low. Therefore, the sensitive IMF is expected to have a high value of λ_n.
Step 5. Find the most sensitive IMF (the s^{th} IMF),

$$\lambda_s = \max_n \lambda_n. \tag{5.2}$$

The IMF having the highest value of sensitivity factor is the most sensitive IMF.

After the sensitive IMF is selected, the full spectrum of this IMF is obtained. A full spectral indicator reflecting the characteristic of planar vibration motion is then extracted for condition monitoring. This indicator to use is problem specific. As will be illustrated later for the problem of impeller vane trailing edge damage in a pump, the ratio between energy of backward components and energy of forward components is used as the indicator. This was based on the characteristic of the velocity field in the pump, as will be described in Section 5.2.3. A flowchart of the proposed method is given in Figure 5.1.

5.2.2 Application to Simulation Data

To illustrate how the proposed indicator generation method works, first this method is applied to simulation data. In the fault diagnosis of rotating machinery, the change of a amplitude at a typical frequency [136] and the appearance of a new frequency [137] are two phenomena that commonly occur when a fault occurs. For this reason, simulation signals were constructed considering these two effects to illustrate and test the proposed method.

Four signals, namely x_0, y_0, x_1 and y_1, are constructed and shown in Equations. (5.3)-(5.6). The x_0, y_0 and x_1, y_1 are two sets of vibration signals measured from the X and Y orthogonal directions for "normal" and "fault" conditions respectively. The vibration

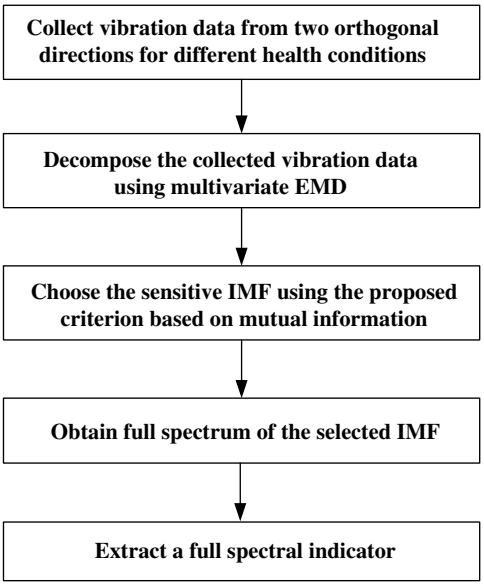


Figure 5.1: Flowchart of indicator generation method I

signals for normal condition have of two parts: the first part is a counter-clockwise rotation at a frequency of 20 Hz with amplitude of 0.5; the second part is an counter-clockwise rotation at 100 Hz with amplitude of 1. Vibration signals for the fault condition consist of four parts: the first part is the same as the first part of the normal signal; the second part is an counter-clockwise rotation at 50 Hz with amplitude of 0.2; the third part is a clockwise rotation at 50 Hz with amplitude of 1; the fourth part is the same as the second part of the normal signal but with an amplitude of 0.9.

$$x_0 = 0.5\cos(40\pi t) + \cos(200\pi t) \tag{5.3}$$

$$y_0 = 0.5\sin(40\pi t) + \sin(200\pi t) \tag{5.4}$$

$$x_1 = 0.5\cos(40\pi t) + 0.2\cos(100\pi t) + \cos(-100\pi t) + 0.9\cos(200\pi t) \tag{5.5}$$

$$y_1 = 0.5\sin(40\pi t) + 0.2\sin(100\pi t) + \sin(-100\pi t) + 0.9\sin(200\pi t) \tag{5.6}$$

First, standard EMD is applied to x_0, y_0, x_1 and y_1, separately; the results are shown in Figure 5.2. The top row in each column shows the raw signals, the bottom row in each column shows the residuals, and the middle rows are IMFs. There are two IMFs for each of x_0 and y_0, while three for each of x_1 and y_1. Thus standard EMD results in an unequal number of IMFs [89], making it impossible to compare normal and fault signals on the 3^{rd} IMF in Figure 5.2. Moreover, the frequency content of the 2^{nd} IMFs of x_0 and y_0 (i.e. 20 Hz) is different from that of the 2^{nd} IMFs of x_1 and y_1 (i.e. 50 Hz). Therefore, the 2^{nd} IMFs of the normal and fault signals can not be compared directly either. From this simple simulation example, it can be seen that it would be difficult to capture fault information from the same IMF number for real vibration signals, because real vibration signals are much more complicated than those simulated with Equations (5.3) - (5.6).

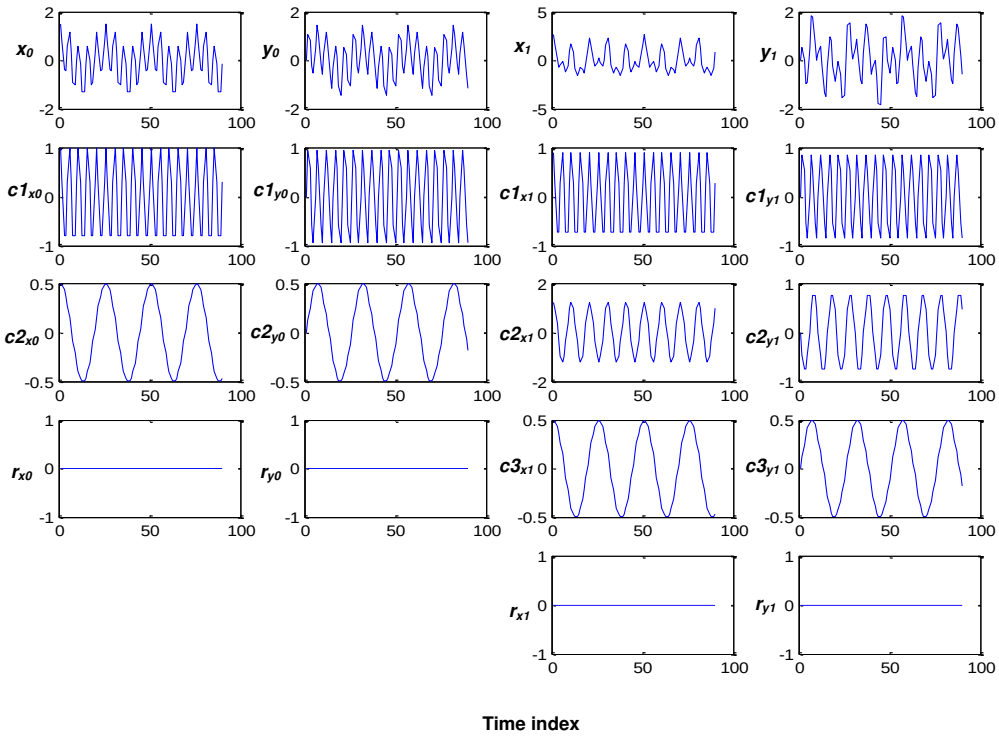


Figure 5.2: Decomposition results of simulated data with standard EMD

As stated in Section 2.2.2, the multivariate versions of EMD can ensure the proper decomposition of signals from multiple sensors. In this example, two signals are measured at two orthogonal directions for each health condition. If bivariate EMD is applied to normal signals (i.e. $x_0(t)\vec{i} + y_0(t)\vec{j}$) and fault signals (i.e. $x_1(t)\vec{i} + y_1(t)\vec{j}$) separately, the results will be the same as those shown in Figure 5.2. The problem of unmatched decomposition in the number of IMFs and the frequency content of an IMF number still exists. The problem of unmatched decomposition could be avoided by decomposing all the signals together. First, all the four raw signals (i.e. x_0, y_0, x_1, and y_1) are combined to form a four-dimensional signal, $x_0(t)\vec{i} + y_0(t)\vec{j} + x_1(t)\vec{k} + y_1(t)\vec{q}$. Multivariate EMD is then conducted to decompose the four-dimensional signal. A set of four-dimensional IMFs, $cn_{x0}(t)\vec{i} + cn_{y0}(t)\vec{j} + cn_{x1}(t)\vec{k} + cn_{y1}(t)\vec{q}$, are obtained, where cn_{x0}, cn_{y0}, cn_{x1}, and cn_{y1} represent the n^{th} IMF of x_0, y_0, x_1, and y_1, respectively. Results are shown in Figure 5.3. The top row shows the raw signals. The bottom row shows the residuals. The 1^{st}, 2^{nd} and 3^{rd} IMFs are shown in the second, third and fourth row for frequencies 100 Hz, 50 Hz and 20 Hz, respectively. The 2^{nd} IMFs of x_0 and y_0 have no element for 50 Hz, therefore their 2^{nd} IMFs are almost 0. Now each row has a common frequency content, which makes the IMFs of normal and fault signals comparable. It can also be seen that the raw complicated motion is decomposed into three simple motions, making further inference regarding the fault-sensitive IMF easier.

The sensitivity factor based on mutual information is then calculated to find the most

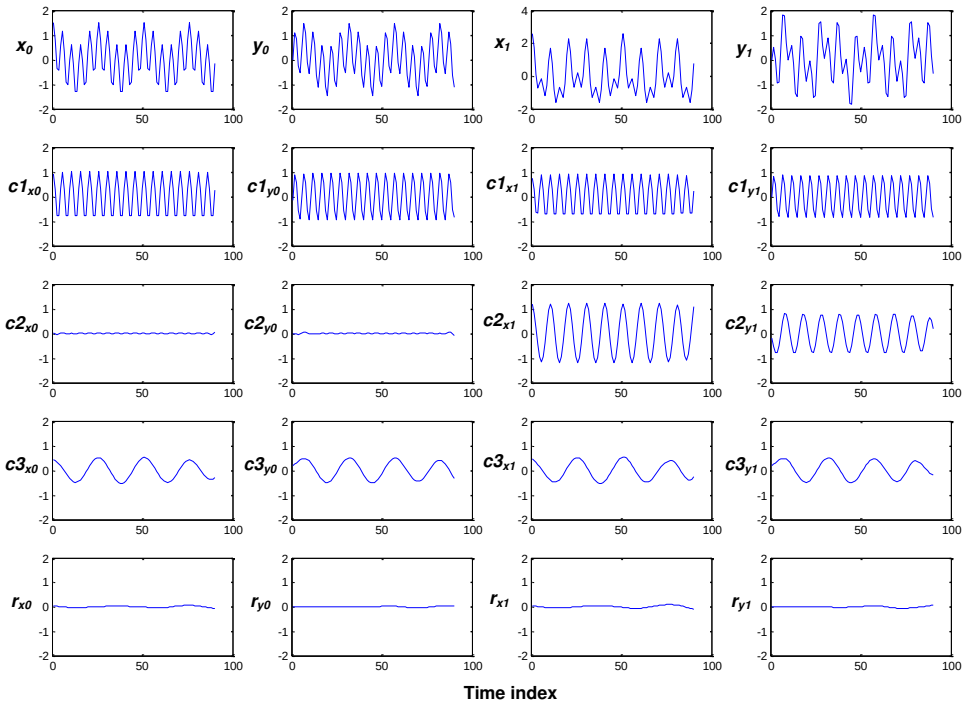


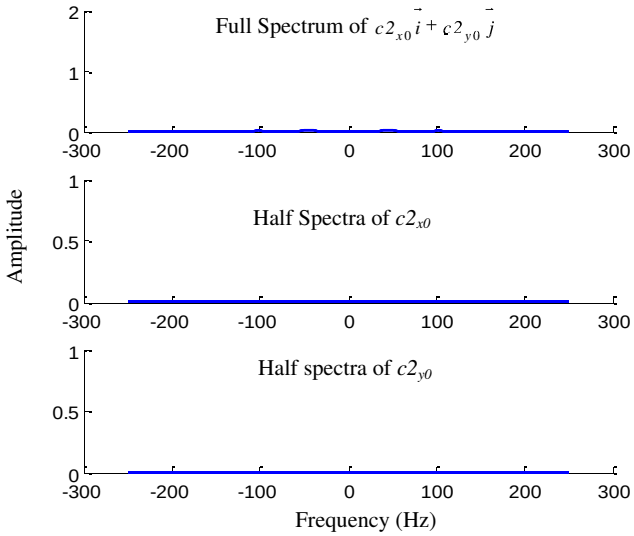
Figure 5.3: Decomposition results of simulated data with multivariate EMD

Table 5.2: Sensitivity factor of each IMF (simulation data)

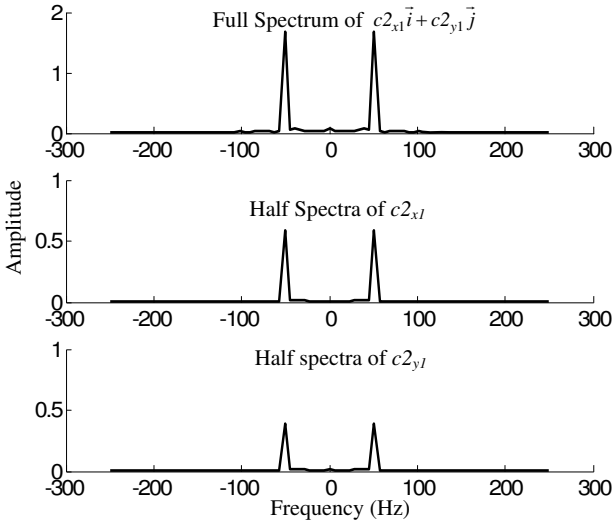
IMF number	Sensitivity factor
1	-1.53
2	**0.51**
3	-2.51

sensitive IMF which can clearly distinguish the fault condition. Table 5.2 shows the details. The 2^{nd} IMF gives the highest value and therefore is chosen as the sensitive IMF. This result is consistent with our intuition because the 2^{nd} IMF, as can be seen from Figure 5.3, shows a clear difference between normal and fault signals.

To express the fault information, full spectrum is conducted on the sensitive IMF (i.e. the 2^{nd} IMF). The 2^{nd} IMFs of x_m and y_m (where $m = 0, 1$) are used as the direct part and the quadrature part, respectively. Thus the forward direction corresponds to the counter-clockwise direction. The full spectra of $c2_{x0}\overrightarrow{i} + c2_{y0}\overrightarrow{j}$ (normal) and $c2_{x1}\overrightarrow{i} + c2_{y1}\overrightarrow{j}$ (fault) respectively are shown on the top of Figure 5.4(a) and Figure 5.4(b). The 2^{nd} IMFs of normal signals are almost zero, so the spectra in the three rows are almost zero for the normal case (Figure 5.4(a)). To facilitate comparison, the half spectra are also shown in the middle row ($c2_{x0}$ and $c2_{x1}$) and the bottom row ($c2_{y0}$ and $c2_{y1}$) of Figure 5.4. The characteristic of full spectrum can be observed by checking the fault case (Figure 5.4(b)). Full spectrum clearly indicates the backward (i.e. clockwise) rotation with amplitude of 1 and the forward



(a) normal condition



(b) fault condition

Figure 5.4: Full spectra and half spectra of the 2^{nd} IMFs of simulated data

(i.e. counter-clockwise) rotation with amplitude of 0.2, as defined in Equations (5.5)-(5.6). The half spectrum doesn't, however, take rotation directivity into account. Therefore, the term $c2_{x1}, 0.2\cos(100\pi t) + \cos(\pi t)$, is simply regarded as $1.2\cos(100\pi t)$ and the amplitude of 1.2 is equally split to the amplitudes at 50 Hz and -50 Hz (i.e. 0.6). The same reasoning applies to $c2_{y1}$, $0.2\sin(100\pi t) + \sin(-100\pi t)$, at -50 Hz and 50 Hz with amplitudes both equal to 0.4. Compared to the half spectrum which gives amplitudes of 0.6 for $c2_{x1}$ and 0.4 for $c2_{y1}$, the full spectrum with amplitude of 1 at -50 Hz and amplitude of 0.2 at 50 Hz reveals the fault information more clearly with regard to both directivity and energy.

The simulated data illustrates the concept of the proposed method I. It shows that the proposed method can help find the fault-sensitive spectral component, and reveal the characteristic of a planar vibration. Next, the proposed method I is applied to the identification of damage levels of impeller vane trailing edge.

5.2.3 Application to Identification of Damage Levels of Impeller Vane Trailing Edge

In this section, the proposed method is applied to the identification of damage levels of impeller vane trailing edge. As introduced in Section 3.1.3, four damage levels are considered: no damage (level 0), slight damage (level 1), moderate damage (level 2) and severe damage (level 3). Vibration data were collected from three accelerometers under each of the four damage levels. Details on experiments refer to Section 3.1.3. In this section, the data collected at the flow rate of 100% BEPQ only is considered.

5.2.3.1 Analysis on Flow Patterns in Pumps

In pumps, hydrodynamic forces are often the major sources of vibration [17]. The velocity distribution at the impeller outlet is non-uniform as a result of the finite vane thickness, the blockage caused by the boundary layers, and possible flow separation. The cutwater (See Figure 3.2 on the structure of a pump) is thus approached by an unsteady flow inducing alternating hydrodynamic forces on the cutwater. These hydrodynamic forces cause the vibration of the pump casing. It is believed that the change of geometry at impeller vane tailing edge affects the velocity and the pressure distributions inside the pump [17]. Consequently, the hydrodynamic forces and therefore the vibrations are affected. To better understand the variation in the flow field, numerical simulations using the computational fluid dynamics (CFD) approach [4] are carried out.

ANSYS CFX 12 [138] was used to conduct steady-state simulations. The pump inlet pipe and the pump outlet pipe were extended to allow possible inlet recirculation and the elliptic influence of the flow (see Figure 5.5). The boundary conditions were set as follows.

- Inlet: the total pressure was applied in the axial direction.
- Outlet: mass flow was imposed.
- Wall: no-slip wall was adopted.

A reference frame refers to a coordinate system within which the position, orientation, and other properties of objects in it are measured. Due to the change between the reference frame of the rotating impeller and the reference frame of the static volute liner, the interaction between the impeller and the liner was simulated with the Frozen-Rotor interface model [138]. In this model, simulation is preformed for a specific relative position of the machine components, and then this relative position is changed step by step in a

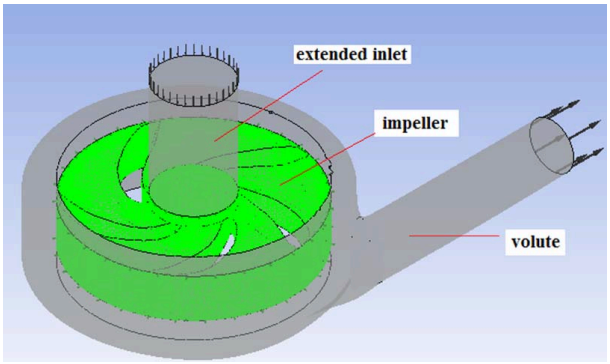


Figure 5.5: Domains for pump CFD simulation

Table 5.3: Parameter settings in CFX 12 for pump simulation [4]

Item	Settings in CFX 12
pumping medium	water
speed	2400 RPM
Domain of simulation	Extended inlet + impeller + volute liner
Impeller grid	Unstructured 355471 nodes
Volute liner grid	Unstructured 296788 nodes
Extended inlet duct grid	Unstructured 120853 nodes
Inlet	Total Pressure = 45043 Pa
Interface inlet pipe/ impeller	Frozen rotor
Interface impeller / volute liner	Frozen rotor
Outlet	455 usgpm (100% BEPQ)
Turbulence model	K-epsilon
Discretization	Second order
Maximum residual convergence criteria	10^{-4} (RMS)

quasi-steady calculation. The parameters for the pump simulation are summarized in Table 5.3.

Figure 5.6 shows the relative velocity fields near the cutwater under the four health conditions obtained from our CFD simulations. The direction of the velocity at a certain location is indicated by the direction of the arrow. The magnitude of the velocity at a certain location is indicated by the length of the arrow. The direction of impeller rotation is clockwise. The flow goes out from the impeller into the volute, experiences a certain degree rotation in the volute, and comes out to the outlet. Thus near the cutwater area, part of the flow is directed into the volute (we call it forward direction as it is in the direction of impeller rotation), and the rest is directed to the outlet (we call it backward direction as it is not in the direction of impeller rotation). At the no damage case, the velocity vectors are well directed towards the outlet and the volute. This directivity is distorted when damage occurs at the vane trailing edge. The degree of distortion increases as the level of damage increases. At the severe case, the flow has a weak trend in the forward direction compared

to the backward direction.

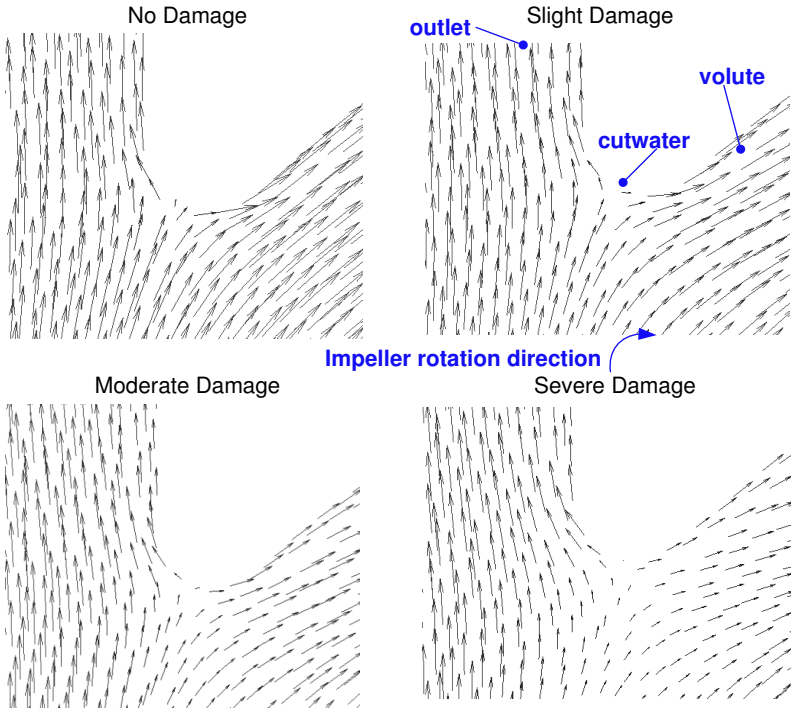


Figure 5.6: Zoomed view of the relative velocity fields near the cutwater area

The vibration sensor labeled as A1 in Figure 3.5 is expected to capture the change of flow directivity indicated in Figure 5.6, because it is located close to the cutwater. Data from sensor A1 in the X direction (donated by channel A1-X) and the Y direction (donated by channel A1-Y) are analyzed because they are in the plane of the main flow (shown in Figure 3.5). Based on the flow patterns shown in Figure 5.6, the forward motion becomes weak compared to the backward motion. Thus, the ratio between the backward whirling energy and the forward whirling energy is expected to increase as the damage level increases. A full spectral indicator in the form of an energy ratio is defined using Equation (5.7) to represent this. In Equation (5.7), $A(f)$ represents the value of amplitude at frequency f in a full spectrum.

$$Er = \frac{\sqrt{\sum_{f<0} A(f)^2}}{\sqrt{\sum_{f>0} A(f)^2}} \tag{5.7}$$

5.2.3.2 Selecting Sensitive Frequency Component

Let x_0, x_1, x_2, and x_3 (respectively y_0, y_1, y_2, and y_3) denote the signal measured by channel A1-X (respectively channel A1-Y) for the no damage, slight damage, moderate damage, and severe damage cases, respectively. Figure 5.6 shows that the impeller rotates clockwise. To make the forward direction the impeller's rotation direction, signals from channel A1-X

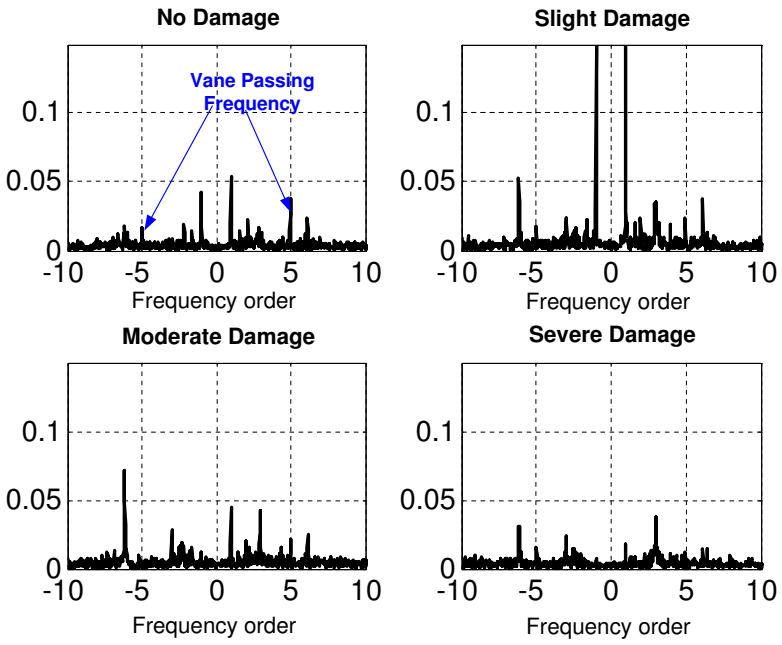


Figure 5.7: Full spectra of raw data for different health conditions

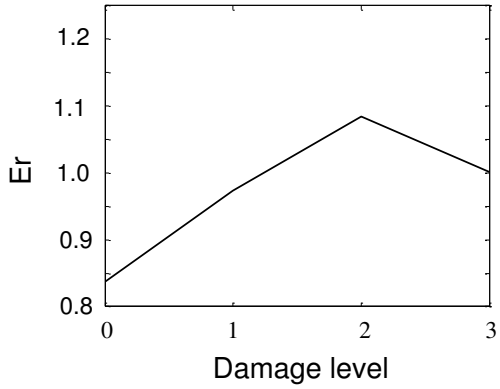
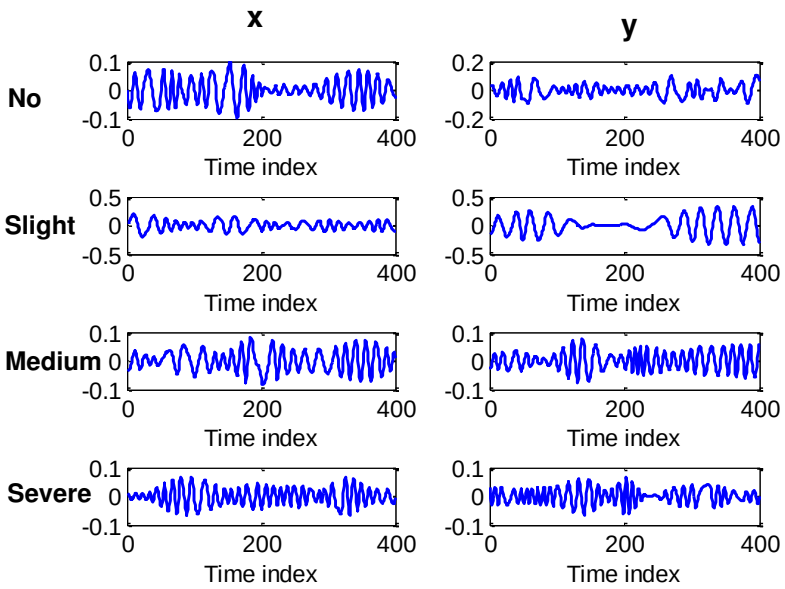


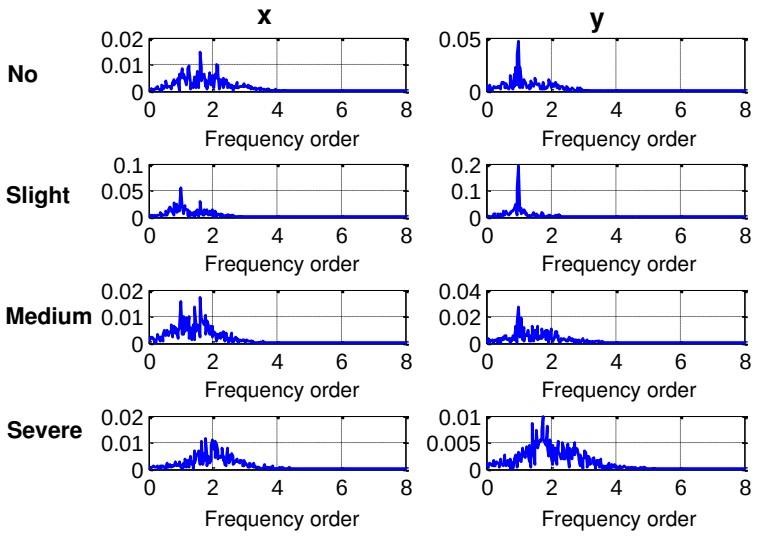
Figure 5.8: Energy ratio of raw signal Versus damage level

and channel A1-Y are used as the quadrature part and the direct part respectively in the calculation of full spectrum (See Section 2.1.2). Full spectra of the raw data are shown in Figure 5.7 under each of the four health conditions and the change in the energy ratio with the damage level is shown in Figure 5.8. It can be seen that the spectra (Figure 5.7) consist of many spectral components. Though the spectra are different for different health conditions, there is no monotonic trend observed in Figure 5.7. Empirical mode decomposition which is capable of filtering spectral components is thus needed.

Although standard EMD could generate the same number (ten in this application) of IMFs for signals from different sources, there still remains an unmatched property problem.



(a) in the time domain

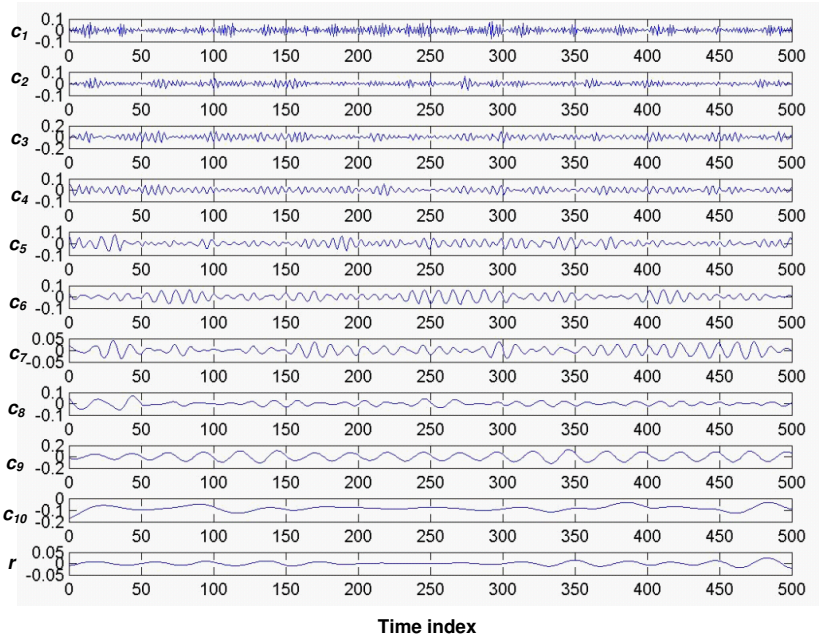


(b) in the frequency domain

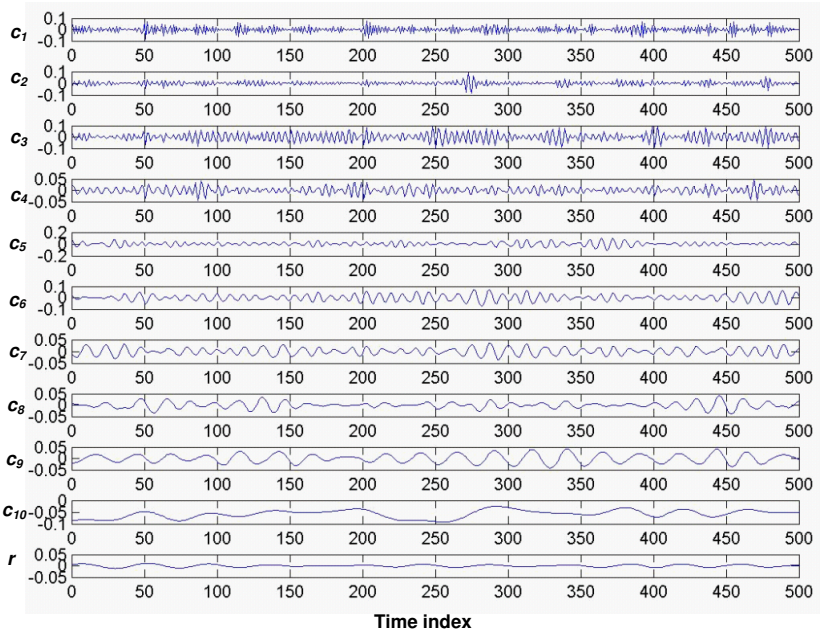
Figure 5.9: The 4^{th} IMFs of different signals (standard EMD)

Figure 5.9 shows the 4^{th} IMF as an example. The frequency contents of the 4^{th} IMFs for different damage levels don't match. The frequency centers of the 4^{th} IMFs for x_0, x_1, and x_2 signals are around 1.5X (i.e. 1.5 times the pump rotation frequency); however, the frequency center of the 4^{th} IMF of x_3 signal is around 2X. This inconsistency also exists in signals measured by A1-Y. At the same damage level, the 4^{th} IMFs of signals measured in different directions don't match either. For example, in the no damage case, the frequency

center of the 4^{th} IMF of x_0 signal is 1.5X, whereas the frequency center of the 4^{th} IMF of y_0 signal is 1X. The inconsistency exists in the slight and moderate damage cases too. To address the unmatched property of IMFs from different sources, multivariate EMD is applied.



(a)



(b)

Figure 5.10: The decomposition results for y_0 (a) and $y_3(b)$ using multivariate EMD

Eight raw signals (i.e. x_0, y_0, x_1, y_1, x_2, y_2, x_3 and y_3) are combined into an eight-

dimensional signal and decomposed together using multivariate EMD algorithm. Figure 5.10 shows the decomposition results for y_0 and y_3 signals (the results for data associated with other health conditions are not presented here). The first ten IMFs are obtained for each signal; the other IMFs have small amplitudes and thus are put into the residual (r). It can be seen from Figure 5.10 that the corresponding IMFs of y_0 and y_3 signals carry the same frequency content. The criterion proposed in Section 5.2.1 is used to evaluate the significance of each of the ten IMFs. The significance factor of the n^{th} IMF is calculated using the averaged mutual information between the n^{th} IMF and its raw signal subtracted by the averaged mutual information between the n^{th} IMF of one health condition and that of other health conditions. Table 5.4 shows the results. The 4^{th} IMF has the highest value and is chosen as the sensitive IMF.

Table 5.4: Sensitivity factor of each IMF (pump data)

IMF number	Sensitivity factor
1	0.09
2	0.16
3	0.19
4	**0.24**
5	0.12
6	0.21
7	0.05
8	-0.11
9	-0.35
10	-0.77

5.2.3.3 Indicator Generation

The 4^{th} IMF of x_i signal and the 4^{th} IMF of y_i signal are used to obtain the full spectrum for each of the four health states (Figure 5.11). At the no damage condition, the forward frequency components are dominant. As the health condition worsens, the strength of the forward components compares unfavorably with that of the backward components. This is further examined by plotting the energy ratio (Equation (5.7)) values in Figure 5.12. It can be seen that the ratio between backward components and forward components monotonically increases as the damage level increases. This agrees with the observation from Figure 5.6. Furthermore, as described at the beginning of Section 5.2.3, the local flow field near the cutwater is subjected to significant variations each time an impeller vane passes it, so the vane passing frequency should be the characteristic frequency. Figure 5.11 shows that the selected IMF is centered around the vane passing frequency (5X). Therefore, this is consistent with our expectation.

In this section, the indicator generation method (Method I) is proposed. This method deals with signals measured from only two orthogonal sensors together. So the fault infor-

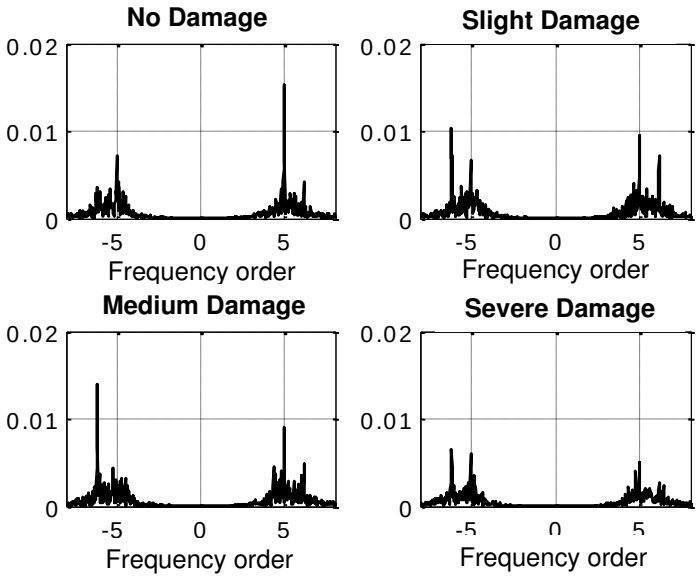


Figure 5.11: Frequency spectra of the 4^{th} IMFs of different signals (multivariate EMD)

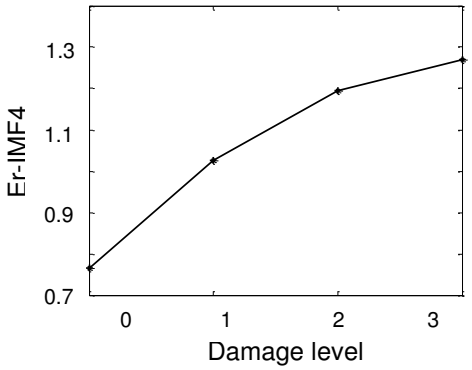


Figure 5.12: Energy ratio of the 4^{th} IMF versus damage level (multivariate EMD)

mation is extracted from a planar vibration. For the impeller vane trailing edge damage, the simulation results show that it affects mostly the main flow plane (a 2D plane). That is why this method works. The disadvantage of Method I is that it can not deal with signals from more than two sensors together. For some other fault type, e.g. impeller vane leading edge damage, the vibration in many locations and directions are affected [17]. Thus more than two sensors need to be involved in the indicator generation process. To do this, method II is proposed next.

5.3 Method II: Fault Identification Using Fuzzy Preference Based Rough Set and Principal Component Analysis

The idea of this method is to extract features first, and then combine features from all sensors together and output a single indicator. The first step (feature extraction) is achieved by signal processing techniques including those for one-dimensional signals (e.g. conventional Fourier spectrum) and for two-dimensional signals (e.g. full spectrum). In the second step, an indicator that has monotonic trend with the fault levels is generated. How to combine the health information (features) from all sensors into an indicator that represents the health condition (i.e. exhibit monotonic trend) is the key issue of this method. To address this issue, 1) a measure to select features exhibiting better monotonic relevance to the fault level; and 2) a strategy to combine the selected features are needed. Next, these two issues are discussed and an indicator generation method is proposed.

5.3.1 Method Description

The first issue is discussed first. As discussed in Chapter 1, there exist many measures for evaluating the significance of a feature in classifying different classes, such as the measures based on correlation coefficients [94], mutual information [53] and rough sets [95]. These measures reflect the relevance between a feature and the class label. The labels in classification problems are nominal variables. But the fault levels (e.g. slight fault, moderate fault, and severe fault) are ordinal variables, as stated in Section 2.3. The features selected for generating an indicator are expected to carry the ordinal information among the fault levels. Thus the evaluation of a feature should base on the ability of this feature in expressing the ordinal information. In another word, the measure evaluates the monotonic relation between the feature and the fault level. Most existing measures, however, do not consider this monotonic relation.

Rough set has been proved to be an effective tool in selecting important features for classification problems. Traditional rough sets consider only the equivalence relation which is suitable for nominal variables. In order to consider the preference relation (monotonic relation), Greco et al. [139] introduced dominance rough sets to measure the monotonic relation qualitatively. Hu et al. [93] extended dominance rough sets and proposed fuzzy preference rough sets which can reflect the monotonicity degree between two variables quantitatively. The mathematical descriptions on dominance rough sets and fuzzy preference rough sets are given in Section 2.5.3 and Section 2.5.4, respectively. In this chapter, the global fuzzy preference approximation quality (global FPAQ (Equation (2.50))) defined in fuzzy preference based rough set is used for the evaluation of monotonic relevance.

The second issue is how to combine the information in different features into one single variable. This process is also called feature fusion. Eigenvalue/Eigenvector analysis has been widely used as a feature fusion method in condition monitoring [140–142]. Pires

et al. [141] found a severity index for stator winding fault, and rotor broken bars were detected from the obtained eigenvalues. Turhan-Sayan [142] used PCA for feature fusion and obtained a single indicator that can effectively represent the health condition of the electromagnetic target of concern.

Principal component analysis (PCA) is a simple eigenvector-based multivariate analysis method. It transforms a number of possibly correlated variables into a number of uncorrelated variables called principal components. Natke and Cempel [140] found that the non-zero eigenvalues, ordered by their magnitudes, can be regarded as fault ranking indices which measure the fault intensity. The first principal component corresponds to the largest eigenvalue, and therefore contains most information on damage conditions. In this chapter, different fault levels of the same fault type are considered. The first principal component is used as a single indicator representing the fault levels.

In this section, a method for generating an indicator of fault levels using both fuzzy preference based rough sets and PCA is proposed. PCA is used to combine information of a set of features into one variable. The global FPAQ (Equation (2.50)) is used to evaluate how much useful health information (monotonic relevance) a variable contributes to the determination of the fault levels. The variable that has the highest value of global FPAQ is the indicator. The process of searching this indicator is described below. Let D be the set of fault levels, C be the set of features, S be the set of selected features, I_s be the indicator generated by S, and e_s represent the monotonic relevance between the indicator and the fault level. The steps are detailed in Table 5.5.

Table 5.5: Indicator generation method II

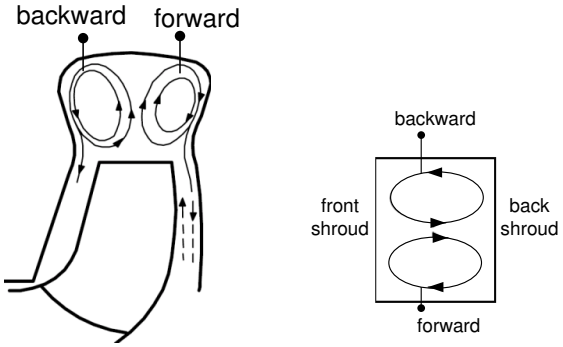
Step 1. Extract features using proper signal processing techniques. The features from all sensors are put together and stored in set $C = [a_1, a_2, \ldots, a_m]$, where m is the total number of features from all sensors.
Step 2. Employ Equation (2.50) to evaluate the monotonic relevance between each feature and D (fault levels). Results are saved as $E = [e_1, e_2, \ldots, e_m]$, where e_i is the monotonic relevance between feature a_i and D.
Step 3. Set $S = \emptyset$. Find a_k such that $e_k = \max_i(e_i)$, put a_k into S (i.e. $S = S \bigcup a_k$), delete it from C (i.e.$C = C - a_k$) and let $I_s = a_k$. and $e_s = e_k$.
Step 4. For each feature, a_i, in C, generate a temporary feature set $T_i = S \bigcup a_i$, where $i = 1, 2, \ldots, p$ and p is the total number of features in C.
Step 5. Compute I_{tempi} (I_{tempi} is the first principal component of T_i).
Step 6. Calculate e_{tempi}, the monotonic relevance between I_{tempi} and D using Equation (2.50).
Step 7. Find T_k that corresponds to the highest monotonic relevance, i.e. $e_{tempk} = \max_i(e_{tempi})$.
Step 8. If $e_{tempk} > e_s$, then let $S = T_k$, $C = C - a_k$, $e_s = e_{tempk}$, $I_s = I_{tempi}$ and go to Step 4; otherwise, go to Step 9.
Step 9. Output the indicator I_s.

The advantage of this method is that it includes information from all sensors. So it is useful for the fault type that affects vibration in many locations and directions. Next, this method is applied to the identification of damage levels of impeller vane leading edge.

5.3.2 Application to Identification of Damage Levels of Impeller Vane Leading Edge

The impeller vane leading edge damage causes more complicated flow inside the pump than the impeller vane trailing edge damage. Because the latter is located in the downstream of the pump flow, so mainly only the volute passage is affected. The former is located in the upstream of the pump flow, the flow patterns in the whole flow passage (including the impeller passage as well as the volute passage) are affected.

In the impeller passage, the flow is deflected by 90 degree from the axial entry to the radial outlet. As the flow progresses, a secondary flow (recirculation from the front shroud to the rear shroud) builds up owing to centrifugal forces as shown in Figure 5.13(b) [17]. The change of geometry at impeller leading edge, affects velocity, thus the centrifugal forces, and therefore the secondary flow in the volute passage. In the volute passage, secondary flows exist as shown in Figure 5.13(a) [17] due to the curvature of volute geometry. The flow pattern has the shape of a double vortex. It becomes increasingly asymmetrical with the growing non-uniformity of the impeller outflow. When there is damage on the impeller leading edge, the impeller outflow, and thus this double vortex in the volute passage is influenced.



(a) Volute passage (b) Impeller passage

Figure 5.13: Secondary flow in the pump flow field [17]

According to Ref. [17], change in flow patterns generates certain kind of hydraulic excitation forces, that can be sensed through the pump impeller (rotor) and pump casing vibrations. To capture this information, three tri-axial accelerometers, shown in Figure 3.5, were mounted on the system. These accelerometers are all affected by the main flow, and are individually sensitive to certain flow pattern. The tri-axial accelerometer A1 is located near the outlet, so it (specifically channel A1-Y) is expected to capture the information on the outlet of the main flow. The tri-axial accelerometer A2 is located on the top of the volute

casing which is closest to the volute passage, so the flow pattern in the volute passage is reflected through this acceleromter. Moreover, A2YZ plane (the plane where channel A2-Y and channel A2-Z are) is believed to be informative, as it is in this plane where secondary flow in volute passage (Figure 5.13(a)) occurs. The tri-axial accelerometer A3 is located on the bearing casing which is directly connected to the impeller rotor. So the secondary flow in the impeller passage can be captured by this accelerometer. Moreover, A3XY plane (the plane where channel A3-X and channel A3-Y are) is in the plane where the secondary flow in impeller passage (Figure 5.13(b)) occurs, so it is believed to be informative.

Four damage levels are considered in this section: no damage (level 0), slight damage (level 1), moderate damage (level 2) and severe damage (level 3). For details on experiments, please refer to Section 3.1.2.

In Method II, there are two steps: feature extraction and indicator generation. The two steps are conducted one by one in the following.

5.3.2.1 Feature Extraction by Half and Full Spectra

As shown in Chapter 4, conventional Fourier spectrum (half spectrum), more specifically, the amplitudes at pump rotating frequency (1X), its 2^{nd} harmonic (2X), and vane passing frequency (5X) carry pump health conditions. Ref. [95] also confirmed that the 2^{nd} harmonic of vane passing frequency (10X) carry useful information on pump conditions. So four features (i.e. the amplitudes at 1X, 2X, 5X and 10X) are extracted from each half spectrum of a vibration signal. Three tri-axial accelerometers produces nine channels. Each channel outputs one vibration signal. So there are 36 (4×9) features from half spectrum.

Section 5.2.3 shows that full spectrum captures the change of a planar flow pattern. The secondary flow (Figure 5.13(a) and 5.13(b)) is believed to contain information on pump health conditions. To capture this information, full spectrum, as stated in Section 2.1.2 is applied. As positive and negative halves are not the mirror images of each other, there are eight features (i.e. amplitude at 1X, 2X, 5X, 10X, -1X, -2X, -5X and -10X) calculated from each full spectrum. Each tri-axial accelerometer has three orthogonal directions (X, Y, Z), and produces three plane combinations (i.e. XY plane, YZ plane and XZ plane). As a result, nine plane combinations from three accelerometers generate 72 features. Therefore, totally a feature set consisting of 108 (72+36) features are generated. Table 5.6 lists the vibration signals and their corresponding ranges of feature No. in the feature set.

It is worth noting that in the pump experiments, the fluid flows clockwise (Figure 3.5), (i.e. A1-Y to A1-X direction). To ensure that the forward components are clockwise, signals from channels A1-Y and A1-X are used as the direct part and the quadrature part, respectively in the full spectrum calculation (Figure 2.2). Similar explanation applies to other tri-axial accelerometers. To make this clear, the signal sources are named, in which the first two letters stand for the tri-axial accelerometer, the third letter for the direct part and the fourth letter for the quadrature part. For example, A1XZ means that in the calculation

of a full spectrum, the direct part is the signal from channel A1-X and the quadrature part is the signal from channel A1-Y.

Table 5.6: Signal source and the corresponding range of feature No.

Signal source	Feature No.	Signal source	Feature No.
A1XZ	1-8	A1-X	73-76
A1YX	9-16	A1-Y	77-80
A1YZ	17-24	A1-Z	81-84
A2XY	25-32	A2-X	85-88
A2YZ	33-40	A2-Y	89-92
A2ZX	41-48	A2-Z	93-96
A3XY	49-56	A3-X	97-100
A3YZ	57-64	A3-Y	101-104
A3ZX	65-72	A3-Z	105-108

5.3.2.2 Indicator Generation

The monotonic relevances of each of the 108 features with the fault levels are evaluated using fuzzy preference based rough sets (see Section 2.5.4). Figure 5.14 shows the results. It can be seen that different features have different monotonic relevance with the fault levels. Now an indicator is generated based on the 108 features. As described in Table 5.5, feature selection and feature fusion are used alternately in the indicator generation process. Results are discussed below.

To reveal the performance of the proposed approach, four methods listed in Table 5.7 are compared. In method-1, the feature that has the highest monotonic relevance with the fault level is used as a single indicator. Feature fusion is not involved. In method-2, the feature fusion (i.e. PCA) is applied directly to all 108 features, and no feature selection is involved. In method-3 and method-4, both feature fusion and feature selection are involved. Dominance rough sets and fuzzy preference based rough sets both selected features in terms of monotonic relations with the fault levels; the first one is found to reveal qualitative information and the latter one is claimed to show quantitative information. To check the usefulness of the one over the other, method-3 of using PCA on features selected by dominance rough sets is adopted. The proposed method (method-4) employs PCA on the features selected through fuzzy preference based rough sets. It is important to note that the method-2 and method-3 follow the same algorithm shown in Table 5.5. The only difference is in the feature evaluation method. Methods 2 and 3 uses the approximation qualities defined in "dominance rough sets" (Equation (2.40)) and "fuzzy preference based rough sets" (Equation (2.50)) for the feature evaluation, respectively. The reason why these methods are compared are explained as follows. Comparison of method-1 and method-4 shows the limitation of one feature. Comparison of method-2 and method-4 demonstrates the necessity

of selecting features for the indicator generation. Comparison of method-3 and method-4 shows the importance of measuring the monotonic relevance when selecting features for indicator generation.

Table 5.7: Different methods to be compared

Method	Description
Method-1	Use one single feature (without PCA for feature fusion)
Method-2	PCA applied to all the 108 features
Method-3	PCA applied to features selected by dominance rough sets
Method-4 (proposed method)	PCA applied to features selected by fuzzy preference based rough sets

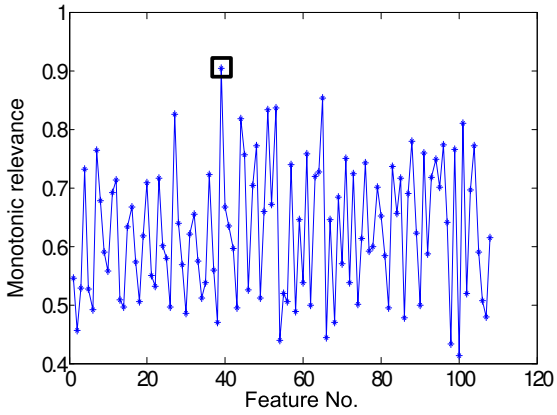


Figure 5.14: Feature evaluation using fuzzy preference based rough sets

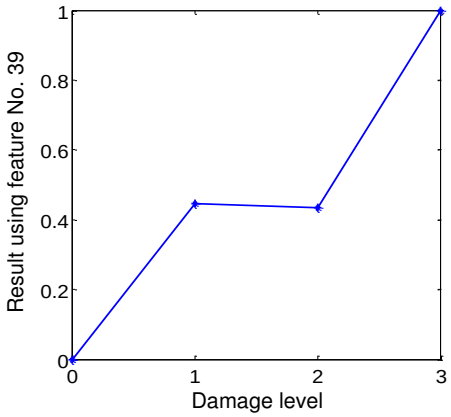


Figure 5.15: Trend of the best feature (No. 39 - Amplitude at -5X in full spectrum of A2YZ, method-1)

Method-1 is checked first. Figure 5.14 shows that feature No. 39 (marked with a small black square) gives the highest value of monotonic relevance. Figure 5.15 shows the value

of feature No. 39 for four damage levels. It can be seen that even with the most relevant feature, the damage levels are still not clearly revealed. The samples for moderate damaged impeller and slight damaged impeller are mixed up. This shows that the monotonic information is not fully contained in a single feature. Feature fusion is needed to extract the monotonic information distributed in different features. PCA is employed herein for this purpose.

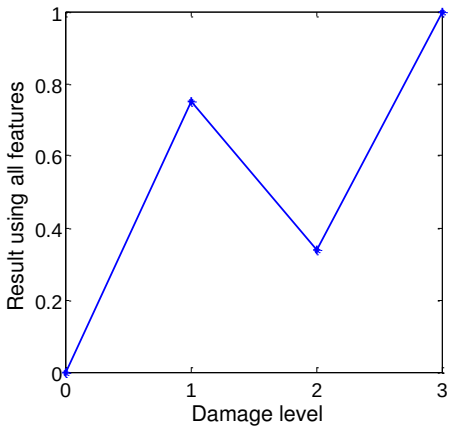


Figure 5.16: The 1^{st} principal component of 108 features Versus damage levels (method-2)

In method-2, PCA is applied directly to all 108 features. The first principal component of the 108 features is shown in Figure 5.16. In this figure, the monotonic trend is not observed. The reason is explained as follows. As shown in Figure 5.16, different features have different performance. The features whose performance values are small give little contribution to the indicator generation, which in turn results in the lost of monotonic trend. This shows the necessity of selecting features for indicator generation.

Method-3 uses dominance rough sets for feature evaluation. The results are shown in Figure 5.17, from which it can be seen that the monotonic trend is now obtained, but the boundary between slight damage and moderate damage is very small. This is because the dominance rough sets consider preference relation qualitatively, thus the selected features are not guaranteed to be the best monotonic ones quantitatively.

Finally, the proposed method (method-4) for indicator generation is tested. Figure 5.18 shows the results, from which different levels are clearly distinguished and more importantly the indicator monotonically increases with damage levels. By comparing Figures 5.15 - 5.17 with Figure 5.18, it can be seen that the proposed method performs best. It can not only distinguish different damage levels, but also keeps a clear monotonic trend with the damage levels. This means that in order to clearly reveal the monotonic trend, quantitative information which can be revealed through fuzzy preference based rough sets is needed.

In method-4 (the proposed method), five features were finally selected to generate an indicator. The physical meanings and positions for the chosen five features are listed in

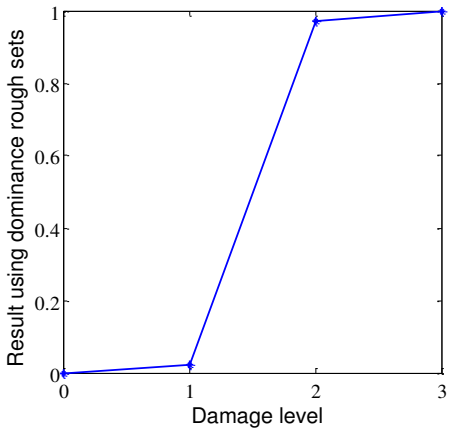


Figure 5.17: Results using dominance rough sets and PCA (method-3)

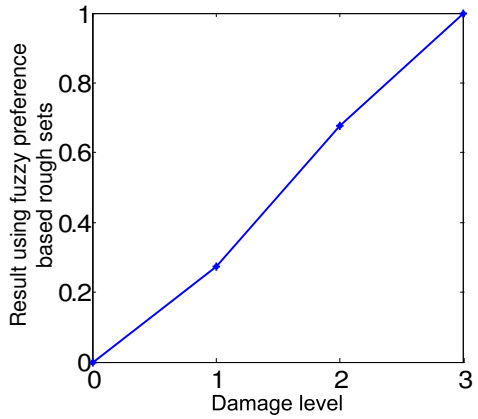


Figure 5.18: Results using fuzzy preference rough sets and PCA (the proposed method)

Table 5.8. Among the five selected features, two are from A3ZX, one each from A2YZ, A3XY and A2ZX planes. These are the planes, as stated at the beginning of Section 5.3.2, that are related to some flow patterns. A2YZ are in the plane of the secondary flow in the volute passage. A3XY is in the plane of the secondary flow in the impeller passage. A2ZX and A3ZX are in the plane of the main flow. This is consistent with our expectation that these planes are sensitive to impeller vane leading edge damage. Moreover, all of the features are from full spectrum; this shows the effectiveness of full spectrum in revealing the pump health information embedded in flow patterns.

This application shows that the proposed method successfully generates an indicator representing the development of damage levels for impeller vane leading edge damage. Fuzzy preference based rough sets help to select features that have better monotonic relevance with fault levels. PCA fuses the information on selected features. This indicator contains information from different sensors (i.e. vibration at different locations and directions). One drawback with this indicator generation method (method II) is that the physical

Table 5.8: Details on the five selected features

Chosen feature No.	Physical meaning	Position
39	Amplitude -5X (backward)	A2YZ
65	Amplitude 1X (forward)	A3ZX
43	Amplitude 5X (forward)	A2ZX
53	Amplitude -1X (backward)	A3XY
69	Amplitude -1X (backward)	A3ZX

meaning of the generated indicator is not as clear as that of method I because of the linear transformation involved in PCA.

5.4 Summary

In this chapter, two methods of generating indicators for fault levels by integrating information from possible sensors are presented.

The first method (method I) regards signals from two sensors and different health conditions as one multivariate signal. Multivariate empirical mode decomposition is adopted to decompose the multivariate signal into a set of IMFs. The fault-sensitive IMF is chosen by a criterion based on mutual information. Then a full spectra based indicator is obtained. The advantage of the indicator generated by method I is that it reveals the characteristic of a planar vibration motion. So it is useful for tracking the fault levels of the fault type that causes the changes of planar vibrations, e.g. impeller vane trailing edge damage. The disadvantage is that it does not work for a fault type that causes a vibration motion in more than two dimensions.

The second method (method II) extracts features first, and then uses the fuzzy preference based rough set to select features having better monotonic relevance with fault levels and PCA to combine information in selected features into a single indicator. The generated indicator makes use of information among different sensors and features, and outperforms each individual feature. This method is general and can work for one, two or more sensors. Thus it is useful for tracking the fault levels of the fault type that affects vibration in various locations and directions. This is the advantage of method II. However, because of the linear transformation induced by PCA, the generated indicator doesn't keep the physical meanings of the original selected features. This is the disadvantage of method II.

The two methods are applied to the identification of damage levels of impeller vane trailing edge damage and impeller leading edge damage, respectively. The lab experimental results show that the indicators generated by the two methods effectively and monotonically represent the damage levels, and therefore are useful in fault identification.

Chapter 6

A Machine-Learning-Based Method for Fault Identification

In Chapter 5, signal-based methods for fault identification are studied. They provide an indicator that monotonically varies with fault levels, and thus the fault level can be estimated by checking the value of this indicator. Expertise on the mechanisms of the fault is required for successful fault identification. In this chapter, machine-learning-based fault identification will be studied. A machine-learning-based method resorts to machine learning algorithms and builds a model in the training process expressing the relation between the features and the fault levels. This model is then used for automatic fault identification.

Researchers have used classification algorithms such as K-nearest neighborhood (KNN) [73] and Neural Network [79] to build a classification model for fault identification. Lei et al. [73] proposed a weighted KNN method for gear crack level identification. They also proposed a method using a combination of multiple classifiers to achieve higher accuracy of gear fault level classification [79]. In [73, 79], the fault level was regarded as a nominal variable, and the problem of tracking gear fault level was treated as a classification problem. This approach ignores the ordinal information among different fault levels. For example, a moderate fault is worse than (<) a slight fault but is better than (>) a severe fault. In classification, however, the severe, moderate and slight faults are parallel to each other and cannot be compared using ">" and "<" operations. Ordinal information is the main characteristic of the fault levels, which makes the diagnosis of fault levels more complicated than diagnosis of the fault types.

In order to keep the ordinal information among fault levels, some researchers build an assessment model using technologies such as fuzzy c-mean [143], Self-Organizing Map (SOM) [144], and Hidden Markov Model (HMM) [145]. The newly collected data is compared with the assessment model and a parameter is produced for estimating of the fault level for the new data. Qiu et al. [144] employed SOM to build a model and the distance between a newly collected data and the best matching unit of the SOM model was used for bearing health assessment. Pan et al. [143] built a model using bearing data collected under two health states (one is the normal state (i.e. the bearing is healthy) and the other is the

failure state (i.e. the bearing fails)) using fuzzy c-means. The distance of a newly collected data to the normal state was checked to estimate the fault level. Li and Limmer [146] built a linear auto-regression model using vibration data collected under the normal state. The distance (or difference) between the model output and the newly collected data was used to estimate the gear damage levels. Ocak [145] utilized a HMM to train a model based on bearing data collected under the normal state. It was found that the HMM yielded probabilities that decreased consistently as the bearing wore in time. In these methods, the model is built by utilizing data from one health status (i.e. normal state) or two heath statuses (normal state and failure state). Information on the intermediate fault levels (e.g. slight fault and moderate fault) is not included in the model.

Ordinal ranking [72] is a machine learning algorithm which generates a ranking model that expresses the ordinal information contained in the training data. A detailed description of ordinal ranking is given in Section 2.6.2.2. Ordinal ranking has been applied in the information retrieval field [85], but its application to fault diagnosis hasn't been reported yet. One issue to be solved in its application to fault diagnosis is feature selection, because most existing feature selection algorithms are for classification.

The objective of this chapter is to develop an intelligent method for diagnosing the fault levels using ordinal ranking to reserve the ordinal information among fault levels. The organization of this chapter is as follows. Section 6.2 introduces the background on ordinal ranking. Section 6.2 proposes a feature selection method for ordinal ranking. Section 6.3 presents a machine-learning-based method for fault identification. Section 6.4 applies the proposed method to pitting levels of planet gears in a planetary gearbox. Finally, summary comes in Section 6.5. The content of this chapter has been published in [84] and submitted to [147]. [1]

6.1 Background

Ordinal ranking generates a ranking model that expresses the ordinal information contained in the training data. Chu et al. [137] proposed an ordinal ranking algorithm (SVOR) based on support vector machine. The concept is briefed as follows. First, the original feature space ($\mathbf{x}$) is mapped into the high dimensional feature space ($\phi(\mathbf{x})$). In the feature space ($\phi(\mathbf{x})$), an optimal projection direction $\mathbf{w}$, and $L-1$ thresholds which define $L-1$ parallel discriminant hyperplanes for the L ranks correspondingly were found, as shown in Figure 1.8. The points satisfying $b_{i-1} < \mathbf{w}\cdot\phi(\mathbf{x}) < b_i$ are assigned the rank i as the label. The ranking model is thus

$$d = (\text{rank})\ i, \quad \text{if } b_{i-1} < \mathbf{w}\cdot\phi(\mathbf{x}) < b_i. \tag{6.1}$$

A detailed description on ordinal ranking and the algorithm SVOR are given in Section

[1] A version of this chapter has been accepted. "Xiaomin Zhao, Ming J Zuo, Zhiliang Liu and Mohammad Hoseini, 2012. Measurement (in press)."

2.6.2. Same as for many other machine learning techniques, feature selection is a necessary procedure for ordinal ranking, particularly because it can enhance accuracy [148]. One popular and powerful feature selection scheme is to select a feature subset that has maximum relevance to the label and minimum redundancy among themselves [53, 100, 148]. Measures are needed to evaluate the relevance between a feature and the label (i.e. feature-label relevance) and redundancy among features (i.e. feature-feature redundancy), respectively.

However, most existing measures are proposed for classification. Nevertheless, the measure for evaluating the relevance in classification is not suitable for ordinal ranking [60, 148], because the label (rank) of ordinal ranking is an ordinal variable, whereas the label of classification is a nominal variable as stated in Section 2.3. Mukras [60] found that the standard information gain (mutual information), though worked well in classification problems, failed in ordinal ranking problems. Baccianella et al. [149] used the idea of mutual information and introduced a filter method (called RRIGOR) for text-related applications of ordinal ranking. RRIGOR involves estimations of probability density functions which need large a number of samples and is computationally intensive. Correlation coefficient is conceptually simple and practically effective. In the next section, a feature selection method using correlation coefficients is proposed for ordinal ranking.

6.2 Proposed Feature Selection Method for Ordinal Ranking

According to the types of variables, several correlation coefficients are defined, as introduced in Section 2.3. The Pearson correlation coefficient evaluates the correlation between two continuous variables. The Polyserial correlation coefficient evaluates the correlation between a continuous variable and a ordinal variable. Their definitions are given in Section 2.4. The value of correlation coefficient varies from -1 to 1. A correlation coefficient of 1 means that the two variables are perfectly correlated; -1 means that the two variables are perfectly inversely correlated; 0 means that the two variables are not correlated. The absolute value of the correlation coefficients range from 0 to 1. If the absolute value of the correlation coefficient between two variables is closer to 0, then the two variables are less correlated.

A feature with a higher absolute value of the Polyserial correlation coefficient is more relevant to the rank, whereas a feature with a lower absolute value contributes little or even adverse information to the rank. Similarly, two features with large absolute value of the Pearson correlation coefficient share more redundant information. The proposed feature selection method follows the maximum relevance and minimum redundancy scheme. Let $T = \left\{x_1, x_2, \cdots, x_m, d\right\}$ be the data set, $C = [x_1, x_2, \cdots, x_m]_{n\times m}$ be the set of features which are represented by a $n \times m$ matrix where n is the total number of samples and m is the total number of features, and d be the variable of ranks for each sample. S is the set of selected features. $V(x_i, d)$ is the relevance between feature x_i and the rank d, $M(x_i, x_j)$ is

the redundancy between two features x_i and x_j. The proposed method selects S based on model (6.2), where t_1 and t_2 are two thresholds. t_1 is selected to avoid features with little or adverse information being included in S. t_2 is set to ensure that the redundancy between any two arbitrary features is below a certain level. The values of t_1 and t_2 are determined by the specific application problems. An example for selecting t_1 and t_2 will be given in Section 6.4.2.

$$\begin{aligned} \text{maximize} \quad & F(S) = \sum_{x_i \in S,\, S \subset C} V(x_i, d) \\ \text{subject to} \quad & V(x_i, d) > t_1. \\ & M(x_i, x_j) < t_2, \quad \exists\, x_i, x_j \in S,\ i \neq j. \end{aligned} \tag{6.2}$$

Table 6.1: The proposed feature selection method for ordinal ranking

Input: $T = \left\{x_1, x_2, \cdots, x_m, d\right\}$ - the data set
t_1, t_2 $(1 \geq t_1, t_2 \geq 0)$ - thresholds;
Output: S - the set of selected features

Step 1. $S = \emptyset$.

Step 2. calculate the relevance vector, $\mathbf{p} = [p_j]_{1\times m}$, whose element, p_j, is the absolute value of the Polyserial correlation coefficient between the j^{th} feature (i.e. x_j) and the rank (d). $j = 1, 2, \ldots, m$.
calculate the redundancy matrix $Q = [q_{ij}]_{m\times m}$, whose element, q_{ij},, is the absolute value of the Pearson correlation coefficient between the i^{th} feature (x_i) and the j^{th} features (x_j).$i, j = 1, 2, \ldots, m$.

Step 3. find the largest element in $\mathbf{p}$, i.e. $p_r = \max(\mathbf{p})$, then put the corresponding feature into S ($i.e. S = S \bigcup x_r$) and set p_r=0.

Step 4. find the features whose redundancy with feature x_r are not smaller than t_2, i.e. $h|q_{hr} \geq t_2$, then set their relevance values to zero (i.e. $p_h = 0$) so that these features won't be selected in future steps.

Step 5. check elements in $\mathbf{p}$. If $p_j \leq t_1$ for each $j = 1, 2, \ldots, m$, then go to Step 6; otherwise, go to Step 3.

Step 6. return S.

A sequential forward search strategy is used to find the solution to model (6.2), because it has the lowest computational load compared to other strategies [65]. The proposed feature selection method is described in Table 6.1. The selection starts with an empty set, $S = \emptyset$. Then the feature whose relevance to the rank is the highest is included into S. The process stops when the relevance of all unselected features are not larger than t_1 or the redundancy

between all unselected features and features in S are larger than t_2. The output is the set of selected features, S.

Note that in Table 6.1, the Polyserial and the Pearson correlation coefficients are used for evaluation of feature-label relevance and feature-feature redundancy, respectively, because the features extracted from vibration data are often continuous variables. In other applications, proper correlation coefficient can be chosen according to the types of features (Table 2.3) if the features are not continuous.

6.3 A Fault Identification Method Using Ordinal Ranking

Ordinal ranking can preserve the ordinal information, which makes it helpful in fault identification. Figure 6.1 shows the flow chart of the proposed fault identification method using ordinal ranking. Firstly, feature extraction is conducted to express the health information of the machinery from raw measured signals. Secondly, feature selection is conducted using the feature selection method proposed in Section 6.4.2. Finally, the selected feature subset is imported into the ordinal ranking algorithm (SVOR) described in Section 2.6.2 to diagnose the fault levels, and the output is the diagnosis results.

For the convenience of description, ranks '1', '2', '3', ..., 'L' are used to denote the baseline ('1') and progressively higher fault levels in subsequent sections. The larger the rank value is, the severer the fault is. The diagnosis results will be quantitatively evaluated using two metrics [11]: the mean absolute (MA) error (Equation 6.3) and the mean zero-one (MZ) error (Equation 6.4). MA error is affected by how wrongly a sample is diagnosed. The further the diagnosed rank is from the true rank, the larger the MA error is. If more ordinal information is preserved in the ranking model, the MA error is more likely to be smaller. MZ error, commonly used in classification problems, is affected only by whether a sample is wrongly diagnosed or not. If each rank is more clearly separated from others, the MZ error is more likely to be smaller. The smaller values of MA and MZ errors mean a better ranking model. In the two equations, n is the total number of samples, $d_i^{'}$ is the diagnosed rank for the i^{th} sample, and d_i is the true rank for the i^{th} sample.

Mean Absolute Error (MA error):

$$\frac{1}{n}\sum_{i=1}^{n}|d_i^{'}-d_i| \tag{6.3}$$

Mean Zero-one Error (MZ error):

$$\frac{1}{n}\sum_{i=1}^{n}t_i \quad \text{where } t_i=\begin{cases}1, & d_i^{'}\neq d_i\\ 0, & \text{otherwise}\end{cases} \tag{6.4}$$

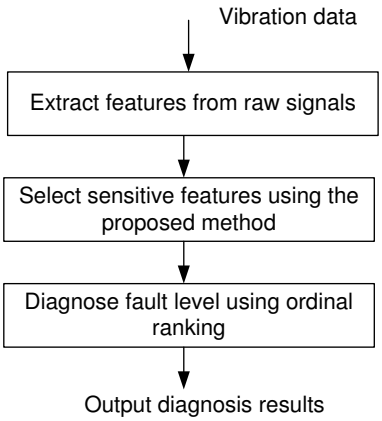


Figure 6.1: Proposed approach for diagnosis of fault levels

6.4 Application to Identification of Pitting Levels for Planetary Gears

The proposed method was used to identify pitting levels of planetary gears using experimental data collected from the planetary gearbox test rig described in Section 3.2. Four levels of pitting were considered, i.e. baseline, slight, moderate, and severe pitting. For details on the experimental design and data collection, refer to Section 3.2.

6.4.1 Feature Extraction

The traditional techniques for vibration-based gear fault diagnosis are typically based on statistical features of the collected vibration signals. The statistical features are easy to compute and effective in detecting gear faults, and thus are widely used [150, 151]. For example, Keller and Grabill [152] found that the two statistical features (sideband index and sideband level factors) were consistently successful in detecting the presence of a crack on a carrier. Many statistical features have been proposed and studied for fixed-shaft gearboxes, however, some of which are not suitable for planetary gearboxes.

For a fixed-shaft gearbox, damage to an individual gear tooth appears in the form of symmetric sidebands around the gear meshing frequency in the frequency domain. For the convenience of statistical feature extraction, regular, residual, difference and envelope signals are defined for fixed-shaft gearboxes [150, 153]. Regular signal is the inverse Fourier transform of the regular components which are defined as the fundamental shaft frequency, the fundamental and harmonics of the gear meshing frequency and their first order sidebands. Difference signal is defined as the time averaged signal excluding the regular mesh components. Residual signal is similar to difference signal but includes the first-order sidebands of the fundamental and harmonics of the gear meshing frequency. Envelope signal is the envelope of the signal band-pass filtered about the gear meshing frequency.

In a planetary gearbox, the dominant frequency component usually does not appear at

the gear meshing frequency because the planet gears are usually not in phase. In fact, the gear meshing frequencies are often completely suppressed, and sidebands are not symmetric around the meshing frequency any more. For description convenience, $f_{p,q}$ will be used to denote the frequency of $f = (p \cdot Z_r + q) \cdot f^c$ where Z_r is the number of ring gear teeth, f^c is the carrier frequency, p and q are integers ($p > 0, q > -p \cdot Z_r$). In an ideal planetary gearbox, only frequency components that appear at sidebands where $p \cdot Z_r + q = kN_p$ (N_p is the number of planets) will survive in a vibration signal [152]. Keller and Grabill [152] referred to the surviving sidebands with two different names: dominant sideband and apparent sideband. For each group of sidebands with the same value of p, there is one dominant sideband (donated by RMC_{p,q_d}), which is the one closest to the p^{th} harmonic of gear meshing frequency. Other surviving sidebands in this group are called apparent sidebands (donated by $RMC_{p,d}$). Let RMC^s denote the shaft frequency and its harmonics, RMC_{p,q_d+1} denote the first-order sideband of RMC_{p,q_d}. The regular ($g(k)$), difference ($d(k)$), residual ($r(k)$) and envelope ($e(k)$) signals for a planetary gearbox are then defined in Equations (6.5) - (6.8) [152]. In these Equations, $x(k)$ is a vibration signal in the time-waveform, F^{-1} is the inverse Fourier transform, $o(k)$ is the signal bandpass filtered about the dominant meshing frequency (RMC_{1,n_d}) and $H(o(k))$ is the Hilbert transform of $o(k)$.

$$g(k) = F^{-1}[RMC^s + RMC_{p,q} + RMC_{p,q_d} + RMC_{p,q_d \pm 1}] \tag{6.5}$$

$$d(k) = x(k) - g(k) \tag{6.6}$$

$$r(k) = x(k) - F^{-1}[RMC^s + RMC_{p,q} + RMC_{p,q_d}] \tag{6.7}$$

$$e(k) = \left| o(k) + iH(o(k)) \right| \tag{6.8}$$

With signals $g(k), d(k), r(k)$ and $e(k)$ defined, features can now be defined and calculated for planetary gearboxes. Sixty-three features are extracted from each vibration signal. These sixty-three features could be divided into three groups: the time-domain features, the frequency-domain features, and features specifically designed for gear fault diagnosis. Table 6.2 lists the definitions of the sixty-three features. For details on these features, refer to [150, 153]. In this table,

(1) The first group contains 18 commonly used time-domain features. They are maximum, minimum, average absolute, peak to peak, mean, root mean square (RMS), delta RMS, variance, standard deviation, skewness, kurtosis, crest factor, clearance factor, impulse factor, shape factor, coefficient of variation, coefficient of skewness, and coefficient of kurtosis.

(2) The second group contains 30 features from the frequency-domain spectrum. The first four features are statistical features: mean frequency, frequency center, root mean square frequency, and standard deviation frequency. The remaining twenty-six features are amplitudes at frequencies $f^1_{p,q} = (p \cdot Z_{r1} + q) \cdot f^{c1}$ and $f^2_{p,q} = (p \cdot Z_{r2} + q) \cdot f^{c2}$, where $p = 1$, q ranges from -6 to 6 with a step of 1, f^{c1} is the carrier frequency of the 1^{st} stage planetary gearbox, f^{c2} is the carrier frequency of the 2^{nd} stage planetary gearbox, Z_{r1} and

Z_{r2} are the number of teeth of ring gears of the 1^{st} stage and 2^{nd} stage planetary gearbox, respectively.

(3) The third group contains 15 features specifically designed for gearbox fault diagnosis. They are energy ratio, energy operator, $FM4$, $M6A$, $M8A$, $NA4$, $NB4$, $FM4^*$, $M6A^*$, $M8A^*$, $NA4^*$, NB^4*, $FM0$, sideband level factor and sideband index.

Table 6.2: Definations of features for planetary gear fault diagnosis

No.	Feature Name	Definition
Time-domain features		
F1	maximum	$x_{\max} = \max(x(k))$
F2	minimum	$x_{\min} = \min(x(k))$
F3	average absolute	$x_{\text{abs}} = \frac{1}{K}\sum_{k=1}^{K} \lvert x(k)\rvert$
F4	peak to peak	$x_{\text{p}} = \max(x(k))$-$\min(x(k))$
F5	mean	$\overline{x} = \frac{1}{K}\sum_{k=1}^{K} x(k)$
F6	RMS	$x_{\text{rms}} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} x(k)^2}$
F7	delta RMS	$x_{\text{drms}} = x_{\text{rms}}^{j} - x_{\text{rms}}^{j-1}$ where j is the current segment of time record and j-1 is the previous segment
F8	variance	$x_{\sigma^2} = \frac{1}{K}\sum_{k=1}^{K}(x(k) - \overline{x})^2$
F9	standard deviation	$x_{\sigma} = \sqrt{x_{\sigma^2}}$
F10	skewness	$x_{\text{sk}} = \frac{\frac{1}{K}\sum_{k=1} K(x(k)-\overline{x})}{x_{\sigma}^3}$
F11	kurtosis	$x_{\text{kur}} = \frac{\frac{1}{K}\sum_{k=1} K(x(k)-\overline{x})^4}{x_{\sigma}^4}$
F12	crest factor	$x_{\text{cf}} = \frac{x_{\max}}{x_{\text{rms}}}$
F13	clearance factor	$x_{\text{clf}} = \frac{\max(\lvert x(k)\rvert)}{(x_{\text{rms}})^2}$
F14	impulse factor	$x_{\text{if}} = \frac{\max(\lvert x(k)\rvert)}{x_{\text{abs}}}$
F15	shape factor	$x_{\text{sf}} = \frac{x_{\text{rms}}}{x_{\text{abs}}}$
F16	coefficient of kurtosis	$x_{\text{cv}} = \frac{\overline{x}}{x_{\sigma}}$
F17	coefficient of skewness	$x_{\text{cs}} = \frac{\frac{1}{K}\sum_{k=1}^{K} x(k)^3}{(x_{\sigma})^3}$
F18	coefficient of kurtosis	$x_{\text{ck}} = \frac{\frac{1}{K}\sum_{k=1}^{K} x(k)^4}{(x_{\sigma})^4}$
Frequency-domain features		
F19	mean frequency	$X_{\text{mf}} = \frac{1}{N}\sum_{n=1}^{N} X(n)$
F20	frequency center	$X_{\text{fc}} = \frac{\sum_{n=1}^{N}(f(n)\cdot X(n))}{\sum_{n=1}^{N} X(n)}$
F21	root mean square frequency	$X_{\text{rmsf}} = \frac{\sum_{n=1}^{N}(f(n)^2\cdot X(n))}{\sum_{n=1}^{N} X(n)}$
F22	standard deviation frequency	$X_{\text{stdf}} = \frac{\sum_{n=1}^{N}((f(n)-X_{\text{fc}})^2\cdot X(n))}{\sum_{n=1}^{N} X(n)}$
F23-F35	root mean square frequency	amplitudes at the following frequencies: $f_{1,p}^{1} = (Z_{r1} + p)\cdot f^{c1}$ where p=-6, -5, …, 6.

Table 6.2: (continued)

F36-F48	root mean square frequency	amplitudes at the following frequencies: $f_{1,p}^2 = (Z_{r2} + p)\cdot f^{c2}$ where p=-6, -5, …, 6.
Features specifically designed for planetary gearboxes		
F49	energy ratio	er = $\frac{\text{RMS(d(k))}}{\text{RMS(r(k))}}$
F50	energy operator	eo = kurtosis($y(k)$) where $y(k) = x(k)^2 - x(k-1)\cdot x(k+1)$
F51	FM4	FM4=kurtosis$d(k)$
F52	M6A	M6A=$\frac{\frac{1}{K}\sum_{k=1}^{K}(d(k)-\overline{d})^6}{(\frac{1}{N}\sum_{k=1}^{N}(d(k)-\overline{d})^2)^3}$
F53	M8A	M8A=$\frac{\frac{1}{K}\sum_{k=1}^{K}(d(k)-\overline{d})^8}{(\frac{1}{K}\sum_{k=1}^{K}(d(k)-\overline{d})^2)^4}$
F54	NA4	NA4=$\frac{\frac{1}{K}\sum_{k=1}^{K}(r(k)-\overline{r})^4}{(\frac{1}{M}\sum_{j=1}^{M}(\frac{1}{K'}\sum_{k=1}^{K'}(r_j(k)-\overline{r_j})^2))^2}$
F55	NB4	NB4=$\frac{\frac{1}{N}\sum_{k=1}^{K}(e(k)-\overline{e})^4}{(\frac{1}{M}\sum_{j=1}^{M}(\frac{1}{K'}\sum_{k=1}^{K'}(e_j(k)-\overline{e_j})^2))^2}$
F56	FM4*	FM4*=$\frac{\frac{1}{N}\sum_{k=1}^{K}(d(k)-\overline{d})^4}{(\frac{1}{M}\sum_{j=1}^{M}(\frac{1}{K'}\sum_{k=1}^{K'}(d_j(k)-\overline{d_j})^2))^2}$
F57	M6A*	M6A*=$\frac{\frac{1}{N}\sum_{k=1}^{K}(d(k)-\overline{d})^6}{(\frac{1}{M'}\sum_{j=1}^{M'}(\frac{1}{K'}\sum_{k=1}^{K'}(d_j(k)-\overline{d_j})^2))^3}$
F58	M8A*	M8A* =$\frac{\frac{1}{N}\sum_{k=1}^{K}(d_i-\overline{\mathbf{d}})^8}{(\frac{1}{M'}\sum_{j=1}^{M'}(\frac{1}{K'}\sum_{k=1}^{K'}(d_j(k)-\overline{d_j})^2))^4}$
F59	NA4*	M6A* =$\frac{\frac{1}{N}\sum_{k=1}^{K}(r(k)-\overline{r})^4}{(\frac{1}{M'}\sum_{j=1}^{M'}(\frac{1}{N}\sum_{k=1}^{K'}(r_j(k)-\overline{\mathbf{r_j}})^2))^2}$
F60	NB4*	NB4* =$\frac{\frac{1}{N}\sum_{k=1}^{K}(e(k)-\overline{e})^6}{(\frac{1}{M'}\sum_{j=1}^{M'}(\frac{1}{K'}\sum_{k=1}^{K'}(e_j(k)-\overline{e_j})^2))^2}$
F61	FM0	FM0= $\frac{\max(x(k))-\min(x(k))}{\sum_{p=1}^{P} RMC_{p,q_d}}$ where p is the total number of harmonics considered
F62	sideband level factor	slf =$\frac{RMC_{p,q_d-1}+RMC_{p,q_d+1}}{x_\sigma}$
F63	sideband index	si =$\frac{RMC_{p,q_d-1}+RMC_{p,q_d+1}}{x_\sigma}2$

6.4.2 Feature Selection

Four accelerometers namely LS1, LS2, HS1 and HS2 (shown in Figure 3.12) are used to collect vibration data. Features calculated from signals measured by each of the four sensors are combined, and totally 252 features are extracted. Features No.1 - No. 63 are from sensor LS1 following the order in Table 6.2, and features No. 64 - No. 126, No. 127 - No. 139, and No. 140 - No. 252 are from sensors LS2, HS1, and HS2, respectively.

The feature-label relevances (i.e. the absolute value of the Polyserial correlation coefficient) between each individual feature and the ranks (fault levels) are shown in Figure 6.2. It can be seen from Figure 6.2 that different features have different relevance values, some of which are very small. A threshold (t_1) is employed to determine whether a feature has

positive contribution to the ranks. Only features with relevance values above t_1 are considered to be useful in learning a ranking model. If t_1 is large, only a few really important features will be kept. If t_1 is small, most features will be kept and some might be useless. The choice of t_1 is problem dependent. Generally speaking, t_1 can be chosen to be 0.5, so that more than half information contained in an individual feature is related to the ranks. In the case that only a few features has the relevance values above 0.5, t_1 can be chosen as a smaller value so as to allow more features to be selected. In this application, t_1=0.5 is chosen. The largest value of the feature-label relevance in Figure 6.2 is 0.765 (feature No. 94), followed by 0.762 (feature No. 31), 0.762 (feature No. 157), and 0.752 (feature No. 220). These top four features (Nos. 94, 31, 157 and 220) are the amplitudes at sideband $(Z_{r1} + 2)\cdot f^{c1}$ from sensors LS2, LS1, HS1 and HS2, respectively. The feature-feature re-

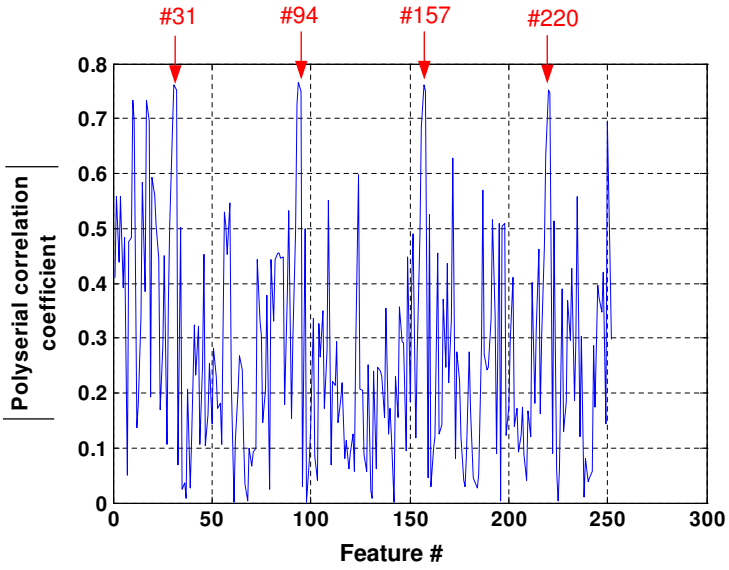


Figure 6.2: Feature-label relevance between damage levels and each of the 252 features

dundancies (i.e. the absolute value of the Pearson correlation coefficient) between the best feature (feature No. 94) and each of the 252 features are shown in Figure 6.3. It can be seen that some features (e.g. Nos. 31, 157 and 220) are highly related to feature No. 94; this means that a large amount of information in those features is also contained in feature No. 94. If these features are selected together with feature No.94, there will be redundant information. A threshold (t_2) is chosen to limit the redundancy among selected features. Features whose redundancy values with selected features are higher than t_2 will be omitted. If t_2 is large, only a few features will be omitted and finally most feature will be selected. If t_2 is small, most features will be omitted and finally only a few features will be selected. The choice of t_2 is problem dependent. Generally speaking, t_2 should be larger than 0.5, so that features that share more than half of the information with other selected features will be omitted. Depending on specific problems, t_2 can be increased to keep the number of

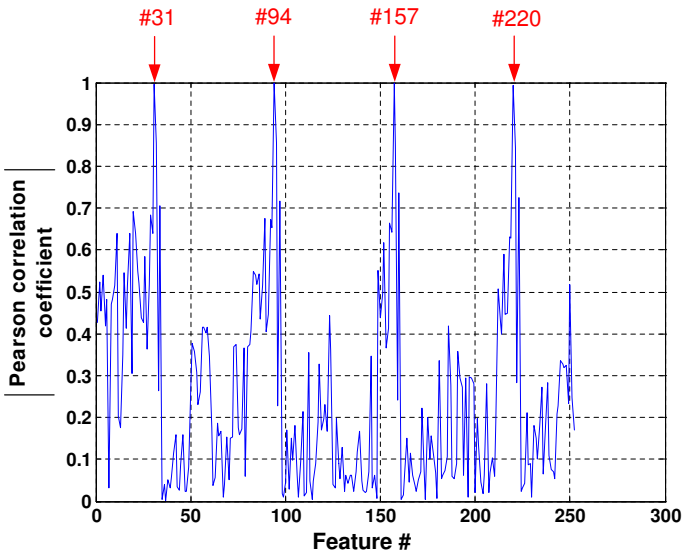


Figure 6.3: Feature-feature redundancy between feature No. 94 and each of the 252 features

selected features to a desired range. By checking Figure 6.3, t_2=0.8 is chosen so that the highly related features (i.e. feature Nos. 31, 157 and 220) are omitted and others can be further considered in next steps. Using the proposed feature selection method (Section 6.2), 11 features listed in Table 6.3 are selected.

Table 6.3: Eleven features selected by the proposed feature selection method

List	Feature No.	Physical meaning	Sensor
1	94	Amplitude at $(Z_{r1}+2)\cdot f^{c1}$	LS2
2	10	Skewness	LS1
3	93	Amplitude at $(Z_{r1}+1)\cdot f^{c1}$	LS2
4	11	Kurtosis	LS1
5	172	Amplitude at $(Z_{r2}+4)\cdot f^{c2}$	HS1
6	124	FM0	LS2
7	4	Peak to peak	LS1
8	89	Amplitude at $(Z_{r1}-3)\cdot f^{c1}$	LS2
9	192	Average absolute value	HS2
10	22	Standard deviation frequency	LS1
11	29	Amplitude at $Z_{r1}\cdot f^{c1}$	LS1

In the above example, the selection of t_1 and t_2 is based on the visual inspection of feature-label relevance plot and feature-feature redundancy plot. A more general way of selecting t_1 and t_2 is by testing different values of t_1 and t_2 and choose the values that generate a feature subset resulting in the smallest diagnosis error. The computational cost of this way of selection is, however, usually high.

6.4.3 Identification of Gear Pitting Levels

For the convenience of description, ranks '1', '2', '3', and '4' are used to denote the baseline, slight pitting, moderate pitting and severe pitting. To test the diagnostic ability of ordinal ranking, the whole data set is split into two subsets: the training set and the test set. The training set is for training a ranking model. The test set is for testing the diagnostic ability of the trained ranking model. In this chapter, three separate scenarios were considered to split the whole data set into two subsets, as listed in Table 6.4.

Table 6.4: Distributions of the training set and the test set in three scenarios

Scenario	Training Set		Test Set	
	No. of samples	ranks for samples	No. of samples	ranks for samples
Scenario 1	320	{'1', '2', '3', '4'}	320	{'1', '2', '3', '4'}
Scenario 2	480	{'1', '3', '4'}	160	{'2'}
Scenario 3	480	{'1', '2', '3'}	160	{'4'}

(1) In scenario 1, the whole data set is randomly split into two equal-sized subsets, one for training and the other for testing. In both the training set and the test set, samples from ranks '1', '2', '3' and '4' are included. This scenario tests the performance of the ranking model when the training set covers the whole rank range.

(2) In scenario 2, samples from only ranks '1', '3' and '4' are included in the training set; and samples from only rank '2' are included in the test set. Practically, it might occur that data of some fault levels are missed in the training set. This scenario examines the case when data of slight fault level are not collected for training. It tests the interpolation ability of the ranking model.

(3) In scenario 3, samples from only ranks '1', '2' and '3' are included in the training set; and samples from only rank '4' are included in the test set. Similar to scenario 2, this scenario examines the case when data of severe fault level are not collected for training. It tests the extrapolation ability of the ranking model.

The algorithm SVOR introduced in Section 2.6.2.2 is employed to train and test the ranking model. The 2^{nd} degree polynomial kernel was used as the kernel function. During the training process, the five-fold cross validation method was employed to determine the optimal value of the regularization cost parameter, C, involved in the problem formulation (2.80). In the five-fold cross validation, an initial search was done on a coarse grid of the region $\{0.1 < C < 100\}$ first, followed by a fine search around the best grid of the initial search. Diagnosis results are discussed next.

6.4.4 Results and Discussion

6.4.4.1 Effect of Feature Selection

To check the performance of the proposed feature selection method, five feature subsets are generated and employed for analyzing the same data in each scenario: (1) all 252 features; (2) top 38 relevant features; (3) 11 features in Table 6.3 (the proposed method); (4) randomly selected 11 features; (5) 11 features selected by the Pearson correlation coefficient.

Details on how and why these feature subsets are obtained are explained as follows. Feature subset (1) doesn't involve feature selection. Feature subset (2) follows the feature selection scheme that uses top ranking features without considering relationships among features [73]. Following this scheme, 38 features whose feature-label relevance values are larger than 0.5 are obtained. Comparison of feature subsets (1) and (2) shows the influence of irrelevant features. Feature subset (3) is generated by the proposed method following the feature selection scheme of maximum relevance and minimum redundancy. Comparison of feature subsets (2) and (3) demonstrates the influence of redundant features. Feature subset (4) chooses 11 features randomly. Comparison of feature subsets (3) and (4) further emphasizes the importance of proper feature selection. Feature subset (5) is generated using a feature-label evaluation measure (i.e. the Pearson (strictly, Point-biserial) correlation coefficient) that is employed in [100] for classification problems. Strictly, this measure is the Point-biserial correlation coefficient. The Point-biserial correlation coefficient is mathematically the same as the Pearson correlation coefficient, and the latter is more popularly called. This chapter follows Ref. [100] and calls measure the Pearson correlation coefficient in the following space. The generation process for feature subset (5) is the same as the proposed method except that the rank is regarded as a nominal variable and the Pearson correlation coefficient is used in evaluation of the feature-label relevance (In the proposed method, the Polyserial correlation coefficient is used). Comparison of feature subsets (3) and (5) indicates the proper evaluation measure for feature-label relevance in ordinal ranking.

Each of the five feature subsets is imported into SVOR to diagnose pitting levels in each scenario. Results are provided in Table 6.5-Table 6.7.

In Scenario 1, the training set and the test set are randomly generated. To reduce the impact of randomness on the test results, 30 runs are conducted. The average and the standard deviation of the 30 test errors of the 30 runs are provided in Table 6.5. Using all 252 features, the mean values of MA error and the MZ error are both 0.099. Using the 38 relevant features, the mean values of the MA error and the MZ error are reduced to 0.078 and 0.077, respectively. This shows that irrelevant features have adverse effects on the ranking model. In the proposed method, some redundant features are further deleted from the 38 features, keeping only 11 features. The mean values of the MA error and the MZ error are further reduced to 0.073 and 0.072, respectively. This shows that the redundant information can reduce the performance of the ranking model, and needs to be excluded. Using the randomly selected 11 features, the mean values of the MA error and the MZ error are 0.229

Table 6.5: Results of scenario 1 - 320 training samples (ranks '1', '2', '3', '4') and 320 test samples (ranks '1', '2', '3', '4')

Features used in ordinal ranking	MA Error (mean±standard deviation)	MZ Error (mean±standard deviation)
all 252 features	0.099±0.022	0.099±0.022
top 38 features	0.078±0.016	0.077±0.016
11 features (the proposed method)	0.073±0.012	0.072±0.012
11 randomly selected features	0.229±0.025	0.220±0.024
11 features selected using the Pearson correlation coefficient	0.083±0.020	0.082±0.019

and 0.220 respectively, which are relatively high. The reason is that not enough relevant information is adopted in these features and there might be redundant information as well. Using the 11 features selected by the Pearson correlation coefficient, the mean values of the MA error and the MZ error are 0.083 and 0.082, respectively. Compared with the results of the proposed method, it can be shown that the Pearson correlation coefficient work less efficiently than the Polyserical correlation coefficient (the proposed method). The reason is that the Pearson correlation coefficient cannot properly reflect the relevance between a continuous feature and an ordinal rank. As a result, relevant features are not correctly selected. In Scenario 1, the proposed method generates the lowest mean and standard deviation of the MA error and the MZ error.

Table 6.6: Results of scenario 2 - 480 training samples (ranks '1', '3', '4') and 160 test samples (rank '2')

Features used in ordinal ranking	No. of samples in predicted ranks			MA error	MZ error
	'1'	'3'	'4'		
all 252 features	0	124	36	1.225	1
top 38 features	0	149	11	1.069	1
11 features (the proposed method)	21	131	8	1.050	1
randomly selected 11 features	0	23	137	1.856	1
11 features selected using the Pearson correlation coefficient	0	71	89	1.556	1

In Scenario 2, the training samples are from ranks '1', '3' and '4' only. The test samples (rank '2') are predicted to be one of the three ranks (i.e. '1', '3', and '4'). Because rank '2' will never be predicted, the MZ error is always 1. MA errors only are checked and compared. In the perfect case, the test samples are all predicted to be either rank '1' or '3',

which are two closest ranks to the true rank (i.e. '2'). In this case, the MA error is 1. In the worst case, the test samples are all predicted to be rank '4', making a MA error of 2. The diagnosed ranks of 160 test samples, along with their MA and MZ errors under each feature subset are listed in Table 6.6. With all 252 features, 124 samples are predicted to be rank '3' and the rest 36 samples are predicted to be rank '4', making a MA error of 1.225. Using the top 38 features, 149 samples are ranked '3' and the rest are ranked '4', resulting in a MA error of 1.069. It can be seen that after deleting irrelevant features, the MA error is reduced. This shows that irrelevant features have negative effect on the interpolation ability of the ranking model. With the proposed method, eight samples are ranked as '4', and others are ranked as either '1' or '3', generating a MA error of 1.050. This indicates that deleting redundant features improves the interpolation ability of the ranking model. With randomly selected 11 features, 137 samples are ranked '4', giving a high MA error of 1.856. This is because randomly selected features contain irrelevant and redundant information. Using 11 features selected by the Pearson correlation coefficient, 89 samples are ranked as '4' and a MA error of 1.556 is generated, which means that the interpolation ability of this ranking model is poor. In Scenario 2, the proposed method demonstrates the best interpolation ability among the five.

Table 6.7: Results of scenario 3 - 480 training samples (ranks '1','2','3') and 160 test samples (rank '4')

Features used in ordinal ranking	No. of samples in predicted ranks			MA error	MZ error
	'1'	'3'	'4'		
all 252 features	0	76	84	1.475	1
top 38 features	0	34	126	1.215	1
11 features (the proposed method)	0	12	148	1.075	1
randomly selected 11 features	0	91	69	1.569	1
11 features selected using the Pearson correlation coefficient	0	47	113	1.294	1

In Scenario 3, the training samples are from ranks '1', '2' and '3' only. The test samples (rank '4') are predicted to be one of the three ranks (i.e. '1', '2', and '3'). Same as in Scenario 2, the MZ error is always 1 because rank '4' will never be predicted. MA errors only are checked and compared. In a perfect case, the test samples are all predicted to be rank '3', which is the closest rank to the true rank (i.e. '4'). In this case, the MA error is 1. In a worst case, the test samples are all predicted to be rank '1', making an MA error of 3. Table 6.7 shows the detailed results. With all 252 features, around half of the test samples (76 samples) are ranked '2' and half are ranked '3', making an MA error of 1.475. The top 38 features put 34 samples in rank '2' and others in rank '3', reducing the MA error to 1.215. This demonstrates that irrelevant features should be excluded in order to improve the

extrapolation ability of the ranking model. The features selected by the proposed method further reduce the MA error to 1.075 by eliminating the redundant information. Randomly selected features put most samples (91) into rank '2' and the rest into rank '3', giving an MA error of 1.569. This shows that if the relevance and redundant information are not considered during the feature selection, a good ranking model is hard to achieve. The 11 features selected using the Pearson correlation coefficient give a MA error of 1.294, indicating a worse extrapolation ability of the ranking model than that of the proposed method. This is because of the improper evaluation of feature-label relevance during the feature selection process. In this scenario, the proposed method generates the lowest MA error, and thus produces a ranking model with the best exploration ability among the five.

Comparisons between results of the proposed method and results of all 252 features, top 38 features, and randomly selected 11 features prove the benefits of deleting irrelevant features and redundant features. Comparisons between results of the proposed method and results of features selected using the Pearson correlation coefficient show the effectiveness of the Polyserical correlation coefficient (used in the proposed method) in evaluating the feature-label relevance for ordinal ranking problems. Using the Pearson (Point-biserial) correlation coefficient, the ranks are regarded as a nominal variable. That is why the Pearson (Point-biserial) correlation coefficient works well for classification problems not for ordinal ranking problems. In all of the three scenarios, the proposed method gives the lowest error, proving its effectiveness in building a ranking model for diagnosis of fault levels.

6.4.4.2 Comparison of Ordinal Ranking and Classification

For comparison purposes, the traditional diagnosis approach [73, 79] which uses a multi-class classifier to diagnose the fault levels is also applied to each scenario. To avoid the influence of the learning machine, support vector machine (SVM) is adopted as a classifier since SVOR (the ordinal ranking algorithm adopted in the proposed diagnosis approach) is based on SVM. One-against-all strategy [103] was used for multi-class classification using SVM. The same 2^{nd} degree polynomial kernel was utilized. The same five-fold cross validation was employed to find the optimal value of regularization cost parameter, C. The procedure of diagnosing pitting levels is also the same as described in Section 6.3 except that the ordinal ranking algorithm is replaced by the classification algorithm. Results of the proposed diagnosis approach (using ordinal ranking) and traditional diagnosis approach (using classification) for three scenarios are listed in Table 6.8.

In Scenario 1, the MA error of ordinal ranking (0.073) is smaller than that of classification (0.088), whereas the MZ error (0.072) is larger than that of classification (0.066). This could be explained as follows. The MZ error treats wrongly ranked samples equally and the value of MZ error isn't influenced by how well the ordinal information is kept. The more separately each rank is classified, the more likely that the MZ error is low. The aim of classification is to classify each rank as separately as possible; therefore classification gives

Table 6.8: Comparison of the proposed approach (ordinal ranking) and traditional approach (classification)

Diagnosis approach	Scenario 1				
	MA error				MZ error
proposed approach (ordinal ranking)	0.073				0.072
traditional approach (classification)	0.088				0.066
	Scenario 2				
	No. of samples in predicted ranks			MA error	MZ error
	'1'	'3'	'4'		
proposed approach (ordinal ranking)	21	131	8	1.050	1
traditional approach (classification)	0	80	80	1.500	1
	Scenario 3				
	No. of samples in predicted ranks			MA error	MZ error
	'1'	'2'	'3'		
proposed approach (ordinal ranking)	0	12	148	1.075	1
traditional approach (classification)	25	19	116	1.431	1

a lower MZ error. However, the MA error is influenced by how well the ordinal information is kept. It penalizes the wrongly ranked samples considering how far a sample is wrongly ranked from its true rank. The more ordinal information is kept in the ranking model, the more likely that MA error becomes small. Classification doesn't guarantee that the ordinal information is kept. Ordinal ranking, on the other hand, aims to express the ordinal information in the feature space, and therefore the ordinal information is largely preserved. That is why ordinal ranking produces a smaller MA error than classification. The above argument is also supported by results in Scenarios 2 and 3. In Scenario 2, the true rank is '2'. Using classification, half of test samples (80 samples) are classified into rank '3' and half into rank '4', generating a MA error of 1.500. Ordinal ranking gives a lower MA error (1.05). In Scenario 3, the true rank is '4'. 25 samples are ranked '2', 19 are rank '2' and the rest are rank '3', resulting in a MA error of 1.431 using classification. Ordinal ranking gives a lower MA error of 1.075.

The above comparisons show that the ordinal ranking results in a lower MA error, and classification generates a lower MZ error. For diagnosis of fault levels, a low MA error is more important than a low MZ error. The reason is explained as follows. A low MA error means that the diagnosed fault level of a new sample is close to its true level. A low MZ error, however, cannot ensure a "closer" distance between the diagnosed fault level and true level. There are chances that a "severe fault" sample is predicted to be a "no fault" sample, which needs to be avoided in fault diagnosis. In this sense, ordinal ranking is more suitable for fault diagnosis than classification. The advantage of ordinal ranking is more obvious

when data of a certain fault level are missing in the training process, as can be seen from Scenarios 2 and 3 in Table 6.8.

6.5 Summary

In this chapter, a machine-learning-based method using ordinal ranking is proposed for fault identification. A feature selection method based on correlation coefficients is developed to improve the diagnosis accuracy using ordinal ranking. The proposed method selects features that are relevant to ranks, and meanwhile ensures that the redundant information is limited to a certain level. The Polyserial correlation coefficient is employed in evaluating the relevance between features and ranks. The Pearson correlation coefficient is utilized in measuring the redundant information between two features. The feature selection method is applied to the detection of pitting levels of planet gears. Results show that the proposed feature selection method efficiently reduces the diagnosis errors, and improve the interpolation and extrapolation abilities of the model trained by ordinal ranking.

The use of ordinal ranking for fault identification (the proposed approach) and the use of classification for the same task are compared. Classification generates a lower mean zero-one error than ordinal ranking. However, ordinal ranking has advantages over classification in terms of lower mean absolute error, better interpolation ability and extrapolation ability. This is because of the unique properties of ordinal ranking and classification: i.e. ordinal ranking is designed to search a monotonic trend in the feature space to reflect the change of ranks; whereas classification is to search a plane to separate each rank. Results on diagnosis of pitting levels of planet gears in a planetary gearbox show the effectiveness of the proposed machine-learning-based method.

The proposed feature selection method is based on correlation coefficient, which considers only linear relationship between two variables. Nonlinear relationships between variables are yet to be investigated.

Chapter 7

A Machine-Learning-Based Method for Fault Detection, Isolation and Identification

Fault detection, isolation, and identification (FDII) are three main tasks in fault diagnosis as stated in Chapter 1. In Chapter 4, a machine-learning-based method for fault detection and isolation (i.e. diagnosis of fault types) is studied. In Chapter 6, a machine-learning-based method for fault identification (i.e. diagnosis of fault levels) is studied. Chapters 4 and 6 consider fault types and fault levels, separately. An integrated method which is capable of diagnosing both fault types and fault levels (i.e. conducting fault detection, isolation and identification (FDII) together) is more helpful in fault diagnosis [21]. This chapter studies this problem.

The organization of this chapter is as follows. Section 7.1 introduces the background. Section 7.2 proposes an integrated method for FDII. Section 7.3 applies the proposed method to impeller damage diagnosis in a pump. Summary comes in Section 7.4.

7.1 Background

Reported work on machine-learning-based FDII treats the fault level in the same way as the fault type, and utilizes classifiers to distinguish them simultaneously [83, 86]. Ge et al. [83] used a support vector machine (SVM) classifier to diagnose faults in sheet metal stamping operations including three fault types (i.e. misfeed, slug, and workpiece) and three fault levels for the workpiece (i.e. too thick, too thin and missing). Lei et al. [86] used an adaptive neuro-fuzzy inference system (ANFIS) to classify bearing faults including three fault types (i.e. outer race, inner race and ball) and two fault levels for each of the three fault types (i.e. defect size of 0.007 inch and 0.021 inch). However, these methods of using classifiers for FDII have two shortcomings. The first shortcoming is that a single classification model is built for diagnosis of fault types and fault levels simultaneously. Thus samples having different fault levels and the same fault type are treated as totally different fault conditions. As

a result, the information on a fault type is not fully expressed in such a classification model. Because these samples actually share the same information on that fault type considering that they have the same fault type (though different fault levels). The second shortcoming is, as discussed in Chapter 6, classifiers fail to keep inherent characteristics of the fault level (that is, the ordinal information among different fault levels). In this chapter, a new integrated diagnosis scheme for FDII is proposed overcoming the two shortcomings.

7.2 A Proposed Method for Fault Detection, Isolation and Identification (FDII)

The aforementioned two shortcomings of existing methods are further elaborated as follows. Fault types and fault levels are different in their influences on measured vibration signals. Fault types refer to different fault locations (e.g. impeller trailing edge and impeller leading edge) or different fault modes (e.g. unbalance and misalignment). Each fault type can be regarded as a source that causes the change of vibration signals. For example, different fault types may result in different characteristic frequencies. Unbalance of a bearing excites 1X (shaft frequency) and parallel misalignment of a bearing excites 2X (the second harmonic of shaft frequency) [29]. On the other hand, fault levels of the same fault type can be regarded as the severity of the source that causes the change of the measured vibration signals. Even though fault levels of the same fault type are different from each other, they all contain information on the same fault type. For example, different levels of the impeller trailing edge damage all affect the vane passing frequency as shown in Section 5.2.3. Therefore, fault levels with the same fault type are not totally different fault conditions. It is reasonable to include all fault levels of the same fault type when building the diagnosis model for the fault type. Considering the effect of fault levels and fault types as discussed above, a two-step diagnosis method is proposed. In the first step, diagnosis of fault types is conducted. Samples having the same fault type (regardless of their fault levels) are diagnosed. After the fault type is known, fault level is diagnosed in the second step.

Coming to the second shortcoming, from the measurement point of view, the fault type is a nominal variable and the fault level is an ordinal variable (shown in Section 2.3). So they need to be treated differently according to their properties. The fault type, because of its nominal property, can be diagnosed by employing classification techniques, as in Chapter 4. The fault level, if it is diagnosed using classifiers, the ordinal information contained among different fault levels can not be kept, as shown in Chapter 6. To overcome this, fault levels can be diagnosed using ordinal ranking techniques, as discussed in Chapter 6. Ordinal ranking [72] is a recently studied supervised learning algorithm, which is described in Section 2.6.2 and is not repeated here.

7.2.1 Method Description

To overcome the two aforementioned shortcomings, an integrated diagnosis method is developed. In this method, diagnosis of fault types is achieved by classification. For each fault type, diagnosis of fault levels is conducted by ordinal ranking. A flow chart of the proposed method is given in Figure 7.1. First, feature extraction is conducted using signal processing techniques. Second, feature selection is conducted for fault type classification, and for fault level ranking of each fault type, respectively. Third, the fault types is diagnosed using the classification model based on selected features. Fourth, the diagnosis results is checked. If the result shows that there is no fault, then the "no fault" is output, meaning that the machine is healthy. Otherwise, the fault level of that classified fault type is further diagnosed using the fault level ranking model of that fault type; the classified fault type and diagnosed fault level are output.

In Figure 7.1, two types of feature selection are involved, one for improving the performance of fault type classification model and the other for improving the performance of fault level ranking model of each fault type. The feature selection methods proposed in Sections 4.3 (for classification) and 6.2 (for ordinal ranking) can be employed, as will be shown in Section 7.3.

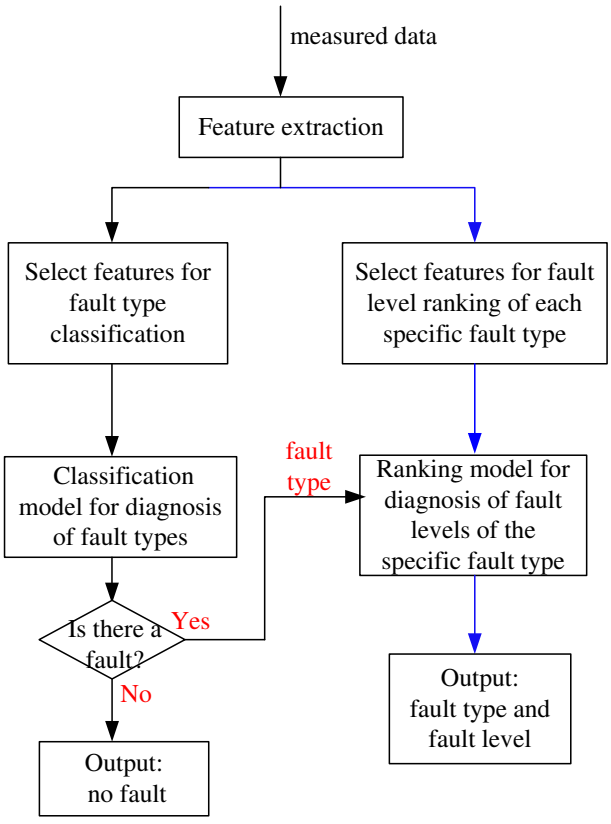


Figure 7.1: An integrated method for fault detection, isolation and identification (FDII)

7.2.2 Measures for Evaluating Diagnosis Results

In FDII, two labels are associated with each sample: the first label is the fault type and the second label is the fault level. The performance of diagnosis of fault types and the performance of diagnosis of fault levels should be evaluated differently, because the fault type and the fault level are different variables. The former is nominal variable and the latter is ordinal variable. Let d_{ti} and d_{ri} be the true fault type and the true fault level for the i^{th} sample, respectively. Let d'_{ti} and d'_{ri} be the diagnosed fault type and fault level for the i^{th} sample, respectively. The diagnosis results are evaluated using (1) the fault type zero-one error (TZ error) defined with Equation (7.1) and (2) fault level absolute error (LA error) defined with Equation (7.2). In these equations, n is the total number of samples, n_t is the total number of samples that meet $d_{ti} = d'_{ti}$. The first measure (TZ error) is actually classification error. It evaluates the diagnosis error of fault types. The second measure (LA error) evaluates the diagnosis error of fault levels based on the samples whose fault types are correctly diagnosed.

(1) fault type zero-one error (TZ error)

$$\text{TZ} = \frac{1}{N}\sum_{i=1}^{N} t_i, \quad \text{where } t_i = \begin{cases} 0, & \text{if } d_{ti} = d'_{ti} \\ 1, & \text{otherwise} \end{cases} \tag{7.1}$$

(2) fault level absolute error (LA error)

$$\text{LA} = \frac{1}{N_t}\sum_{i=1}^{N_t} t_i, \quad \text{where } t_i = \begin{cases} |d_{ri} - d'_{ri}|, & \text{if } d_{ti} = d'_{ti} \\ 0, & \text{otherwise} \end{cases} \tag{7.2}$$

In the following, the proposed method is applied to the fault diagnosis of impellers in slurry pumps. The experimental design and data collection are given in Section 3.1.

7.3 Application to FDII of Impellers in Slurry Pumps

In this application, three fault types (i.e. no damage (ND), impeller tailing edge damage (TED) and impeller leading edge damage (LED)) and three fault levels (slight, moderate and severe) for each fault type are considered. Data collected for each health status are listed in Table 7.1. The labels representing the fault type and the fault level are listed in parentheses in the first two columns of Table 7.1.

Same as in Section 5.3.2.1, four features (i.e. amplitudes at 1X, 2X, 5X and 10X) are extracted from the half spectrum, and eight features (i.e. amplitudes at -1X, -2X, -5X, -10X, 1X, 2X, 5X and 10X) are extracted from the full spectrum. From three tri-axial accelerometers, nine half spectra (from nine channels A1X, A1Y, A1Z, A2X, A2Y, A2Z, A3X, A3Y, A3Z) and nine full spectra (from nine planes A1YX, A1XZ, A1YZ, A2XY, A2ZX, A2YZ, A3XY, A3ZX, A3YZ) are obtained. So there are totally 108 ($4 \times 9 + 8 \times 9$) features extracted.

Table 7.1: Number of samples collected for each health condition [2]

Fault types (classes)	Fault levels (ranks)	Number of samples	
No damage (0)	No (0)	162	
Impeller TED (1)	Slight (1)	162	486
	Moderate (2)	162	
	Severe (3)	162	
Impeller LED (2)	Slight (1)	162	486
	Moderate (2)	162	
	Severe (3)	162	
Total		1134	

To show the effectiveness of the proposed diagnosis method, three methods are compared.

(1) The first method doesn't treat fault types and fault levels differently and uses a classifier to distinguish them simultaneously. There is only one single classification model built for diagnosis. This method is currently commonly used in literature [83, 86].

(2) The second method diagnoses fault types first, and then diagnoses fault levels for the specific fault type. That is, diagnosis of fault types and fault levels is conducted in two separate steps. Classification techniques are used in both steps. This method overcomes the first shortcoming of the commonly used methods as discussed in Section 7.2. But it doesn't solve the problem of the second shortcoming, i.e. ordinal information among fault levels does not kept in diagnosis of fault levels.

(3) The third method diagnoses fault types and fault levels in two separate steps as the second method does. The fault type is diagnosed using classification, and the fault level is diagnosed using ordinal ranking. This method follows the proposed diagnosis scheme. It overcomes the two short shortcomings discussed in Section 7.2.

Comparison of the first method and the second method shows the benefits of diagnosing fault types and fault levels separately. Comparison of the second method and the third method shows the benefits of utilizing ordinal ranking but classification for diagnosing fault levels.

Feature selection is considered in each method. In the first method, diagnosis of fault types and fault levels is treated as a seven-class classification problem. The feature selection method for classification introduced in Section 4.3 is applied, and four features are selected. They are: 'A3xzf_Amp1', 'A2z_Amp5', 'A1y_Amp2', and 'A2xzf_Amp5'. The features are named in the format of 'LETTER1_LETTER2'. 'LETTER1' stands for the source of this feature. It consists of three or five letters. If 'LETTER1' consists of three letters, then this feature is from a half spectrum, and 'LETTER1' stands for the channel name. If there are five letters, then this feature is from a full spectrum, the first four letters stands for two sensors and the last letter stands for forward ('f') or backward ('b') component in the full spectrum. The fourth (and five) digits in 'LETTER2' stands for certain harmonic of pump

rotating frequency. For example, 'A3xzf_Amp1' represents the amplitude at 1X (forward component) of the full spectrum of two signals measured from channel A3X and channel A3Z.

In the second method and the third method, diagnosis of fault types and fault levels are conducted separately. So feature selection need to be conducted for diagnosis of fault types and diagnosis of fault levels, separately. For diagnosis of faut types, the feature selection method for classification introduced in Section 4.3 is used. Three features ('A1y_Amp5', 'A2xzb_Amp1', and 'A2yzf_Amp5') are selected. For diagnosis of fault levels, feature selection method for ordinal ranking introduced in Section 6.2 is used. Four features ('A1xyf_Amp5', 'A3xzf_Amp5', 'A2xzf_Amp5', 'A1x_Amp5') are selected for the impeller TED levels, and twelve features ('A3xzf_Amp1', 'A3xzf_Amp10', 'A1y_Amp10', 'A3xyf_Amp5', 'A3xzf_Amp5', 'A1xyf_Amp5', 'A3xzf_Amp2', 'A3xyb_Amp10', 'A1xyb_Amp1', 'A2yzb_Amp5', 'A2xzf_Amp5', 'A1y_Amp2') are selected for the impeller LED levels.

Table 7.2: Results (mean ± standard deviation) of diagnosing pump fault types and fault levels

Method	TZ error	LA error
(1) Diagnose fault types and fault levels simultaneously using classification	0.060 ±0.021	0.105 ±0.072
(2) Diagnose fault types using classification first, and then diagnose fault levels using classification	0.044 ±0.0126	0.124 ±0.0781
(3) Diagnose fault types using classification first, and then diagnose fault levels using ordinal ranking (proposed method)	0.044 ±0.0126	0.050 ±0.005

The diagnosis errors for each of the three methods using their selected features as mentioned above are summarized in Table 7.2. The TZ error evaluates the performance of diagnosis of fault types. The first method produces the highest mean TZ error (0.060). The second method and the third method reduce the mean TZ error to 0.044. The mean LA error evaluates the performance of diagnosis of fault levels. The second method uses classification for diagnosis of fault levels, and generates a high mean LA error (0.124). The third method uses ordinal ranking for the same, and the mean LA error is reduced to 0.050. The third method produces the smallest TZ error and LA error (mean and standard deviation). This shows the effectiveness of the proposed diagnosis method.

The reasons why the proposed method works well are summarized below. The proposed diagnosis method takes the characteristics of fault types and fault levels into consideration, and thus overcomes the two shortcomings of the traditional methods as addressed in Section 7.2. Diagnosis of fault types, regardless of fault levels, is conducted first, such that (1) the information of the fault type is fully expressed, and (2) a complicated problem (i.e. a seven-class classification problem in this application) is changed to a relatively simple problem

(i.e. a three-class classification problem). As a result, the diagnosis error of fault types (TZ error) is reduced. Moreover, diagnosis of fault level is conducted by ordinal ranking. So, the ordinal information among fault levels is kept. As a result, the diagnosis error of fault levels (LA error) is reduced.

7.4 Summary

In this chapter, a machine-learning based method is proposed for fault detection, isolation and identification (i.e. diagnosis of both fault types and fault levels). The different characteristics of fault types and fault levels are analyzed and considered in the proposed method.

In the proposed method, the fault type is diagnosed first. Samples with a specific fault type and different fault levels are treated as having the same label (i.e. fault type). In such a way, the information on fault types is fully included in the classification model for fault types and the classification model is simplified. After the fault type is diagnosed, the fault level of the specific fault type is diagnosed using ordinal ranking. Application to diagnosis of impeller damage in slurry pumps shows the effectiveness of the proposed method.

The proposed method does not work for the case where multiple fault types are dependent on one another. That is, it does not consider the interaction among different fault types.

Chapter 8

Conclusions and Future Work

Fault diagnosis is of prime importance for the safe operation of mechanical systems. My research focuses on data-driven fault diagnosis. There are two main approaches in data-driven fault diagnosis: signal-based and machine-learning-based. Fault diagnosis consists of three tasks: fault detection, fault isolation and fault identification. Following the two approaches, this thesis studies the three tasks, with special attention paid to fault identification. This chapter summarizes my contributions to data-driven fault diagnosis, describes some problems that remain to be further addressed, and suggests directions for future work.

8.1 Summary and Conclusion

8.1.1 Signal-Based Fault Identification

In the signal-based approach, fault diagnosis is achieved by checking the values of fault indicators which are sensitive enough to represent the health condition of machines. The generation of fault indicators is the key issue in signal-based fault diagnosis. Among the three tasks of fault diagnosis, the third task (i.e. fault identification or diagnosis of fault levels) is the most difficult one. Utilizing the ordinal information is essential for fault identification. In order to do so, the fault indicator for fault identification needs to demonstrate a monotonic trend with the fault severity levels. Such an indicator, however, is often difficult to find. This thesis works on this problem and proposes two indicator generation methods for fault identification of rotating machinery by integrating information from multiple sensors.

The first proposed method extracts an indicator by processing signals from two sensors simultaneously. This method was inspired by the fact that the joint information of two sensors might be lost if the signal of each sensor is processed individually. This thesis adopts the full spectrum technique which processes signals from two sensors simultaneously. In order to focus on the sensitive frequency range, multivariate empirical mode decomposition and mutual information are used. Then, full spectrum analysis is applied to the selected frequency range, from which an indicator is generated. Application of this approach to

the identification of impeller vane trailing edge damage shows that the generated indicator clearly reflects the change of flow patterns in pumps, and exhibits a monotonic trend. The physical explanations of this indicator can be verified using numerical simulations of the flow fields in pumps through computational fluid dynamics (CFD). The indicator generated by this method reveals the characteristics of a planar vibration motion. It is useful for tracking the fault levels of the fault type that causes changes of planar vibrations. However, this method can deal with signals from two sensors only.

The second proposed method extracts an indicator integrating information from multiple sensors (possibly more than two). The idea of this method is to first extract features from each individual sensor or from two sensors together, and then combine features from all sensors and output a single indicator. The first step is achieved by signal processing techniques including those for one-dimensional signals (e.g. conventional Fourier spectrum) and for two-dimensional signals (e.g. full spectrum). In the second step, an indicator that has monotonic trend with the fault level is generated. How to combine the health information (features) from all sensors into a single indicator that represents the health condition (i.e. exhibits monotonic trend) is the focus of this method. The global fuzzy preference approximation quality defined in a fuzzy preference based rough set is used to evaluate the performance of features in terms of monotonic relevance with fault levels, and principal component analysis (PCA) is used to combine the information contained in the selected features. Application of this approach to the identification of impeller vane leading edge damage shows that the indicator generated clearly and monotonically reflects the damage levels of vane leading edge. This method is useful for the fault types that cause vibration in multiple directions and locations. The disadvantage, however, is that the physical meaning of the indicator is hard to interpret.

In summary, the contribution of this thesis to the signal-based fault diagnosis (specifically fault identification) are:

- Developed an indicator generation method for fault identification by integrating information from two sensors. This method is especially useful for tracking the fault levels of a certain fault type that causes the changes of planar vibrations.
- Developed an indicator generation method for fault identification by integrating information from multiple sensors. This method is useful for tracking the fault levels of a certain fault type that affects vibration in various locations and directions (more than two).

8.1.2 Machine-Learning-Based Fault Diagnosis

The machine-learning-based fault diagnosis consists of three steps: feature extraction, feature selection and machine learning. The focus of this thesis is feature selection.

Fault detection and isolation (i.e. diagnosis of fault types) can be regarded as a classification problem. Rough set is a powerful tool for feature selection in classification problems.

In this thesis, feature selection based on neighborhood rough set is studied. Neighborhood size is a key factor that affects the feature selection results. The effect of neighborhood size is analyzed, based on which the problem of applying the original neighborhood rough set for feature selection is discussed. That is, features are likely to be wrongly estimated if only one neighborhood size is used for all features which are obtained from different sensors. To overcome this, the original neighborhood rough set is modified with multiple neighborhood sizes. The modified neighborhood rough set considers the physical meaning of neighborhood size, i.e. neighborhood size is the noise level that the feature encounters. Thus, in the modified neighborhood rough set, each feature is associated with a neighborhood size that stands for its noise level. A feature selection algorithm based on the modified neighborhood rough set is then introduced. To test its performance in practical applications, the feature selection algorithm was applied to the diagnosis of three fault types in slurry pumps. It is found that the feature selection algorithm based on the modified neighborhood rough set produced lower classification errors than the feature selection algorithm based on the original neighborhood rough set.

Fault identification (i.e. diagnosis of fault levels) has an important characteristic, that is, there is ordinal information among different fault levels. In order to preserve the ordinal information, fault identification is regarded as an ordinal ranking problem in this thesis. A feature selection method based on correlation coefficient is proposed for ordinal ranking problems. Then a diagnosis method is proposed using ordinal ranking and the proposed feature selection algorithm. The diagnosis method is applied to identify four pitting levels of planet gears. Results show that the proposed feature selection method efficiently reduces the diagnosis errors, and improves the interpolation and extrapolation abilities of the ranking model. The use of ordinal ranking for fault identification (the proposed approach) and the use of classification for the same task have been compared. It is found that ordinal ranking has advantages over classification in terms of lower mean absolute error, better interpolation ability and better extrapolation ability.

Furthermore, an integrated diagnosis scheme that is capable of fault detection, fault isolation and fault identification is proposed. This proposed diagnosis scheme consists of two steps. In the first step, only fault type is diagnosed. Samples with a specific fault type but different fault levels are treated as having the same label (i.e. fault type), and diagnosed through classification. In such a way, the information on fault types is fully expressed and the classification model is simplified. In the second step, the fault levels are diagnosed for each fault type using the ordinal ranking technique. Application to impeller damage diagnosis in pumps shows that the proposed scheme produces smaller errors compared with the traditional diagnosis method in which diagnosis of fault types and fault levels is conducted using classification.

In summary, the contribution of this thesis to the machine-learning-based fault diagnosis are:

- Modified the neighborhood rough set considering the physical meaning of neighborhood size, and introduced a feature selection method based on the modified neighborhood rough set for fault detection and isolation.
- Proposed a feature selection algorithm for ordinal ranking, and developed a diagnosis method for fault identification using ordinal ranking. The proposed diagnosis method keeps the ordinal information among different fault levels, which has been largely ignored in the literature.
- Proposed an integrated diagnosis scheme for fault detection, isolation and identification. This scheme considers the different characteristics of fault types and fault levels.

8.2 Future Work

8.2.1 Signal-Based Fault Identification

In this thesis, signal-based fault identification is conducted when only one single fault type exists. If multiple fault types exist simultaneously, the identifications of the overall fault level and the fault level for each individual fault type need to be studied. This work is extremely difficult if different fault types interact with each other.

8.2.2 Machine-Learning-Based Fault Diagnosis

This thesis applies ordinal ranking to fault identification, and finds that ordinal ranking outperforms classification. The feature selection methods for ordinal ranking, however, haven't been well studied in the literature. This thesis proposes a feature selection method based on correlation coefficients. Other feature selection methods for classification, such as mutual information [53], FOCUS [54, 55] and Fisher criterion [50], might be generalized for ordinal ranking, which needs further study.

Moreover, in this thesis, supervised-learning algorithms are used for machine-learning-based fault diagnosis, because it fully uses the information of historical data. One main drawback of supervised-learning is that it can only diagnose the fault types / levels that are included in the training data. Thus, the historical data of all possible fault types and fault levels must be available for training. However, in some cases, it may be extremely difficult to acquire data for certain fault types / levels from real systems. To handle such cases, unsupervised-learning can be used. The combination of supervised learning and unsupervised learning taking advantage of both can be a direction for machine-learning-based fault diagnosis.

Bibliography

[1] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chap man and Hall, 1986.

[2] T. H. Patel and M. J. Zuo, "The pump loop, its modifications, and experiments conducted during phase II of the slurry pump project," tech. rep., Reliability Research Lab, University of Alberta, Edmonton, Alberta, 2010.

[3] M. Hoseini, D. Wolfe, Y. Lei, and M. J. Zuo, "Planetary gearbox test rig: system features and procedures for operation and assembly," tech. rep., Reliability Research Lab and Syncrude Research Center, 7 October 2008.

[4] X. Zhao, T. Patel, A. Sahoo, and M. J. Zuo, "Numerical simulation of slurry pumps with leading or trailing edge damage on the impeller," tech. rep., University of Alberta, August 2010.

[5] C. D. Bocaniala and V. Palade, "Computational intelligence methodologies in fault diagnosis: review and state of the art," in *Computational Intelligence in Fault Diagnosis* (V. Palade, L. Jain, and C. D. Bocaniala, eds.), Advanced Information and Knowledge Processing, Springer London, 2006.

[6] A. K. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mechanical Systems and Signal Processing*, vol. 20, no. 7, pp. 1483–1510, 2006.

[7] G. P. S. Raghava, "Principal of support vector machine." http://imtech.res.in/raghava/rbpred.

[8] P. Goldman and A. Muszynska, "Application of full spectrum to rotating machinery diagnostics," *Orbit*, vol. First Quarter, pp. 17–21, 1999.

[9] N. Rehman and D. P. Mandic, "Multivariate empirical mode decomposition," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, vol. 466, pp. 1291–1302, 2010.

[10] Q. Hu, D. Yu, J. Liu, and C. Wu, "Neighborhood rough set based heterogeneous feature subset selection," *Information Sciences*, vol. 178, pp. 3577–3594, 2008.

[11] W. Chu and S. Keerthi, "Support vector ordinal regression," *Neural Computation*, vol. 19, pp. 792–815, 2007.

[12] W. Slken, "Introduction to pumps - centrifugal pump." http://www.wermac.org.

[13] M. J. Zuo, S. Wu, Z. Feng, A. Aulakh, and C. Miao, "Condition monitoring of slurry pumps for wear assessment," tech. rep., Reliability Research Lab, University of Alberta, Edmonton, Alberta, March 30, 2007.

[14] Wapcaplet, "Epicyclic gear ratio." WikiPedia, December 2008.

[15] M. Pandey, T. Patel, X. Liang, T. Tian, and M. J. Zuo, "Descriptions of pitting experiments, run-to-failure experiments, various load and speed experiments, and crack experiments carried out on the planetary gearbox test rig," tech. rep., Department of Mechanical Engineering, University of Alberta, Edmonton, Alberta, 2011.

[16] M. Hoseini, Y. Lei, D. V. Tuan, T. Patel, and M. J. Zuo, "Experiment design of four types of experiments: pitting experiments, run-to- failure experiments, various load and speed experiments, and crack experiments," tech. rep., Department of Mechanical Engineering, University of Alberta, Edmonton, Alberta, 2011.

[17] J. F. Gulich, *Centrifugal pumps*. New York: Spinger, 2010.

[18] *Warman Centrifugal Slurry Pumps - Slurry Pump Handbook*. Weir Slurry Group, Inc., 2006.

[19] A. Heng, S. Zhang, A. C. Tan, and J. Mathew, "Rotating machinery prognostics: state of the art, challenges and opportunities," *Mechanical Systems and Signal Processing*, vol. 23, no. 3, pp. 724–739, 2009.

[20] R. Isermann and P. Ball, "Trends in the application of model-based fault detection and diagnosis of technical processes," *Control Engineering Practice*, vol. 5, no. 5, pp. 709–719, 1997.

[21] E. Sobhani-Tehrani and K. Khorasani, *Fault Diagnosis of Nonlinear Systems Using a Hybrid Approach*. Springer, 2009.

[22] E. P. Carden and P. Fanning, "Vibration based condition monitoring: a review," *Structural Health Monitoring*, vol. 3, no. 4, pp. 355–377, 2004.

[23] J. Chen and R. Patton, *Robust model-based fault diagnosis for dynamic systems*. Kluwer Academic Publishers, 1999.

[24] Y. Feng, Y. Qiu, C. J. Crabtree, and P. J. Tavner, "Predicting gearbox failures in wind turbines," tech. rep., Energy Group, School of Engineering and Computing Sciences, Durham University, 2011.

[25] T. Miyachi and J. Kato, "An investigation of the early detection of defects in ball bearings using vibration monitoring - pratical limit of detectability and growth speed of defects," in *The International Conference on Rotordynamics, JSMEIFTOMM*, (Tokyo), pp. 14–17, 1986.

[26] B. Tao, L. Zhu, H. b. Ding, and Y. Xiong, "Renyi entropy-based generalized statistical moments for early fatigue defect detection of rolling-element bearing," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 221, no. 1, pp. 67–79, 2007.

[27] B. B. Sassi S. and T. M., "TALAF and THIKAT as innovative time domain indicators for tracking ball bearings," in *Proceedings of the 24nd Seminar on machinery vibration* (M. Thomas, ed.), pp. 404–419, Canadian Machinery Vibration Association, 2005.

[28] "Iso standard 10816," in *Mechanical Vibration-Evaluation of Machine Vibration by measurements on non-rotating parts*, Geneva, Switzerland: International Organization for Standardization, 1995.

[29] M. P. Norton and D. G. Karczub, *Fundamentals of Noise and Vibration Analysis for Engineers*. Cambridge University Press; second edition, 2003.

[30] P. D. McFadden, "Determining the location of a fatigue crack in a gear from the phase of the change in the meshing vibration," *Mechanical Systems and Signal Processing*, vol. 2, no. 4, pp. 403 – 409, 1988.

[31] A. Oppenheim and R. Schafer, "From frequency to quefrency: a history of the cepstrum," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 95–99, 2004.

[32] A. McCormick and A. Nandi, "Bispectral and trispectral features for machine condition diagnosis," *IEE Proceedings -Vision, Image and Signal Processing*, vol. 146, pp. 229–234, oct 1999.

[33] F. E. H. Montero and O. C. Medina, "The application of bispectrum on diagnosis of rolling element bearings: a theoretical approach," *Mechanical Systems and Signal Processing*, vol. 22, no. 3, pp. 588–596, 2008.

[34] L. Qu, X. Liu, G. Peyronne, and Y. Chen, "The holospectrum: a new method for rotor surveillance and diagnosis," *Mechanical Systems and Signal Processing*, vol. 3, no. 3, pp. 255–267, 1989.

[35] W. Li, G. Zhang, T. Shi, and S. Yang, "Gear crack early diagnosis using bispectrum diagonal slice," *Chinese Journal of Mechanical Engineering (English Edition)*, vol. 16, no. 2, pp. 193–196, 2003.

[36] C. W. Lee and Y. S. Han, "The directional wigner distribution and its applications," *Journal of Sound and Vibration*, vol. 216, pp. 585–600, 1998.

[37] T. H. Patel and A. K. Darpe, "Use of full spectrum cascade for rotor rub identification," *Advances in Vibration Engineering*, vol. 8, pp. 139–151, 2009.

[38] J. Lin and M. J. Zuo, "Gearbox fault diagnosis using adaptive wavelet filter," *Mechanical Systems and Signal Processing*, vol. 17, pp. 1259–1269, NOV 2003.

[39] Z. Peng and F. Chu, "Application of the wavelet transform in machine condition monitoring and fault diagnostics: a review with bibliography," *Mechanical Systems and Signal Processing*, vol. 18, no. 2, pp. 199–221, 2004.

[40] Q. Gao, C. Duan, H. Fan, and Q. Meng, "Rotating machine fault diagnosis using empirical mode decomposition," *Mechanical Systems and Signal Processing*, vol. 22, no. 5, pp. 1072–1081, 2008.

[41] D. Dyer and R. M. Stewart, "Detection of rolling element bearing damage by statistical vibration analysis," *Journal of Mechanical Design*, vol. 100, no. 2, pp. 229–235, 1978.

[42] Z. Feng, M. J. Zuo, and F. Chu, "Application of regularization dimension to gear damage assessment," *Mechanical Systems and Signal Processing*, vol. 24, no. 4, pp. 1081 – 1098, 2010.

[43] S. Loutridis, "Instantaneous energy density as a feature for gear fault detection," *Mechanical Systems and Signal Processing*, vol. 20, no. 5, pp. 1239–1253, 2006.

[44] X. Zhao, T. H. Patel, and M. J. Zuo, "Multivariate EMD and full spectrum based condition monitoring for rotating machinery," *Mechanical Systems and Signal Processing*, vol. 27, pp. 712–728, 2011.

[45] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. 14, pp. 55–63, jan 1968.

[46] J. Emilyn and Er.K.Ramar, "Rough set based clustering of gene expression data: a survey," *International Journal of Engineering Science and Technology*, vol. 12, pp. 7160–7164, 2010.

[47] J. H. Friedman and W. Stuetzle, "Projection pursuit regression," *Journal of the American Statistical Association*, vol. 76, no. 376, pp. 817–823, 1981.

[48] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[49] S. Yvan, I. Iaki, and L. Pedro, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[50] W. Malina, "On an extended fisher criterion for feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-3, pp. 611–614, sept. 1981.

[51] K. Kira and L. Rendell., "The feature selection problem: traditional methods and a new algorithm," in *Proceedings of Ninth National Conference on Artificial Intelligence*, 1992.

[52] M. A. Hall, *Correlation-based feature selection for machine learning*. PhD thesis, University of Waikato, 1999.

[53] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226–1238, 2005.

[54] H. Almuallim and T. G. Dietterich, "Learning with many irrelevant features," in *In Proceedings of the Ninth National Conference on Artificial Intelligence*, pp. 547–552, AAAI Press, 1991.

[55] M. Dash, H. Liu, and H. Motoda, "Consistency based feature selection," in *Knowledge Discovery and Data Mining. Current Issues and New Applications* (T. Terano, H. Liu, and A. Chen, eds.), vol. 1805 of *Lecture Notes in Computer Science*, pp. 98–109, Springer Berlin / Heidelberg, 2000.

[56] J. Richard and S. Qiang, "Rough set based feature selection: a review," *Advanced Reasoning Group*, pp. 1–50, 2007.

[57] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, pp. 131–156, 1997.

[58] Selwyn and Piramuthu, "Evaluating feature selection methods for learning in data mining applications," *European Journal of Operational Research*, vol. 156, no. 2, pp. 483–494, 2004.

[59] K. Michalak and H. Kwasnicka, "Correlation-based feature selection strategy in neural classification," in *Intelligent Systems Design and Applications, 2006. ISDA '06. Sixth International Conference on*, vol. 1, pp. 741 –746, oct. 2006.

[60] R. Mukras, N. Wiratunga, R. Lothian, S. Chakraborti, and D. Harper, "Information gain feature selection for ordinal text classification using probability re-distribution," in *The IJCAI07 workshop on texting mining and link analysis, Hyderabad IN*, 2007.

[61] H. Almuallim and T. G. Dietterich, "Learning boolean concepts in the presence of many irrelevant features," *Artificial Intelligence*, vol. 69, pp. 279–305, 1994.

[62] E. Tsang, D. Chen, D. Yeung, X. Z. Wang, and J. Lee, "Attributes reduction using fuzzy rough sets," *Fuzzy Systems, IEEE Transactions on*, vol. 16, pp. 1130–1141, oct. 2008.

[63] A. Arauzo-Azofra, J. Benitez, and J. Castro, "Consistency measures for feature selection," *Journal of Intelligent Information Systems*, vol. 30, pp. 273–292, 2008.

[64] Z. Pawlak, *Rough sets: Theoretical aspects of reasoning about data*. Kluwer academic publisher, Dordrecht, 1991.

[65] J. Doak, "Intrusion detection: the application of input selection, a comparison of algorithms, and the application of a wide area network analyzer," Master's thesis, University of California, 1992.

[66] P. Somol, P. Pudil, and J. Kittler, "Fast branch amp; bound algorithms for optimal feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 7, pp. 900–912, 2004.

[67] Y. Li and Y. Liu, "A wrapper feature selection method based on simulated annealing algorithm for prostate protein mass spectrometry data," in *Computational Intelligence in Bioinformatics and Computational Biology, 2008. CIBCB '08. IEEE Symposium on*, pp. 195 –200, sept. 2008.

[68] W. Cohen, R. Greiner, and D. Schuurmans, *Probabilistic hill climbing*, vol. 2 of *Computational Learning Theory and Natural Learning Systems*. MIT Press, MA, 1994.

[69] C. L. Huang and C. J. Wang, "A GA based feature selection and parameters optimizationfor support vector machines," *Expert Systems with Applications*, vol. 31, no. 2, pp. 231–240, 2006.

[70] X. Wang, J. Yang, X. Teng, W. Xia, and R. Jensen, "Feature selection based on rough sets and particle swarm optimization," *Pattern Recognition Letters*, vol. 28, no. 4, pp. 459–471, 2007.

[71] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Springer, 1998.

[72] H. Lin, *From ordinal ranking to binary classification*. PhD thesis, California Institute of Technology, Pasadena, Unites States, 2008.

[73] Y. Lei and M. J. Zuo, "Gear crack level identification based on weighted K nearest neighbor classification algorithm," *Mechanical Systems and Signal Processing*, vol. 23, no. 5, pp. 1535–1547, 2009.

[74] W. Sun, J. Chen, and J. Li, "Decision tree and PCA-based fault diagnosis of rotating machinery," *Mechanical Systems and Signal Processing*, vol. 21, no. 3, pp. 1300–1317, 2007.

[75] Q. Hu, Z. He, Z. Zhang, and Y. Zi, "Fault diagnosis of rotating machinery based on improved wavelet package transform and SVMs ensemble," *Mechanical Systems and Signal Processing*, vol. 21, no. 2, pp. 688–705, 2007.

[76] zge Uncu and I. Trksen, "A novel feature selection approach: combining feature wrappers and filters," *Information Sciences*, vol. 177, no. 2, pp. 449–466, 2007.

[77] D. F. Specht, "Probabilistic neural networks," *Neural Networks*, vol. 3, no. 1, pp. 109–118, 1990.

[78] P. Wasserman, *Advanced Methods in Neural Networks*. Van Nostrand Reinhold, New York, 1993.

[79] Y. Lei, M. J. Zuo, Z. He, and Y. Zi, "A multidimensional hybrid intelligent method for gear fault diagnosis," *Expert Systems With Applications*, vol. 37, no. 2, pp. 1419–1430, 2010.

[80] B. S. K. A. Balushi and S.A.A1Araimi, "Aritificial neural networks and support vector machines with genetic algorithm for bearing fault detection," *Engineering Applications of Aritificial Intelligence*, vol. 16, pp. 657–665, 2003.

[81] B. Li, M. Chow, Y. Tipsuwan, and J. Hung, "Neural-network-based motor rolling bearing fault diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 47, no. 5, pp. 1060–1069, 2000.

[82] S. F. Yuan and F. L. Chu, "Support vector machines-based fault diagnosis for turbo-pump rotor," *Mechanical Systems and Signal Processing*, vol. 20, no. 4, pp. 939–952, 2006.

[83] M. Ge, R. Du, G. Zhang, and Y. Xu, "Fault diagnosis using support vector machine with an application in sheet metal stamping operations," *Mechanical Systems and Signal Processing*, vol. 18, no. 1, pp. 143–159, 2004.

[84] X. Zhao, M. J. Zuo, and Z. Liu, "Diagnosis of pitting damage levels of planet gears based on ordinal ranking," in *IEEE International Conference on Prognostics and Health management, Denver, U.S., June 20-23*, 2011.

[85] C. He, C. Wang, Y. X. Zhong, and R. F. Li, "A survey on learning to rank," in *International Conference on Machine Learning and Cybernetics*, 2008.

[86] Y. Lei, Z. He, and Y. Zi, "A new approach to intelligent fault diagnosis of rotating machinery," *Expert Systems with Applications*, vol. 35, no. 4, pp. 1593–1600, 2008.

[87] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., second ed., 1999.

[88] N. E. Huang, Z. Shen, S. R. Long, M. L. C. Wu, H. H. Shih, Q. N. Zheng, N. C. Yen, C. C. Tung, , and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London Series A-Mathematical Physical and Engineering Sciences*, vol. 454, pp. 903–995, 1998.

[89] D. Looney and D. P. Mandic, "Multiscale image fusion using complex extensions of EMD," *IEEE Transactions on Signal Processing*, vol. 57, pp. 1626–1630, 2009.

[90] G. Rilling, P. Flandrin, P. Goncalves, and J. M. Lilly, "Bivariate empirical mode decomposition," *IEEE Signal Processing Letters*, vol. 14, pp. 936–939, 2007.

[91] N. Rehman and D. P. Mandic, "Empirical mode decomposition for trivariate signals," *IEEE Transactions on Signal Processing*, vol. 58, pp. 1059–1068, 2010.

[92] S. S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, no. 2684, pp. 677–680, 1946.

[93] Q. Hu, D. Yu, and M. Guo, "Fuzzy preference based rough sets," *Information Sciences*, vol. 180, no. 10, pp. 2003–2022, 2010.

[94] I. Guyon, *Practical Feature Selection: from Correlation to Causality*. IOS Press, 2008.

[95] X. Zhao, Q. Hu, Y. Lei, and M. J. Zuo, "Vibration-based fault diagnosis of slurry pump impellers using neighbourhood rough set models," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 224, no. 4, pp. 995–1006, 2010.

[96] C. Chiu, N. H. Chiu, and C. Hsu, "Intelligent aircraft maintenance support system using genetic algorithms and case-based reasoning," *The International Journal of Advanced Manufacturing Technology*, vol. 24, pp. 440–446, 2004.

[97] R. E. Schumacker and R. G. Lomax, *A beginner's guide to structural equation modeling*. New Jersey: Lawrence Erlbaum Associates, Inc., second ed., 2004.

[98] D. Muijs, *Doing quantitative research in education with SPSS*. London: Sage Publications Ltd, second ed., 2010.

[99] K. G. Calkin, "Applied statistics." http://www.andrews.edu/ calkins, 2005.

[100] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, December 2004.

[101] U. Olsson, F. Drasgow, and N. Dorans, "The polyserial correlation coefficient," *Psychometrika*, vol. 47, pp. 337–347, 1982.

[102] X. Zhao, M. J. Zuo, and T. H. Patel, "Generating an indicator for pump impeller damage levels using half and full spectra, fuzzy preference based rough sets, and PCA," *Measurement Science and Technology*, vol. 23, no. 4, pp. 1–11, 2012.

[103] V. N. Vapnik, *Statistical learning theory*. New York: Wiley-Interscience, 1998.

[104] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations," *Philosophical Transactions of the Royal Society A*, vol. 209, pp. 415–446, 1909.

[105] J. Zhong, Z. Yang, and S. Wong, "Machine condition monitoring and fault diagnosis based on support vector machine," in *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM),*, pp. 2228–2233, 2010.

[106] A. Widodo and B.-S. Yang, "Support vector machine in machine condition monitoring and fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 21, no. 6, pp. 2560–2574, 2007.

[107] A. Agresti, *Analysis of ordinal categorical data*. John Wiley & Son, 1984.

[108] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regressionfilter," in *Advances in Large Margin Classifiers* (P. J. Bartlett, B. Schölkopf, D. Schuurmans, and A. J. Smola, eds.), pp. 115–132, MIT Press, 2000.

[109] Y. Freund, R. Iyer, Schapire, R.E.c, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of Machine Learning Research*, vol. 4, no. 6, pp. 933–969, 2004.

[110] H. T. Lin and L. Li, "Large-margin thresholded ensembles for ordinal regression: Theory and practice," *Algorithmic Learning Theory, Proceedings*, vol. 4264, pp. 319–333, 2006.

[111] K. Crammer and Y. Singer, "Pranking with ranking.," *Advances in Neural Information Processing Systems*, vol. 14, pp. 641–647, 2001.

[112] A. Shashua and A. Levin, "Ranking with large margin principle: two approaches," in *Proceedings of Advances in Neural Information Processing Systems*, 2002.

[113] L. Li and H. T. Lin, "Ordinal regression by extended binary classification," in *Advances in Neural Information Processing Systems*, pp. 865–872, 2007.

[114] J. Cardoso and J. Pinto Da Costa, "Learning to classify ordinal data: the data replication method," *Journal of Machine Learning Research*, vol. 8, pp. 1393–1429, 2007.

[115] B. Y. Sun, J. Li, D. D. Wu, X. M. Zhang, and W. B. Li, "Kernel discriminant learning for ordinal regression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 906–910, 2010.

[116] Y. A. Khalid and S. M. Sapuan, "Wear analysis of centrifugal slurry pump impellers," *Industrial Lubrication and Tribology*, vol. 59, no. 1, pp. 18–28, 2007.

[117] M. Inalpolat and A. Kahraman, "A theoretical and experimental investigation of modulation sidebands of planetary gear sets," *Journal of Sound and Vibration*, vol. 323, no. 3-5, pp. 677–696, 2009.

[118] J. McNames, "Fourier series analysis of epicyclic gearbox vibration," *Transaction of the ASME Journal of Vibration and Acoustics*, vol. 124, no. 1, pp. 150–152, 2002.

[119] M. R. Hoseini and M. J. Zuo, "Literature review for creating and quantifying faults in planetary gearboxes," tech. rep., Department of Mechanical Engineering, University of Alberta, Edmonton, Alberta, 2009.

[120] ASM, *Friction, Lubrication, and Wear Technology Handbook*. ASM International; 10th edition edition, 1992.

[121] Q. Hu, *Rough computation models and algorithms for knowledge discovery from heterogenous data*. PhD thesis, Harbin Institute of Technology, 2008.

[122] F. Stern, M. Muste, and M. L. Beninati, "Summary of experimental uncertainty assessment methodology with example," tech. rep., Lowa Institute of Hydraulic Research, College of Engineering The University of Iowa, Iowa City, 1992.

[123] Y. T. Su and S. J. Lin, "On initial fault detection of a tapered roller bearing: frequency domain analysis," *Journal of Sound and Vibration*, vol. 155, no. 1, pp. 75–84, 1992.

[124] "Warman technical bulletin," 1991, 1, 3.

[125] Q. Zheng and S. Liu, "Numerical investigation on the formation of jet-wake flow pattern in centrifugal compressor impeller," *American Society of Mechanical Engineers, International Gas Turbine Institute, Turbo Expo (Publication) IGTI*, vol. 6B, pp. 755–764, 2003.

[126] X. Zhao, M. J. Zuo, and R. Moghaddass, *Diagnostics and Prognostics of Engineering Systems: Methods and Techniques*, ch. Generating indicators for diagnosis of fault levels by integrating information from two or more sensors. IGI Global, 2012.

[127] F. Q. Wu and G. Meng, "Compound rub malfunctions feature extraction based on full-spectrum cascade analysis and SVM," *Mechanical Systems and Signal Processing*, vol. 20, pp. 2007–2021, 2006.

[128] T. H. Patel and A. K. Darpe, "Vibration response of a cracked rotor in presence of rotorstator rub," *Journal of Sound and Vibration*, vol. 317, no. 3-5, pp. 841–865, 2008.

[129] S. Zhang, M. Hodkiewicz, L. Ma, and J. Mathew, "Machinery condition prognosis using multivariate analysis," in *Engineering Asset Management* (J. Mathew, J. Kennedy, L. Ma, A. Tan, and D. Anderson, eds.), pp. 847–854, Springer London, 2006.

[130] C. Cempel, "Multidimensional condition monitoring of mechanical systems in operation," *Mechanical Systems and Signal Processing*, vol. 17, no. 6, pp. 1291–1303, 2003.

[131] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical mode decomposition as a filter bank," *IEEE Signal Processing Letters*, vol. 11, pp. 112–114, 2004.

[132] X. Fan and M. J. Zuo, "Machine fault feature extraction based on intrinsic mode functions," *Measurement Science and Technology*, vol. 19, no. 4, p. 045105, 2008.

[133] C. C. Lin, P. L. Liu, and P. L. Yeh, "Application of empirical mode decomposition in the impact-echo test," *NDT and E International*, vol. 42, pp. 589–598, 2009.

[134] Y. Lei and M. J. Zuo, "Fault diagnosis of rotating machinery using an improved HHT based on EEMD and sensitive IMFs," *Measurement Science and Technology*, vol. 20, no. 12, p. 125701, 2009.

[135] A. Kraskov, H. Stogbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, pp. 1–16, 2004.

[136] X. Fan and M. J. Zuo, "Gearbox fault detection using hilbert and wavelet packet transform," *Mechanical Systems and Signal Processing*, vol. 20, no. 4, pp. 966–982, 2006.

[137] F. Chu and W. Lu, "Experimental observation of nonlinear vibrations in a rub-impact rotor system," *Journal of Sound and Vibration*, vol. 283, no. 3-5, pp. 621–643, 2005.

[138] ANYSYS, *ANSYS CFX 12 Help Document*. ANSYS INC., 275 Technology Ddrive, Canonsburg, PA, April 2009.

[139] S. Greco, M. Inuiguchi, and R. SlowiÅski, "Dominance-based rough set approach using possibility and necessity measures," in *Rough Sets and Current Trends in Computing* (J. Alpigini, J. Peters, A. Skowron, and N. Zhong, eds.), vol. 2475 of *Lecture Notes in Computer Science*, pp. 84–85, Springer Berlin / Heidelberg, 2002.

[140] H. Natke and C. Cempel, "The symptom observation matrix for monitoring and diagnostics," *Journal of Sound and Vibration*, vol. 248, no. 4, pp. 597–620, 2001.

[141] V. F. Pires, J. Martins, and A. Pires, "Eigenvector/eigenvalue analysis of a 3D current referential fault detection and diagnosis of an induction motor," *Energy Conversion and Management*, vol. 51, no. 5, pp. 901–907, 2010.

[142] G. Turhan-Sayan, "Real time electromagnetic target classification using a novel feature extraction technique with pca-based fusion," *IEEE Transactions on Antennas and Propagation*, vol. 53, pp. 766–776, feb. 2005.

[143] Y. Pan, J. Chen, and X. Li, "Bearing performance degradation assessment based on lifting wavelet packet decomposition and fuzzy c-means," *Mechanical Systems and Signal Processing*, vol. 24, no. 2, pp. 559–566, 2010.

[144] H. Qiu, J. Lee, J. Lin, and G. Yu, "Robust performance degradation assessment methods for enhanced rolling element bearing prognostics," *Advanced Engineering Informatics*, vol. 17, no. 3-4, pp. 127–140, 2003.

[145] H. Ocak, K. A. Loparo, and F. M. Discenzo, "Online tracking of bearing wear using wavelet packet decomposition and probabilistic modeling: a method for bearing prognostics," *Journal of Sound and Vibration*, vol. 302, no. 4-5, pp. 951–961, 2007.

[146] C. Li and J. Limmer, "Model-based condition index for tracking gear wear and fatigue damage," *Wear*, vol. 241, no. 1, pp. 26–32, 2000.

[147] X. Zhao, M. J. Zuo, Z. Liu, and M. Hoseini, "Diagnosis of artificially created surface damage levels of planet gear teeth using ordinal ranking," *Measurement*, 2012 (in press).

[148] X. Geng, T. Y. Liu, T. Qin, and H. Li, "Feature selection for ranking," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 407–414, 2007.

[149] S. Baccianella, A. Esuli, and F. Sebastiani, "Feature selection for ordinal regression," in *Proceedings of the ACM Symposium on Applied Computing*, pp. 1748–1754, 2010.

[150] P. D. Samuel and D. J. Pines, "A review of vibration-based techniques for helicopter transmission diagnostics," *Journal of Sound and Vibration*, vol. 282, no. 1-2, pp. 475–508, 2005.

[151] S. Loutridis, "Gear failure prediction using multiscale local statistics," *Engineering Structures*, vol. 30, no. 5, pp. 1214–1223, 2008.

[152] J.Keller and P.Grabill, "Vibration monitoring of a UH-60A transmission planetary carrier fault," in *The American Helicopter Society 59th Annual Forum*, 2003.

[153] A. S. Sait and Y. I. Sharaf-Eldeen, "A review of gearbox condition monitoring based on vibration analysis techniques diagnostics and prognostics," in *Rotating Machinery, Structural Health Monitoring, Shock and Vibration, Volume 5* (T. Proulx, ed.), vol. 8 of *Conference Proceedings of the Society for Experimental Mechanics Series*, pp. 307–324, Springer New York, 2011.