25

Pattern Recognition based Fault Diagnosis in Industrial Processes: Review and Application

Thomas W. Rauber, Eduardo Mendel do Nascimento,
Estefhan D. Wandekokem and Flávio M. Varejão
Universidade Federal do Espírito Santo
Brazil

1. Introduction

The detection and diagnosis of faults in complex machinery is advantageous for economical and security reasons (Tavner et al., 2008). Recent progress in computational intelligence, sensor technology and computing performance permit the use of advanced systems to achieve this objective. Two principal approaches to the problem exist: model-based techniques and model-free techniques. The model-based line of research (Isermann, 2006; Simani et al., 2003) needs analytical model of the studied process, usually involving time dependent differential equations. One advantage is that the faults are an intrinsic part of the model. Deviations from the expected values are recorded in a residual vector which represents the state of health of the process. Frequently, the post-processing of the residual vector is approached by computational intelligence based techniques like statistical classifiers, artificial neural networks, and fuzzy logic. The use of these techniques however should not cause the impression that the classification of the process state is based solely on knowledge extracted from example data. An important drawback of model-based approaches is the necessity to establish an analytical model of the process which is a nontrivial problem. An experimental process setup in a controlled laboratory environment can be described by a mathematical model. Often the process is embedded in a control loop which naturally demands that inputs, controlled variables, and sensor outputs are modeled. In real-world processes the availability of an analytical model is often unrealistic or inaccurate due to the complexity of the process, so that false diagnosis can be caused by inappropriately designed models. Hence, the model-free techniques are an alternative method in case where an analytical model is not available.
In this chapter we describe model-free fault diagnosis in industrial process by pattern recognition techniques. We use the supervised learning paradigm (Bishop, 2007; Duda et al., 2001; Theodoridis & Koutroumbas, 2006) as the primal mechanism to automatically obtain a classifier of the process states. We will present a pattern recognition methodology developed for automatic processing of information and diagnostic decision making on industrial process. The fundamental drawback of the model-free approach is the necessity to provide a statistically significant number of labeled example data for each of the considered process classes. If only a small number of patterns are available in the training phase, the statistical classifiers might be misled and very sensitive to noise. Nevertheless, the extraction of knowledge about the process states principally from a set of example patterns has some attractive properties

and permits the application of well studied pattern recognition paradigms to this important problem of fault detection and diagnosis. Sometimes in the model-free approach a partial model of the faults is used (for instance in bearing fault diagnosis, where the expected frequency features are calculated from the specification of the bearing and the shaft frequency (Li et al., 2000)). This fact however should not mislead the reader that a model-based approach to fault diagnosis is used in that case. The central mechanism to describe a process situation is a d-dimensional feature vector $\mathbf{x} = (x_1 \cdots x_j \cdots x_d)^\mathsf{T}$ together with a class label ω_i. This feature vector is the result of an information processing pipeline depicted in Fig. 1. A similar information processing philosophy was proposed by Sun et al. (2004), however without any feature selection which we consider fundamental for an optimized performance of the fault diagnosis system. Some raw measurements delivered from sensors attached to the process can be immediately used as features x_j without any pre-processing. For instance a thermometer or manometer attached to some chemical reactor tank will provide the continuously valued temperature or pressure which can be passed to the next information processing stage without further treatment. The health of an electrically powered machine can often be characterized by its current consumption. In vibration analysis (Scheffer & Girdhar, 2004), the accelerometer is the main sensor and delivers displacement values $X(t)$ in the time domain which can mathematically be derived to velocity $\dot{X}(t)$ and acceleration $\ddot{X}(t)$ values. Usually the raw time domain signal is submitted to statistical processing (Samanta & Al-Balushi, 2003), Fourier transform (Li et al., 2000), wavelet coefficient extraction (Paya et al., 1997), envelope analysis (McFadden & Smith, 1984), or any other method that provides stationary values. This process is known as feature extraction on the measurement level. For instance, an accelerometer velocity signal $\dot{X}(t)$ of a motor transformed to the frequency domain by a Fourier transform produces frequency values $F(u)$ which can be considered extracted features at each of the frequencies u. This leads us to an essential problem of feature extraction, namely the production of large amounts of features. We need subsequent steps after the feature extraction on the measurement level to reduce the dimension of the finally used feature vector $\mathbf{x}$ to a reasonable size. This can be achieved again by feature extraction, this time on the information processing level, and finally by feature selection, to retain only a relatively small amount of features which additionally are the most discriminative ones. For instance, all frequencies u of a Fourier spectrum could be submitted to Principal Component Analysis (a linear feature extractor based on statistics) to reduce the dimensionality of the data (Jolliffe, 2002). Finally, the retained principal components could be fed to a feature selection algorithm in order to produce, for instance, a unique feature vector $\mathbf{x}$ to describe the whole process situation.
The basic strategy of obtaining a highly discriminative, low dimensional process state descriptor in the form of a single feature vector can be resumed in the following main steps:

1. Get as many raw measurements from as many sensors of a process as possible;
2. Submit the raw measurements to as many feature extraction techniques as possible, plausibly applicable to the specific problem;
3. Reduce the high dimensional data to only a few final features which simultaneously are the most discriminative descriptors of the process;
4. Induce a classifier for fault diagnosis.

The first item is usually restricted by the available sensors which can be used for data acquisition. For instance a motor pump produces electrical, acoustic, and vibrational patterns (Al Kazzaz & Singh, 2003) which all could be used if appropriate sensors are available. The second methodology opens up a huge variety of signal processing techniques which are only

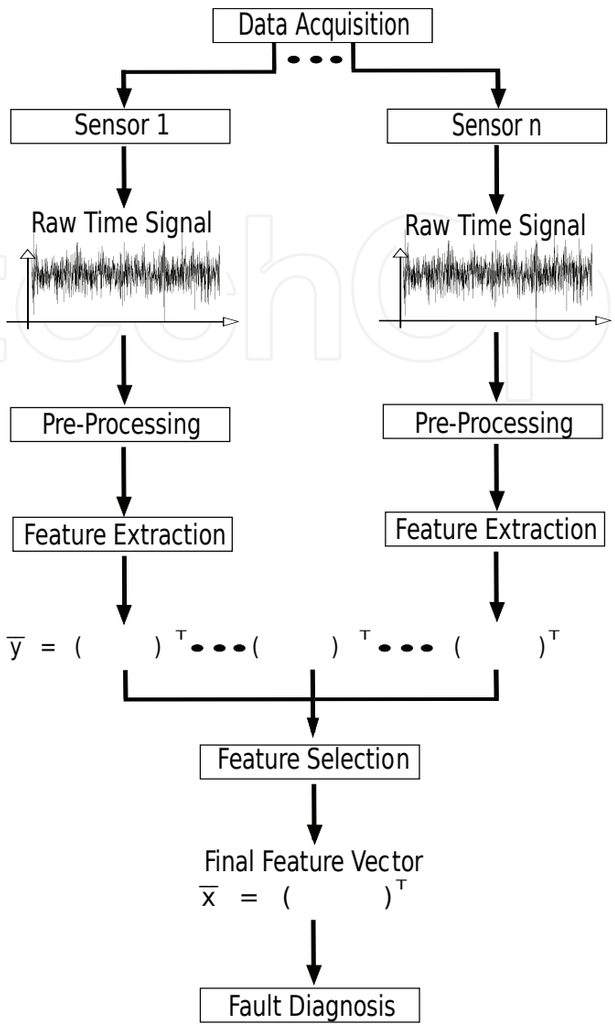


Fig. 1. Information processing pipeline to obtain a feature vector **x** which is of relatively low dimension and contains the most discriminative information about the state of a process.

restricted by the nature of the signal, the available software, and computing resources. From a vibration signal one might calculate wavelet coefficients (Paya et al., 1997), Fourier coefficients (Li et al., 2000), statistical measurements (Samanta & Al-Balushi, 2003), and frequencies obtained from envelope analysis (McFadden & Smith, 1984).

A review of some well-known processing techniques, such as statistical features in the time and frequency domain, to extract detailed information for induction machine diagnosis will be presented. Fourier methods to transform the time-domain signal to the frequency domain, where further analysis is carried out, is also discussed in this chapter. Using these techniques the state of machine can be constantly monitored and detailed analysis may be made concerning the health of the machine. The use of as many feature extraction methods as possible raises the chances that the most discriminative information is somehow captured by some of these features. If a prior restriction to only a limited set of features is defined, one might loose

valuable aspects hidden in the signal. Finally, we use feature selection techniques to emphasize the importance of each feature for the classification task and this information could be used for instance to retain the most discriminative information in a low dimensional vector.
The final stage of the construction of the monitor of the process condition is the training of a classifier, using labeled examples. Different strategies for the construction of the final classifier based on the result of the feature selection stage can be followed. The simplest one is to use the selected feature set as a filter for training and classification. More sophisticated techniques determine the final classification as the result of a multi-level decision process based on an *ensemble* of distinct classifiers (Duda et al., 2001). The goal is to elevate the performance of the final classifier relative to the individual classifier performances. The individual classifiers might differ for instance in the feature set that is used.
The chapter is organized in the following manner: Section 2 gives an overview of existing techniques to calculate features from raw signals. Then the calculus of new features from existing ones by information processing methods is approached in section 3. Special attention is given to the information filtering done by feature selection in section 4. When the feature model has been defined, we are interested in the expected quality of our fault classifier. This question is analyzed in section 5. As a practical benchmark of some of the presented techniques, section 6 illustrates their application to an interesting real-world, complex diagnosis task in the context of oil rig motor pump fault diagnosis. We investigate the described fault diagnosis methodology using pattern recognition techniques using real examples of rolling element bearing fault and misalignment fault of rotating machines. Final conclusions are drawn in section 7.

2. Measurement level feature extraction

Any conclusion about the condition of a process is based on the patterns which are obtained from sensorial devices attached to the dynamic system that in general constantly changes its state. Which sensors will be used should be considered as an integral part of the design of the whole diagnostic system. An ideal situation is a continuous on-line monitoring with many distinct sensors which deliver the data about the electrical, acoustic, and vibration activities. Usually this ideal situation is not encountered due to technical or budget restrictions or since the specificity of the application requires particular sensors. After preprocessing, the sensor patterns are available as digital information that can be processed by a specialized hardware or general purpose computer. A principal distinction is made with respect to the domain of the signal. The original continuous signal $s(t)$ in the time domain is discretized into n samples $s_1, \ldots, s_n$ that were acquired during a finite sampling interval. The number of samples depends on the duration of the acquisition and the sampling frequency. The Fourier transform provides the signal in the frequency domain. A mixed time-frequency domain is encountered when time dependent signals are processed by short-term Fourier transforms or wavelet transforms. In the following we compile some representative feature extraction methods that are widely used in the literature related to pattern-recognition based fault diagnosis.

2.1 Statistical features in the time domain

When we consider the original discretized time domain signal, some basic discriminative information can be extracted in the form of statistical parameters from the n samples $s_1, \ldots, s_n$ (Stefanoiu & Ionescu, 2006).

2.1.1 Root Mean Square (RMS)

One of the most important basic features that can be extracted directly from the time-domain signal is the RMS which describes the energy of the signal. It is defined as the square root of the average squared value of the signal and can also be called the *normalized energy* of the signal:

$$\mathrm{RMS} = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1} {s_i}^2}. \tag{1}$$

Especially in vibration analysis the RMS is used to perform fault *detection*, i.e. triggering an alarm, whenever the RMS surpasses a level that depends on the size of the machine, the nature of the signal (for instance velocity or acceleration), the position of the accelerometer, and so on. After the detection of the existence of a failure, fault *diagnosis* is performed relying on more sophisticated features. For instance the ISO 2372 (VDI 2056) norms define three different velocity RMS alarm levels for four different machine classes divided by power and foundations of the rotating machines.

2.1.2 Peak-to-Valley (PV) alias Peak-to-Peak (PP)

Another important measurement of a signal, considering a semantically coherent sampling interval, for instance a fixed-length interval or one period of a rotation, is the peak-to-valley value which reflects the amplitude spread of a signal:

$$\mathrm{PV} = \frac{1}{2}\left(\max_{i=0}^{n-1} s_i - \min_{i=0}^{n-1} s_i\right). \tag{2}$$

2.1.3 Peak

If we consider only the maximum amplitude relative to zero $s_{ref} = 0$ or a general reference level s_{ref}, we get the peak value:

$$\mathrm{peak} = \max_{i=0}^{n-1} s_i - s_{ref}. \tag{3}$$

Often the peak is used in conjunction with other statistical parameters, for instance the peak-to-average $\left(\mathrm{peak} \Big/ \frac{1}{n}\sum_{i=0}^{n-1} s_i\right)$ or peak-to-median $\left(\mathrm{peak} \Big/ \underset{i=0}{\overset{n-1}{median}}\, s_i\right)$ rates (Ericsson et al., 2005).

2.1.4 Crest factor

When we relate the peak value to the RMS of the signal, we obtain the crest factor:

$$\mathrm{CF} = \mathrm{peak}/\mathrm{RMS}, \tag{4}$$

which expresses the *spikiness* of the signal. The crest factor is also known as peak-to-average ratio or peak-to-average power ratio and is used to characterize signals containing repetitive impulses in addition to a lower level continuous signal. The modulus of the signal should be used in the calculus.

2.1.5 Kurtosis

The analytic definition of the kurtosis is $\kappa = -3 + \mu_4/\sigma^4$ with $\mu_4 = E\{(s - E\{s\})^4\}$ being the fourth moment around the mean and σ^4 being the square of the variance. Considering the n samples $s_1, \ldots, s_n$, we determine the sample kurtosis as:

$$\kappa = -3 + \frac{\frac{1}{n}\sum_{i=0}^{n-1}(s_i - \bar{s})^4}{\left[\frac{1}{n}\sum_{i=0}^{n-1}(s_i - \bar{s})^2\right]^2}, \tag{5}$$

where $\bar{s}$ denotes the estimated expected value of the signal (average). The kurtosis expresses an aspect of *spikiness* of the signal, although in a higher order than the crest factor, and describes how peaked or float the distribution is. If a signal contains sharp peaks with a higher value, then its distribution function will be sharper.

2.1.6 Further statistical parameters

Besides the RMS, variance, and kurtosis, Samanta & Al-Balushi (2003) further present the skewness (normalized third central moment) $\gamma_3 = \mu_3/\sigma^3$ and the normalized sixth central moment $\gamma_6 = \mu_6/\sigma^6$ as statistical features in bearing fault detection. In the context of gearbox fault detection, Večeř et al. (2005) describe further statistical features frequently used as condition indicators: the energy operator

$$\mathrm{EO} = \frac{n^2\sum_{i=0}^{n-1}\left((s_{i+1}^2 - s_i^2) - \bar{s}\right)^4}{\left[\sum_{i=0}^{n-1}\left((s_{i+1}^2 - s_i^2) - \bar{s}\right)^2\right]^2}\ ; \tag{6}$$

energy ratio, that is, ratio of the standard deviations of the difference signal and the raw signal

$$\mathrm{ER} = \sigma(d)/\sigma(s), \tag{7}$$

where difference signal d is defined as the remainder of the vibration signal after the regular meshing components are removed; sideband level factor (sum of the first order sideband about the fundamental gear mesh frequency divided by the standard deviation of the time signal average), sideband index (average amplitude of the sidebands of the fundamental gear mesh frequency); zero-order figure of merit

$$\mathrm{FM0} = \mathrm{PV}\Big/\sum_{i=0}^{M} A_i, \tag{8}$$

where A_i is the amplitude of the i-th gear mesh frequency harmonics; kurtosis of the differential signal (FM4 parameter); kurtosis of the residual signal (a synchronous averaged signal without the gear mesh frequency, its harmonics, drive shaft frequency, and its second harmonics) or envelope normalized by an average variance (NA4 parameter and NB4 parameter). The reader interested in more details is referred to (Lei & Zuo, 2009; Večeř et al., 2005).

2.2 Statistical features in the frequency domain

Especially when leading with signals from processes that produce periodic signals, like rotating machinery, the Fourier transform (Bracewell, 1986) of a one-dimensional signal is an information conserving fundamental analytic functional $\mathscr{F}\{s(t)\} = F(u)$ that decomposes the signal into additive sine and cosine terms in the complex domain, i.e. $F(u) = \Re(u) + j\Im(u)$. The phase angle information $\phi(u) = \arctan(\Im(u)/\Re(u))$ is a valuable source of information when for instance the relative movement of different parts of a rotating machine at a given frequency u should be analyzed (Scheffer & Girdhar, 2004). The vast majority of analytic work is however done with the magnitude $|F(u)| = \sqrt{\Re^2(u) + \Im^2(u)}$ of the Fourier transform, also known as the Fourier spectrum or generally the frequency spectrum or simply the spectrum of $s(t)$. The Fourier spectrum will be symmetric and hence only one half has to be kept. Usually the discrete signal buffer of $n\prime$ samples is interpolated to a length n which is a power of two in order to be able to apply the Fast Fourier Transform algorithm of complexity $O(n \log n)$.
The spectrum constitutes a new discrete signal of $n/2$ samples $f_1, \ldots, f_{n/2}$ which serves as the basis to extract more features. For the sake of simplicity we once again presume that we have n samples, instead of $n/2$. The Root Mean Square (RMS) can also be calculated in the frequency domain. Often we are interested on RMS of particular bands interest, for instance the bands around the harmonics in rotating machinery, i.e. the multiples of the fundamental shaft rotation frequency. In order to calculate the features of specific bands, the interval of the particular spectra has to be considered, either as absolute intervals or as percentages (for instance 2% of the frequency value to lower and higher frequencies).
It should be clear that in terms of statistical pattern recognition, the features x_j can assume any of the numerical descriptors that can be obtained from the frequency domain, such as: single frequencies, RMS of bands or the whole signal, and all conceivable functional mappings of the signal. We do not question the information content of the obtained features at this moment. Surely, some features will contain much more valuable discriminative power than others. Some features may even contaminate the descriptive behavior of the feature vector that describes that condition of the process. The information filtering will later be done by the feature selection step.
Especially for the analysis of acoustic signal which might for instance represent the noise emissions caused by a fault, the *Cepstrum* is the inverse Fourier transform of the logarithm of the magnitude of the Fourier transform of signal s, i.e. $\mathscr{F}^{-1}\{\log(|\mathscr{F}\{s(t)\}|)\}$, (Theodoridis & Koutroumbas, 2006).

2.3 Time-frequency domain features

There are analytic techniques that try to capture frequency content at different time instances, as it were a hybrid representation of the signal regarding the changing intensities over time and simultaneously looking at repetitive patterns during limited time intervals because the signal is non-stationary. In fault diagnosis there certainly exist fault signatures that respond adequately to these techniques, for instance the brush seizing faults in a DC servo motor described by Sejdić & Jiang (2008).

2.3.1 Short-time Fourier transform (STFT)

When we multiply a signal with a finite window function and take the Fourier transform of the product, we calculate the STFT. For instance, Al Kazzaz & Singh (2003) used the STFT to obtain the spectra of overlapping signal segments of a vibration signal and then averaged

the set of spectra in order to reduce random vibration and the noise in the signal. When the window function is a Gaussian, we obtain the Gabor transform (Gabor, 1946).

2.3.2 Wavelet transform

The wavelet transform (Chui, 1992) has the advantage of a flexible resolution in the time and frequency domains when compared to the short-time Fourier transform which has a fixed window size. Like in the case of the Fourier transform, we can distinguish among continuous wavelet transform, wavelet series expansion, and the discrete wavelet transform. The latter is used in software implementation in digital signal processing to obtain the wavelet coefficients which describe the signal in terms of multiples of wavelet basis function. An important distinction is between orthogonal and non-orthogonal basis functions. The dyadic wavelet family furthermore facilitates the efficient implementation of the discrete wavelet transforms, since the translations and scales are powers of two. In the majority of research work, the orthogonal dyadic wavelet basis is used (Loparo & Lou, 2004; Paya et al., 1997; Singh & Al Kazzaz, 2004; 2008). A good introduction to the theory of wavelets with an application of gearbox fault detection can be found in (Wang & McFadden, 1996). In that application non-orthogonal wavelet basis functions are used to capture the transients of a fault signal, justified by two drawbacks of the orthogonal bases, namely not having enough scales to describe the scaled versions of the transients and the different calculated coefficients describing the same transients at different time instants. We restrict the definition to the discrete orthogonal dyadic wavelet transform (Castleman, 1995). Given a basis function, or mother wavelet $\psi(s)$ of the signal function s, the set of functions which are scaled and translated versions of the mother wavelet $\psi_{l,k}(s) = 2^{l/2}\psi_{l,k}\left(2^l s - k\right)$ that form an orthonormal basis of $L^2(\mathbb{R})$, with $-\infty < l,k < \infty$, $l,k \in \mathbb{N}$, are the dyadic orthogonal wavelet functions. The coefficients of the discrete dyadic wavelet transform can be obtained as:

$$c_{l,k}(s) = \sum_{i=0}^{n-1} s(i\Delta t)\psi_{l,k}(i\Delta t), \tag{9}$$

where the signal s is sampled at n discrete instances at intervals Δt and $l = 0, 1, \ldots, \log_2 n - 1$, $k = 0, 1, \ldots, 2^l - 1$. In the context of signal analysis this transform can also be viewed as a multiresolution recursive filter bank for different scaled and translated versions of the same signal signature. Dyadic wavelets that furthermore have a compact support are for instance the Haar or Daubechies wavelets (Castleman, 1995).

2.3.3 Other time-frequency analysis techniques

Recently, Yan et al. (2009) have presented the Frequency Slice Wavelet Transform (FSWT). It constitutes a parameterized generalization, and can be specialized into the Fourier transform, the Gabor transform, the Morlet wavelet transform (Goupillaud et al., 1984), and the Wigner-Ville distribution (Wigner, 1932).

2.4 Final Considerations of feature extraction on the measurement level

It should be clear that from a raw signal a variety of features can be extracted, starting from simple statistical parameters, like the RMS, until sophisticated mathematical transforms. The envisaged application is of course the main motivation to use a certain feature extractor or another. On the other hand, we could adopt a strategy to obtain a great variety of new, possibly useful information from the original signal, sending it to a battery of feature extractors. Why

should it not be reasonable to use some statistical features from the time domain together with a few wavelet coefficients, possibly from different wavelet types, then further joining some RMS value from several different bands of a frequency spectrum ? One could argue that this produces a huge amount of features, possibly worthless for diagnostic purposes, introducing features that contaminate the valuable parameters of our system. The answer to this objection is the use of subsequent information filtering steps in the processing pipeline. Later we will describe the use of feature selection to retain only the most useful features that finally characterize the process states. For now we can think of merging all available features produced by the procedures described above in to a new feature vector $\mathbf{x} := (x_1 \ \cdots \ x_j \ \cdots \ x_d)^\mathsf{T}$ that will be processed by the next information processing methods.

3. High level feature extraction

In the previous section we gave a slight overview for methods which can calculate features from a raw signal acquired from a technical process with the intention to use it for fault diagnosis. There exist a series of methods which take existing feature vectors and transform them into other feature vectors eventually reducing the dimension of the original descriptor and/or improving its discriminative behavior with respect to the fault classes. One principal distinction can be made between linear and non-linear methods.

3.1 Linear methods

When the components y_l, $l = 1, \ldots d'$ of the new feature vector $\mathbf{y}$ are all linear combinations $y_l = \sum_{j=1}^{d} m_{lj} x_j = \mathbf{m}_l^\mathsf{T} \mathbf{x}$ of the components x_j of the original feature vector $\mathbf{x}$, then we obtain a linear feature extractor. The m_{lj} are the real valued elements of a matrix $\mathbf{M}$ that implements the extraction easily as $\mathbf{y} = \mathbf{M}^\mathsf{T} \mathbf{x}$. Linear methods have the advantage that they are mathematically tractable by well studied applied linear Algebra. Linear feature extraction is an unsupervised technique, working without the knowledge to which class a pattern belongs. This can easily destroy discriminative information. For instance in Principal Component Analysis it can happen that the variance in a newly extracted component is higher than in another, ranking it before the lower variance component, although the class separability in the second less important component might be better.

3.1.1 Principal Component Analysis

Principal Component Analysis (PCA) (Jolliffe, 2002) is the most well studied technique to map the existing feature vectors $\mathbf{x}$ to linearly uncorrelated and lower dimensional feature vectors $\mathbf{y}$, eventually sacrificing new components with a small variance. The first two or three principal components additionally can be visualized, exposing the mutual relationship of the patterns. From the total of n d-dimensional pattern samples $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ that describe process situations we estimate the mean $\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}^{(k)}$ and covariance matrix $\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1} \sum_{k=1}^{n} (\mathbf{x}^{(k)} - \hat{\boldsymbol{\mu}})(\mathbf{x}^{(k)} - \hat{\boldsymbol{\mu}})^\mathsf{T}$. The symmetric $d \times d$ matrix $\hat{\boldsymbol{\Sigma}}$ is then submitted to an eigenanalysis which delivers d eigenvalues λ_j and the corresponding eigenvectors $\boldsymbol{\phi}_j$ which are ordered following a descending order of the corresponding eigenvalues to form the feature extractor $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_d)$. If we extract $\mathbf{y} = \boldsymbol{\Phi}^\mathsf{T} \mathbf{x}$ we obtain a linearly uncorrelated feature vector $\mathbf{y}$ of the same dimension d. If we delete the columns $d'+1, \ldots, d$ from $\boldsymbol{\Phi}$ with $1 \leq d' < d$, we obtain the $d \times d'$ matrix $\boldsymbol{\Phi}'$ that extracts the d' principal components as $\mathbf{y}' = \boldsymbol{\Phi}'^\mathsf{T} \mathbf{x}$. The approximation error $\mathcal{E}$ committed by discarding the $d - d'$ low variance components is $\mathcal{E}(d') = \sum_{j=d'}^{d} \lambda_j$. Synonymous for PCA are Karhunen-Loève Transform (KLT),

Hotelling transform, or Proper Orthogonal Decomposition (POD). When the data has been centered to its global mean and all components are preserved, PCA is also equivalent to Singular Value Decomposition (SVD).

3.1.2 Independent Component Analysis

Another linear feature extractor is Independent Component Analysis (ICA) (Hyärinen et al., 2001) which has the task to separate independent signals from mixed signals. This could be interesting to recognize the contributions of different faults in a signal, or find the latent independent variables that mix together to the observable variables (features) in a process. Again, from the original feature vector $\mathbf{x}$ a new feature vector $\mathbf{s}$ is extracted by a linear transform (matrix multiplication) as $\mathbf{s} = \mathbf{W}\mathbf{x}$, where the extractor $\mathbf{W}$ is a $d' \times d$ matrix, called the demixing matrix. The d' independent components s_j are maximized with respect to their non-Gaussianity, an information based criterion for independence, often measured by kurtosis or negentropy. A few representative applications of ICA in the context of fault diagnosis are for instance (Jiang & Wang, 2004; Lee et al., 2006; Pöyhönen et al., 2003).

3.2 Non-linear methods

Any non-linear mapping $\mathbf{y} = \mathbf{\Phi}(\mathbf{x})$, where the original d-dimensional feature vector $\mathbf{x}$ is transformed to the H-dimensional extracted feature vector $\mathbf{y}$ can be considered a non-linear generator of new features that might be more discriminative than the original information. As an example take the classical XOR problem, linearly not separable, where the mapping

$$\mathbf{\Phi}\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} \phi_1\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) \\ \phi_2\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) \\ \phi_3\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{pmatrix} \tag{10}$$

from the original bi-dimensional space enables linear separability in the mapped tri-dimensional space. When a classifier $g(\mathbf{x}) = \mathbf{w}^{\mathsf{T}}\mathbf{\Phi}(\mathbf{x})$ is a linear combination of the basis functions ϕ_h we deal with a Generalized Linear Discriminant Function (GLDF) (Duda et al., 2001). A great variety of feature extraction techniques falls in this category, for instance polynomial combinations of the original features, radial basis functions, Multilayer Perceptrons (when omitting the activation function in the output layer), or any other calculus of new features not definable as a matrix multiplication of a linear extraction matrix with the original feature vector (Theodoridis & Koutroumbas, 2006). When we need only a similarity measure between two mapped feature vectors $\mathbf{\Phi}(\mathbf{x})$ and $\mathbf{\Phi}(\mathbf{x'})$ without having to define the mapping explicitly, but only its dot product, we define a *kernel* $k(\mathbf{x}, \mathbf{x}')$. Examples are the polynomial mapping $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + 1)^r$ calculating all monomials up to degree r and the Gaussian radial basis function $k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma||\mathbf{x} - \mathbf{x}'||^2)$ with shape parameter γ. These kernels are especially needed in Support Vector Classification and Regression (Theodoridis & Koutroumbas, 2006).

If we solve a regression problem by replicating the input vector $\mathbf{x}$ as the target vector in a Perceptron with one hidden layer, and if the number of neurons H in the hidden layer is smaller than the dimension d of $\mathbf{x}$, we have an auto-associative feedforward neural network, applied for example by (Skitt et al., 1993) in the context of machine condition monitoring. The

extracted feature vector is composed by the components $y_h = z(\mathbf{w}_h^T \mathbf{x})$, $h = 1, \ldots, H$, where $z(.)$ is the activation function, usually the logistic function:

$$z(a) = \frac{1}{1 + e^{-a}} \tag{11}$$

or the hyperbolic tangent:

$$z(a) = \tanh a = \frac{e^a - e^{-a}}{e^a + e^{-a}}. \tag{12}$$

Kohonen's Self-Organizing Map (SOM) (Kohonen, 1998) organizes neurons in a 1-D, 2-D, 3-D, or higher-dimensional topological map. After the map is trained there is a single responding neuron (the winner) inside the maps lattice. The discrete position could be taken as an extracted feature vector. High-dimensional patterns can be mapped to a lower dimensional map by the Sammon plot (Sammon Jr., 1969) which tries to preserve the mutual distances among all patterns in the original and mapped pattern space. Although the mapping is principally conceived for visualization in two or three dimensions, theoretically any dimension lower than the original d can be chosen for feature extraction. The auto-associative feedforward neural network, Kohonen map, and Sammon map have been applied to non-linear feature extraction by (De Backer et al., 1998), besides a collection of methods called multidimensional scaling.

4. Feature selection

In the previous two sections we gave a slight overview of methods which can calculate features from a raw signal acquired from a technical process with the intention to use it for fault diagnosis. There exist a series of methods which take existing feature vectors and transform them into other feature vectors eventually reducing the dimension of the original descriptor and/or improving its discriminative behavior with respect to the fault classes. This was called high level feature extraction. On the other hand, we can filter out some of the existing features and retaining others, such forming again a new feature vector, without modifying the original individual features. This is feature selection. This stage is of great importance because the feature extraction by several distinct extraction methods probably generates a large quantity of features that can have a marginal importance for classification or even jeopardize the classifier performance if they are used. Hence, the objective of this stage is to express the importance of the features in the classification task. The quality criterion can be individual, but the multidimensional nature of the process descriptors demands the analysis of feature sets, because of the interdependency relations within a feature set.

An early excellent compilation of feature selection techniques is given by Devijver & Kittler (1982). Guyon & Elisseeff (2003) give an introductory treatment about this subject in the field of Machine Learning. A paper collection about computational methods of feature selection is Liu & Motoda (2007). A feature selection algorithm is basically composed of two ingredients, namely a search strategy algorithm and a selection criterion. The search strategy is about which of the feature subsets are analyzed and the selection criterion associates a numerical value to each of these subsets, thus permitting the search for the subset that maximizes the criterion.

Kudo & Sklansky (2000) do an extensive experimental comparison on standard machine learning databases with basically all existing algorithms at the time of the publication. The simplest evaluation of a quality criterion of features is by individually calculate the criterion and then rank the whole set of available features. This strategy of course ignore completely the multidimensional nature of the descriptors of the process. Nevertheless, when the original feature set

is very large, it is a useful preprocessing step. If, for instance, the classification performance of a single feature is not better than random classification, this suggests that it contains no information at all and can be discarded prior to a more sophisticated search. We call this search strategy Best Features (BF).
Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) are greedy algorithms that add or delete a feature at a time without permission to later on eliminate or join it again relative to the currently selected feature set. Their generalized version that permit an exhaustive search within a small candidate set added or deleted at a time are Generalized Sequential Forward Selection (GSFS) and Generalized Sequential Backward Selection (GSBS). The Plus-*L*-Take Away-*R* (PLTR) strategy permits backtracking for a fixed number of steps, thus allowing to eliminate previously selected features or joining previously deleted features. When allowing backtracking as long as it improves the quality criterion of the currently selected feature set, we have the floating versions of PLTR, called Sequential Forward Floating Selection (SFFS) and Sequential Backward Floating Selection (SBFS) (Pudil et al., 1994).
When the criterion always increases monotonically when joining an additional feature, we are able to apply the Branch-and-Bound (BB) strategies. The most prominent algorithm is the original of Narendra & Fukunaga (1977) which was refined later to the improved BB (Yu & Yuan, 1993) and relaxed BB methods (Kudo & Sklansky, 2000). The biggest drawback of BB methods is that the most natural selection criterion is the error rate and this is *not* monotonically increasing or its complement decreasing. Besides, even with the exhaustive implicit visit of all leaves of the search tree, the algorithm is computationally still expensive. Another line of research for search strategies are Genetic Algorithms (GA) (Estevez et al., 2009; Oh et al., 2004; Pernkopf & O'Leary, 2001). The basic idea is to combine different feature subsets by a crossover procedure guided by a fitness criterion. The advantage of GAs for feature selection is the generation of quite heterogeneous feature sets that are not sequentially produced.
Each feature selection subset search strategy must define a criterion J which defines the quality of the set. When a classification or regression task will be performed based on the features, and the very performance of the system is used as the quality criterion, the whole feature selection is called a *wrapper*. The most common wrapper methods use the estimated classification accuracy of the subsequent classifier as the quality criterion. This implies that a classification architecture and an error estimation algorithm must be defined. For instance a 10-fold cross validation together with a Support Vector Machine could be used. A regression wrapper would probably take the expected approximation error as J. A *filter* selection algorithm defines the selected subset before a regression of classification is done by the set. Selection criteria related to filters can principally be grouped into interclass distances, probabilistic distances and information-based criteria (Devijver & Kittler, 1982). A more detailed description for search strategies and selection criteria is given in the following sections.

4.1 Search algorithms in feature selection

If we want to select d features $X_d = \{x_{k_i} | i = 1, \ldots, d; k_i \in \{1, \ldots, D\}\}$, from an available pool of D features $Y = \{x_j | j = 1, \ldots, D\}$, an exhaustive search would take $\binom{D}{d}$ iterations which is computationally unfeasible for even moderate number of features. So we have to rely on suboptimal search strategies. Let us suppose that we a have a selection criterion J that is able to evaluate the quality of a candidate set Ξ_k, composed of k features. We are interested in that candidate set of cardinality k that maximizes the criterion, hence we are looking for the set X_k with

$$J(X_k) = \max_{\{\Xi_k\}} J(\Xi_k). \tag{13}$$

For $k = d$ we obtain a satisfactory feature set that maximizes the selection criterion for d features.

Best Features (BF) is simply evaluating $J(\{x_j\})$ for each feature $j = 1, \ldots, D$, ordering the features in descending order relative to J and setting the selected set X_d to the first d features of the ordered set. This mechanism ignores the multidimensionality of the problem, but on the other hand in an $O(D)$ complexity selects a feature set. As it was already mentioned the application of BF is recommended as a preprocessing step if an extremely large number D of features is available.

Sequential Forward Selection (SFS) (Devijver & Kittler, 1982) starts with an empty set. Consider that k features have already been selected by SFS which are included in the feature set X_k. If Y is the total set of all D features $Y \setminus X_k$ is the set of $D - k$ candidates ξ_j. Test each candidate together with the already selected features and rank them following the criterion J, so that $J(X_k \cup \{\xi_1\}) \geq J(X_k \cup \{\xi_2\}) \geq \ldots \geq J(X_k \cup \{\xi_{D-k}\})$. Then the updated selected feature set is given as $X_{k+1} = X_k \cup \{\xi_1\}$. The algorithm is initialized by $X_0 = \emptyset$ and stops at X_d. Although SFS is an algorithm that considers mutual dependencies among the involved features, it has the main drawback of not allowing the posterior elimination of a feature ξ_j, once it has been selected.

Sequential Backward Selection (SBS) (Devijver & Kittler, 1982) which starts with all features Y as being selected and then discards one feature at a time, until $D - d$ features have been deleted, i.e. d features have been retained as the selected ones. Consider that k features have already been discarded from $\bar{X}_0 = Y$ to form feature set $\bar{X}_k$. In order to obtain the feature set with one more feature discarded, rank the features ξ_j contained in set $\bar{X}_k$, so that $J(\bar{X}_k \setminus \{\xi_1\}) \geq J(\bar{X}_k \setminus \{\xi_2\}) \geq \ldots \geq J(\bar{X}_k \setminus \{\xi_{D-k}\})$. Then the updated selected feature set is given as $\bar{X}_{k+1} = \bar{X}_k \setminus \{\xi_1\}$, i.e. the worst feature ξ_1 is discarded. The SBS strategy is computationally more demanding than the SFS algorithm, since the criterion J has to be evaluated with generally more features. As in the case of SFS, the SBS does not allow the posterior reintegration of a feature ξ_j, once it has been discarded.

The Plus L-Take Away R selection algorithm (Devijver & Kittler, 1982; Pudil et al., 1994) tries to overcome the drawbacks of the SFS and SBS methods by allowing to discard a feature that has already been selected and reintegrate a feature into the selected pool after it has been discarded thus avoiding the nested nature of the sets chosen by SFS and SBS. If the parameter L is greater than R then the we start with the SFS algorithm to join L features to the selected pool. Then R features are discarded by the SBS procedure to get a set X_{L-R}. We repeat the joining L features by SFS and discarding R features by SBS until d features have been reached in X_d. The parameters L and R must appropriately be chosen to match d and not overreach the minimum 0 and maximum D. If $L < R$ then we start with the whole feature set Y and the SBS algorithm to first discard R features. Then L features are joined again by SFS. The SBS-SFS pair is repeated until reaching d features. PLTR in contrast to the greedy forward and backward algorithms SFS and SBS allows a backtracking step, although with a fixed size that can discard already selected or include already discarded features.

If we allow the backtracking for an arbitrary number of times as long as the quality criterion J is improving, we arrive at the *floating* techniques (Pudil et al., 1994). As a representative for a complete sequential search strategy algorithm, we present the Sequential Forward Floating Search (SFFS) in algorithm 1. The second condition in line 9 is a very simple mechanism to avoid looping. It remembers the last included feature by SFS and does not allow the immediate exclusion by SBS. More sophisticated looping prevention techniques could be conceived.

Algorithm 1 is easily simplified to the SFS algorithm by omitting the inner loop between line 6 and line 14.

Algorithm 1 SFFS

Input: A set Y of D available features $Y = \{x_j | j = 1, \ldots, D\}$, the number of desired features d.

Output: The feature set X with cardinality $|X| = d$ that maximizes the selection criterion.

```
1: Select one feature x_SFS from Y by Sequential Forward Selection (SFS)
2: X ← {x_SFS}; Y ← Y \ {x_SFS};
3: repeat
4:     {Select one feature x_SFS from Y by SFS}
5:     X ← X ∪ {x_SFS}; Y ← Y \ {x_SFS};
6:     repeat
7:         conditional_Exclusion ← false;
8:         Determine best candidate for exclusion x_SBS from X using SBS
9:         if J(X \ {x_SBS}) > J(X) AND x_SBS not included in the last SFS step then
10:            conditional_Exclusion ← true;
11:            {Excluding the feature improves criterion J}
12:            X ← X \ {x_SBS}; Y ← Y ∪ {x_SBS};
13:        end if
14:    until ( NOT conditional_Exclusion OR |X| = 1 )
15: until |X| = d
```

All search algorithms previously presented in this section are based on the idea of building a unique feature set that presents an optimal classification performance, and in each iteration the set is improved by the insertion or removal of some features. The output of such method leads naturally to the idea of a order of importance (ranking) among the selected features. However, one may be interested in obtaining several potentially good feature sets, in order to assign the final classification label after the inspection of the performance of each of those sets, rather than relying on a unique set. *Genetic Algorithms* (GA) (Opitz, 1999) can naturally be used to meet this requirement. They are is inspired by Darwin's biological natural selection, in which different individuals combine and compete among themselves in order to transmit their winner genes to future generations. In GA-based feature selection, each individual is a feature set, and their individual quality (fitness) can be measured, for example, as the error rate of a classifier built on the corresponding features. If $|Y|$ is the cardinality of the global pool of features to be selected, a natural way of representing an individual is as a binary string composed of $|Y|$ bits, so that the value 1 in the i-th bit indicates the presence of the i-th feature in the feature set represented by that individual, and the value 0 indicates its absence. The algorithm starts with the creation of the initial individuals (by a random or heuristic-based method), and in each generation, the best individuals are more likely to be selected to transmit their bits (genes) to future generations. This transmission involves the combination of the bits of each "parent" in order to create a "child", which will have, for instance, the first half of its bits coming from the first parent, and the second half coming from the second parent. To increase the diversity of individuals, the mutation operator can be used, so that each bit has a small probability of being flipped. Finally, after the passage of several generations, the population will be composed of distinct individuals with a high fitness value, or equivalently several

feature sets that possibly present good classification results and diversity among themselves with respect to the chosen features.

4.2 Selection criteria in feature selection

The previously defined selection algorithms all need to define a multivariate selection criterion J which judges the quality of the candidate feature set. One of the obvious choices in a classification context is to choose those features that minimize the error of the classifier, thus creating the basic wrapper feature selector. Since we do not have any parametric description of the class specific stochastic processes, we have to estimate the error rate. To do that we need first to define the classifier architecture and then the error estimation method. As a simple rule of thumb one can use the 1-Nearest Neighbor classifier together with the Leave-One-Out error estimation procedure (Devijver & Kittler, 1982; Duda et al., 2001; Theodoridis & Koutroumbas, 2006). The use of a multilayer perceptron as the classifier together with a cross-validation error estimation method is computationally unfeasible, since during selection, for discarding of a single feature or joining another, we would have to completely train the network and estimate its classification accuracy.

Filter criteria $J(X)$ of a feature set X allow to obtain the selected feature set before a regression or classification is done. One can expect that there is a strong correlation between a high value of a filter criterion and a good regressor or classifier. For instance when we have two classes and the distances among the samples of the same class (intraclass distance) is low and the distances between each sample of one class and the samples of the other class are high (interclass distance), we can expect a low error rate of a classifier.

For the definition of an interclass feature selection criterion, one has to define first a metric $\delta(x_i^{(k)}, x_j^{(l)})$ between a sample $x^{(k)}$ which belongs to class ω_i and sample $x^{(l)}$ which belongs to class ω_j. Usually the Euclidean distance $||x_i^{(k)} - x_j^{(l)}||$ is used. Other choices are Minkowski of order s as $\left[\sum_{f=1}^{F} |x_{if}^{(k)} - x_{jf}^{(l)}|^s\right]^{1/s}$, City Block (i.e. Minkowski of order 1), Chebychev $\max_f |x_{if}^{(k)} - x_{jf}^{(l)}|$ or some other nonlinear metric. The index $f = 1, \ldots, F$ is over the current candidate set X. The selection criterion is then the average interclass distance among all data points of all classes $J(X) = \frac{1}{2n^2} \sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \delta(x_i^{(k)}, x_j^{(l)})$.

Consider two probability density functions $p_a(\mathbf{x}; \boldsymbol{\theta}_a)$ and $p_b(\mathbf{x}; \boldsymbol{\theta}_b)$ of a d-dimensional continuous random variable $\mathbf{x}$, defined by their functional forms p_a, p_b and parameter vectors $\boldsymbol{\theta}_a, \boldsymbol{\theta}_b$ respectively. A probabilistic distance measure J between the two probability density functions is a functional that measures the difference Δ integrated over the domain $\mathbb{R}^d$ of $\mathbf{x}$:

$$J(p_a, p_b, \boldsymbol{\theta}_a, \boldsymbol{\theta}_b) = \int_{\mathbf{x}} \Delta[(p_a, p_b, \boldsymbol{\theta}_a, \boldsymbol{\theta}_b)] d\mathbf{x}. \tag{14}$$

The metric Δ should be positive, zero if the values of the two functions coincide, and correlated to the their absolute difference (Devijver & Kittler, 1982). In the context of classification the a priori probabilities $P_a = Pr[\mathbf{x} \in \omega_a]$, $P_b = Pr[\mathbf{x} \in \omega_b]$ for the two classes ω_a, ω_b can additionally be incorporated into the probabilistic distance, hence in this case we have $J = J(P_a, P_b, p_a, p_b, \boldsymbol{\theta}_a, \boldsymbol{\theta}_b)$. In general J is defined for probability density functions (pdf) that could come from distinct functional families, for instance a univariate Normal distribution $p_a(x; \mu, \sigma^2)$ and a Gamma distribution $p_b(x; k, \theta)$. In practice however only pdfs with the same functional form $p_a = p_b$ are compared. Hence the pdf p under consideration becomes implicit and the functional forms are dropped from the argument list, so we have a function

of the parameters $J(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b)$. For instance the distance between two Gaussians in the univariate case with their means and variances given is $J([\ \mu_a \quad \sigma_a^2\ ], [\ \mu_b \quad \sigma_b^2\ ])$. The probabilistic distance between two multivariate Gaussian is the best studied distance. Assuming that the data obey this distribution one can obtain closed form distance measure based only on the means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$ of the classes ω_i. For an overview of probabilistic distances see for instance Rauber et al. (2008).
Information based criteria for feature selection, especially mutual information was applied by Estevez et al. (2009). The idea is to measure the increment of information that a feature produces and rank the feature candidate sets following this information gain.

5. Performance estimation

A very important step in the design of a fault diagnosis system that is based on supervised learning of an automatic classifier is to estimate the quality of the resulting diagnosis system. Once again, like in the case of feature selection we can devise performance criteria and cross validation algorithms that give us an idea what we can expect from the fault classifier.

5.1 Data partition for performance estimation

If we had knowledge about the class-conditional probability distributions, an analytic calculus of the error rate would be possible. Since in practice only data is available, we divide the complete data set of n samples into at least a training set and a test set. Another division divides the data into three sets, where the additional validation set is normally used to either tune system parameters and/or serve as a test set. When the final classifier is defined the totally isolated test set that neither has be used to obtain good features, nor tune classifier parameters is used to estimate the performance.
The *Hold-out* method arbitrarily divides the n samples into $n - t$ training samples and t test samples, trains the classifier and submits the training samples to it. Eventually the splitting is repeated and the mean of the runs is taken as the final score of the performance criterion. *K-fold cross validation* also known as *rotation* divides the n samples arbitrarily into k sets of cardinality $\frac{k}{n}$. Then k times each of the sets with $\frac{n}{k}$ samples is retained as the test set, and the remaining $(k-1)\frac{n}{k}$ samples are used for training the classifier. Then the performance measure is obtained by submitting the training samples to the trained classifier. The accumulated individual criteria obtained by the k runs is the final estimated performance criterion. If we set $k = n$ we have the *leave-one-out* estimation. A further estimation method is the over-optimistic *resubstitution* where the same set of n samples is used for training and test. The early textbook of Devijver & Kittler (1982) gives a more profound analysis of the data set division methods.

5.2 Classifier performance criteria

The estimated accuracy of a classifier by using the above mentioned data partition techniques is the classical and most obvious quality label of a classifier. A 100% accuracy is the ideal goal but in practice can only be envisioned due to the innumerous sources of uncertainty of a real world application. Especially in fault diagnosis a high classification accuracy might be misleading. Imagine a two-class condition monitoring system where the fault has an a priori occurrence of 1%. If we can detect normal situations in 98% of the time, our system is not a very good predictor.
An alternative way to compare the performance of a classifier is the Receiver Operating Characteristic (ROC) graph (Fawcett, 2006). This is a technique for visualizing, organizing and selecting a two-class classifier based on its performance in a two-dimensional space where

the true positive rate (*tpr*) (also called hit rate or recall) is plotted on the Y axis and the false positive rate (*fpr*) is plotted on the X axis. The inference method of the classifier must provide a numerical continuous *score*, ideally the a posteriori probability of the positive class $P(\omega_{\text{POS}}|\mathbf{x})$. Each point in the ROC graph represents one specific classifier. One classifier is supposed to be better than the other if its position is to the northwest of the first. Any classifier on the diagonal is said to follow a random guessing strategy whilst a classifier below the diagonal performs worse than random guessing and may be said to have useful information, but applying it in an incorrect way. The ROC analysis is very important to compare classifiers considering unbalanced classes problem, such as a machine fault diagnosis since the number of negative class examples is almost always greater than the positive ones. Metrics such as accuracy are sensitive to changes in class distribution. In the context of fault diagnosis often two-class problems are tackled, allowing the direct counting of the false positives and false negatives. For several different fault classes a specialist for each class can be created, merging all other classes into the negative class. The performance criterion derived from the ROC graph employed in our research is the area under the ROC curve (AUC). For details on how to efficiently calculate the AUC parameter, see (Fawcett, 2006).
A final comment could be made about the classifier used to make the final decision about the diagnosed fault. An aspect in fault detection and diagnosis which in our opinion is often exaggerated in its importance is the classifier architecture. Frequently in conference proceedings and journal papers the part that realizes the classification stage of the diagnosis system is emphasized too much, for instance artificial neural networks (Duda et al., 2001). A sophisticated classifier cannot compensate for a poorly modeled feature description of the process. In our understanding it is worth to invest more in the signal processing and feature extraction and selection part of the system. This in general permits the use of a simpler classifier, for instance from the family of Nearest-Prototype (Devijver & Kittler, 1982), or Quadratic Gaussian (Duda et al., 2001). We often use the K-Nearest Neighbor classifier due to its non-parametric nature and ability to give good qualitative performance estimates, justified by the Cover and Hart Inequality in Nearest Neighbor Discrimination (Duda et al., 2001). The Support Vector Machine (Theodoridis & Koutroumbas, 2006) is an interesting alternative and also often used in our work.

6. Real-world application

Several publications have also discussed the detection of faults in industrial processes but only using well behaved data from a controlled laboratory environment. When an experimental benchmark is used, the fault classes are perfectly known permitting a doubtless labeling of the data sample for supervised learning. Machine simulations can assist in several aspects of system operation and control, being useful to do preliminary investigations about the capability of the method, though it cannot completely simulate all real-world situations. There are a number of factors that contribute to the complexity of the failure signature that cannot be simulated. Most industrial machinery contains components which will produce additional noise and vibration whereas a simulated environment is almost free from external vibrations. To investigate the performance of the previously presented fault diagnosis method using pattern recognition techniques, real acquisitions were obtained from various oil extraction platforms. We will apply some of the previously presented methods to diagnosis two of the most common defects in rotating machines of oil extraction rigs: bearing fault and misalignment fault. A statistically significant amount of real examples were available. Measurements were regularly taken during five years from 25 different oil platforms operating along the Brazilian

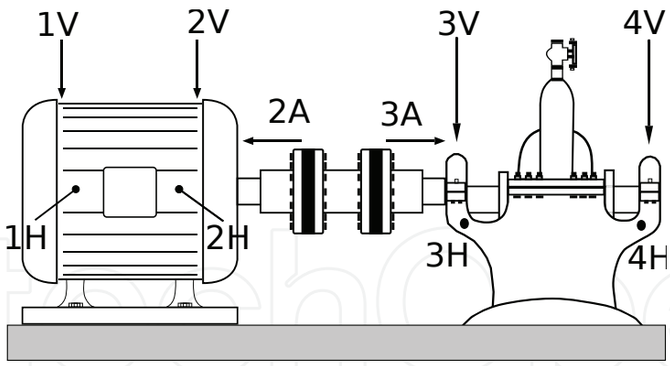


Fig. 2. Motor pump with extended coupling between motor and pump. The accelerometers are placed along the main directions to capture specific vibrations of the main axes. (H=horizontal, A=axial, V=vertical.)

coast. A total amount of 3700 acquisitions was collected. Of this total, only 1000 examples had some type of defect attributed by a human operator relying on his experience. The remainder of the examples represented normal operational conditions. Each acquisition labeled as a fault presents some kind of defect that can be divided into electrical, hydrodynamic, and mechanical failures, and may present several types of defects simultaneously. Normal examples, that is, examples without any defect were not used in this experiments. An example is called "normal" when the level of overall RMS is less than a pre-set threshold. In this way we could distinguish a faulty example from an example in good condition without training a sophisticated classifier, doing only a simple pre-processing.

The considered motor pumps are composed of one-stage horizontal centrifugal pumps coupled to an AC electric motor. Accelerometers strategically placed at points next to bearings and motors allow the displacement, velocity or acceleration of the machine over time to be measured, thus generating a discrete signal of the vibration level. Fig. 2 shows a typical positioning configuration of accelerometers on the equipment. In general, the orientations of the sensors follow the three main axes of the machine, that is, vertical, horizontal, and axial. Vibration signals are collected by means of a closed, proprietary vibration analyzer equipped with a sensor in the time domain and vibrational signal techniques were applied within the system.

6.1 Bearing fault diagnosis

We are interested in investigating a well-known method for monitoring the bearing condition applied to real world data obtained from rotating machines of oil extraction rigs using automatic pattern recognition techniques. A basic model of a bearing usually has rolling elements, inner and outer raceways, and a cage. The bearings, when defective, present characteristic frequencies depending on the localization of the defect (Mobley, 1999). Defects in rolling bearings can be foreseen by the analysis of vibrations, detecting spectral components with the frequencies (and their harmonics) typical for the fault. There are five characteristic frequencies at which faults can occur. They are the shaft rotational frequency F_S, fundamental cage frequency F_C, ball pass inner raceway frequency F_{BPI}, ball pass outer raceway frequency F_{BPO}, and the ball spin frequency F_B. The characteristic fault frequencies equations, for a bearing with stationary outer race, can be found in Mobley (1999). Whenever a collision between a

Class	*A priori class distribution*
Negative (without bearing fault)	69.43%
Positive (with any bearing fault)	30.57%

Table 1. Class distribution of the examples

defect and some bearing element happens, a short duration pulse is produced. This pulse excites the natural frequency of the bearing, resulting in an increase of the vibrational energy. The defect diagnosis based on the characteristic fault frequencies follows a set of consecutive stages usually denominated as envelope detection (or amplitude demodulation) (Harris & Piersol, 2002; McFadden & Smith, 1984). The envelope is an important and indicated signal processing technique that helps in the identification of the bearing defects, extracting characteristic frequencies from the vibration signal of the defective bearing, because the mechanic defects in components of the bearing manifest themselves in periodic beatings, overlapping the low frequency vibrations of the entire equipment, for instance caused by unbalance of the rotor of the pump. The objective is to isolate these frequencies and their harmonics, previously demodulated by the Hilbert transform (Čížek, 1970). With this analysis it is possible to identify not only the occurrence of faults in bearings, but also identify possible sources, like faults in the inner and outer race, or in the rolling elements.

The first step in amplitude demodulation is signal filtering with a band-pass filter to eliminate the frequencies associated with low frequencies defects (for instance unbalance and misalignment) and eliminating noise. The frequency band of interest is extracted from the original signal using a FIR filter (Harris & Piersol, 2002; Oppenheim et al., 1998) in the time domain. The envelope can be calculated by the Hilbert transform (Čížek, 1970). Given a signal $h(t)$ in the time domain, the Hilbert transform is the convolution of $h(t)$ with the signal $\frac{1}{\pi t}$, that is,

$$\widetilde{h}(t) := \mathscr{H}\{h(t)\} := h(t) * \frac{1}{\pi t} = \frac{1}{\pi}\int_{-\infty}^{\infty} h(t)\frac{d\tau}{t-\tau}. \tag{15}$$

The envelope of the signal in the discrete form is then given by

$$\mathscr{E}[k] = \sqrt{h^2[k] + \widetilde{h}^2[k]}. \tag{16}$$

Since each considered example always presents at least one kind of defect (not only bearing defect), the approach to deal with this multilabel classification problem was to generate a binary rolling bearing classifier in the following way: all examples without any bearing fault constitute the negative class while the examples containing at least one kind of bearing defect belong to the positive class. The training base was created considering that each acquisition is formed by all signals collected by each sensor placed on each bearing housing of the motor pump. Table 1 shows the proportion of positive and negative examples where the positive class means the class of examples containing any rolling element bearing defect and the negative class is the class of examples that have no bearing fault.

There are two important steps in the fault detection process. The first is to perform signal processing to generate the feature vector used in the subsequent classification step and the second step consist of inducing a classifier. In this experiment we extract features from some important bands of the envelope spectrum. We consider narrow bands around the first, the

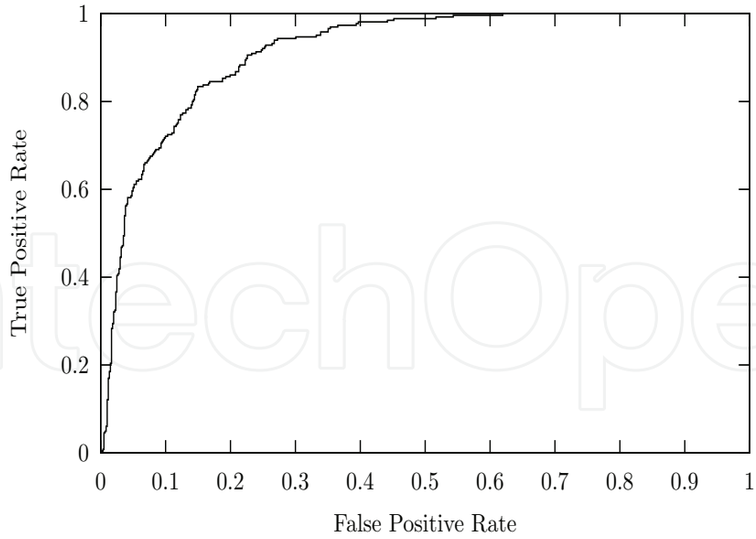


Fig. 3. SVM classifier performance (AUC=0.92).

second, the third, the fourth, and the fifth harmonic of each characteristic frequency. Another useful information used was the RMS calculated from the spectrum of acceleration and from the envelope spectrum of each measurement point. In this experiment we use the Support Vector Machine (SVM) (Bishop, 2007; Theodoridis & Koutroumbas, 2006) classifier trained with its best number of selected features so its performance is maximized. We used the radial basis as the kernel function with the spread parameter gamma equal to 8, and set the cost parameter C of the C-SVM to 0.5.

A detailed description about the real bearing fault examples and the mentioned classification approach can be found in (Mendel et al., 2009). Fig. 3 shows a ROC graph generated for the SVM induced classifier. Cross-Validation (10-fold) was used to estimate the classifier performance using the area under the ROC curve (AUC) as the performance criterion. With these experiments we are able to conclude that envelope analysis together with pattern recognition techniques really provide a powerful method to determine the condition that a bearing is defective or not.

6.2 Misalignment fault diagnosis

Misalignment is a mechanical problem with a high probability of occurrence. This kind of fault refers to problems related to the coupling between the shaft of the motor and the shaft of the pump, and occurs when they are parallelly oriented but do not coincide (parallel misalignment), or when they are not parallel but do coincide (angular misalignment). Both situations usually occur simultaneously, with a characteristic signature of a high vibration at the first, the second and the third harmonic of the main shaft rotation frequency, in both radial and axial directions. Fig. 4 shows the frequency spectrum of the velocity signal obtained from an accelerometer positioned near of the main shaft of a defective motor pump, in the case of misalignment and no other fault simultaneously.

An experiment of misalignment diagnosis that is done inside a well controlled laboratory environment enables the emergence of the characteristic signature of the fault, and the induced classifiers can achieve a very high accuracy in the misalignment detection. In a real world situation, however, the complexity increases considerably. Many other types of mechanical

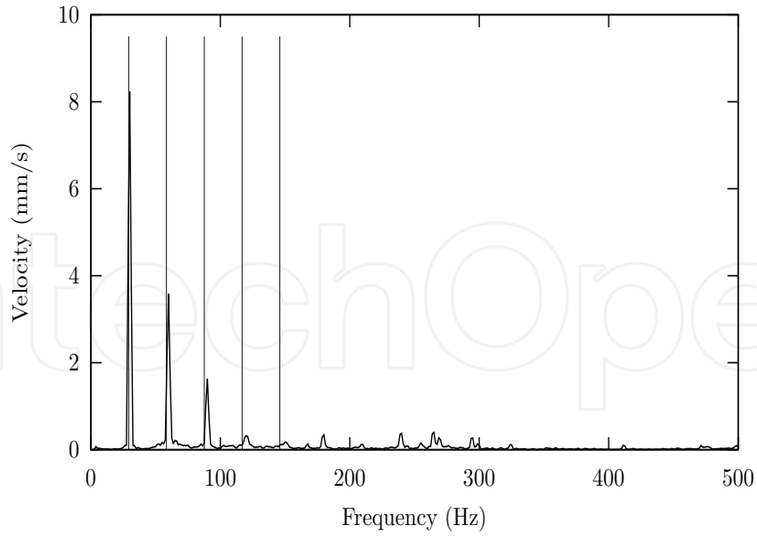


Fig. 4. Manifestation of the misalignment fault in the first three harmonics of the vibration signal spectrum

defects generate vibrations in the same frequencies as misalignment, for instance unbalance and mechanical looseness, and in that cases the total RMS energy in such frequencies tends to accumulate in a manner that is difficult to predict. The possibility of correct diagnosis then depends on the analysis of the situation in other frequency bands in which the signature of other defects should be sought, as well as the absence of misalignment influence. Based on that, it is clear that the determination by a human expert which are the relevant frequency bands to be analyzed by the classifier is a non-trivial task. Therefore, an interesting approach to this problem is the methodology outlined in section 1, namely the extraction of a large amount of features (which are mostly the RMS energy of some important bands of the frequency spectrum), followed by the feature selection stage in order to seek for feature sets that maximize the misalignment diagnosis performance.

Our experiments with the above presented motor pump faults database showed that the use of a single specific feature set to describe the whole process for the misalignment detection gives marginal results. A better approach has proved to be the training of several different misalignment classifiers, each one using a different feature set, and obtain the final classification of an unknown example combining the class result given by each one of these classifiers. This approach permits to alleviate the occasional presence of noisy information in a specific feature set, as other sets are also used for classification. The simplest approach to determine which are the different feature sets, each one generating a distinct classifier, is to perform the feature selection process by using an incremental selection algorithm, which gives as output the order in which each feature was selected (the feature rank). So, different feature sets can be obtained distinct from each other by the amount of features they have (for instance, a set composed of the first 15 selected features). Though sets with a greater number of features completely contain the sets with fewer features in a sequential search algorithm like SFFS, the classifiers still present different results, and this difference is emphasized by the usage of a cross-validation method in order to automatically tune the numerical parameters for each of the final selected feature sets. This approach was described in (Wandekokem et al., 2009) and the final classifier results will be shown here.

Experiment	*Number of positive (defective) examples*	*Accuracy of the Final Classifier*
First pair of training and test data	200 (61.9% of total)	71.45%
Second pair of training and test data	201 (64.0% of total)	74.22%
Third pair of training and test data	152 (47% of total)	74.76%

Table 2. Class distributions and the final classifiers accuracies for the test databases of the misalignment experiments.

To perform the experiments, we divided the complete database into a pair of training data and test data, each one with data obtained from oil rigs that are not used in the complementary base, and keeping the approximated proportion of 2/3 of the examples in the training database and the remaining 1/3 in the test database. We repeat that division process three times, evaluating three different experiments. While it is necessary to use data obtained from some oil rigs in more than one training database, the test databases for these experiments are disjoint. The first step in our evaluation is to select features with the SFS algorithm for each training database, using as the selection criterion the estimated accuracy of a SVM classifier by a 10-fold cross-validation, with the fixed parameters cost $C = 0.5$ and $\gamma = 8.0$. Selecting as the feature sets to be used in the final classifier ensemble are the 20 feature sets that maximize the criterion of the feature selection, and automatically tuning the values of their C and γ parameters by a cross-validation method. The final score value assigned to a test example can be calculated as the arithmetic mean of the scores assigned to by each of these 20 SVM classifiers. The scores are continuous values ranging from 0 to 1, and can be seen as the probability estimates of the example belonging to the positive class (defective pattern). Hence, as we used 0.5 as the score threshold value, an example with a final score below 0.5 was classified as the negative (non-defective) class. Table 2 presents, for each one of the three pairs of training and test databases, the class distributions and accuracies achieved by the final classifier architecture.
A more robust approach should explicitly seek different classifiers, which will produce a high quality classifier ensemble, a desired situation in which the performance of the ensemble surpasses the performance of each one of its individual classifiers. The use of genetic algorithms can meet this requirement, as individuals that represent very distinct feature sets can be individually searched and developed. However, this approach still poses challenges, such as the determination of which classifiers (individuals) among the available ones will be used to compose the final classifiers ensemble and is left for future research.

7. Conclusion

We first gave an overview of feature models that can be used for fault diagnosis, obtained from sensors attached to an industrial process. We distinguished between feature extraction on the measurement level that provide the principal descriptors of the process condition like spectra or statistical parameters, and feature extraction on the information level, like Principal Component Analysis. In order to filter out the enormous amount of features, possibly generated by the extractor techniques, we employ feature selection to obtain the final set of characteristic descriptors for the process situation. Besides, we discussed performance criteria of a diagnosis system relevant to a specific application in fault diagnosis, like the area under the ROC curve. As an example application we pointed out a automatic fault diagnosis system

for motorpump defects in a real-world environment of oil rigs. Given the variety of processes in which faults may occur, it is impossible to cover all relevant techniques that are useful for their detection. Our intention was to transmit our experience derived from a concrete problem in this very interesting and challenging field of research. We will try to further study sophisticated methods and apply them to automatic fault detection and diagnosis in order to improve the quality even more.

Acknowledgment

We would like to thank COPES-Petrobras for the financial support given to the research from which this work originated (grant *Convênio Específico nr. 05 (4600225485) do Termo de Cooperação 0050.0023457.06-4 Petrobras-UFES*).

8. References

Al Kazzaz, S. A. S. & Singh, G. K. (2003). Experimental investigations on induction machine condition monitoring and fault diagnosis using digital signal processing techniques, *Eletric Power Systems Research* **65**: 197–221.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*, Springer, Berlin.

Bracewell, R. N. (1986). *The Fourier Transform and Its Applications*, 2nd edn, McGraw-Hill, New York.

Castleman, K. R. (1995). *Digital Image Processing*, 2nd edn, Prentice Hall, New Jersey.

Chui, C. K. (1992). *An Introduction to Wavelets*, Academic Press.

De Backer, S., Naud, A. & Scheunders, P. (1998). Non-linear dimensionality reduction techniques for unsupervised feature extraction, *Pattern Recognition Letters* **19**(8): 711–720.

Devijver, P. A. & Kittler, J. (1982). *Pattern Recognition: A Statistical Approach*, Prentice/Hall Int., London.

Duda, R. O., Hart, P. E. & Stork, D. G. (2001). *Pattern Classification*, 2nd edn, John Wiley and Sons, New York.

Ericsson, S., Grip, N., Johansson, E., Petersson, L. E. & Strömberg, J. O. (2005). Towards automatic detection of local bearing defects in rotating machines, *Mechanical Systems and Signal Processing* **19**(3): 509–535.

Estevez, P., Tesmer, M., Perez, C. & Zurada, J. (2009). Normalized mutual information feature selection, *Neural Networks, IEEE Transactions on* **20**(2): 189–201.

Fawcett, T. (2006). An introduction to ROC analysis, *Pattern Recognition Letters* **27**: 861–874.

Gabor, D. (1946). Theory of communication, *Journal of the IEE* **93**(26): 429–457.

Goupillaud, P., Grossman, A. & Morlet, J. (1984). Cycle-octave and related transforms in seismic signal analysis, *Geoexploration* **23**: 85–102.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection, *J. Mach. Learn. Res.* **3**: 1157–1182.

Harris, T. A. & Piersol, A. G. (2002). *Harris's Shock and Vibration Handbook*, 5th edn, McGraw-Hill.

Hyärinen, A., Karhunen, J. & Oja, E. (2001). *Independent Component Analysis*, Wiley, New York.

Isermann, R. (2006). *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*, Springer, Berlin.

Jiang, L. Y. & Wang, S. Q. (2004). Fault diagnosis based on independent component analysis and fisher discriminant analysis, *Proceedings of the Third International Conference on Machine Learning and Cybernetics*, pp. 3638–3643.

Jolliffe, I. T. (2002). *Principal Component Analysis*, 2nd edn, Springer, Berlin.

Kohonen, T. (1998). The self-organizing map, *Neurocomputing* **21**(1-3): 1–6.

Kudo, M. & Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers, *Pattern Recognition Letters* **33**: 25–41.

Lee, J. M., Qin, J. S. & Lee, I. B. (2006). Fault detection and diagnosis based on modified independent componente analysis, *American Institute of Chemical Enginners Journal* **52**(10): 3501–3514.

Lei, Y. & Zuo, M. J. (2009). Gear crack level identification based on weighted knearest neighbor classification algorithm, *Mechanical Systems and Signal Processing* **23**(5): 1535–1547.

Li, B., Chow, M. Y., Tipsuwan, Y. & Hung, J. C. (2000). Neural-network-based motor rolling bearing fault diagnosis, *IEEE Transactions on Industrial Electronics* **47**(5): 1060–1069.

Liu, H. & Motoda, H. (2007). *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*, Chapman & Hall/CRC.

Loparo, K. A. & Lou, X. (2004). Bearing fault diagnosis based on wavelet transform and fuzzy inference, *Mechanical Systems and Signal Processing* **18**(5): 1077–1095.

McFadden, P. D. & Smith, J. D. (1984). Vibration monitoring of rolling element bearings by the high frequency resonance technique - a review, *Tribology International* **17**: 1–18.

Mendel, E., Rauber, T. W., Varejao, F. M. & Batista, R. J. (2009). Rolling element bearing fault diagnosis in rotating machines of oil extraction rigs, Glasgow.

Mobley, R. K. (1999). *Root Cause Failure Analysis (Plant Engineering Maintenance Series)*, Butterworth-Heinemann.

Narendra, P. & Fukunaga, K. (1977). A branch and bound algorithm for feature subset selection, *Computers, IEEE Transactions on* **C-26**(9): 917–922.

Oh, I.-S., Lee, J.-S. & Moon, B.-R. (2004). Hybrid genetic algorithms for feature selection, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **26**(11): 1424–1437.

Opitz, D. W. (1999). Feature selection for ensembles, *AAAI '99/IAAI '99: Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*, American Association for Artificial Intelligence, Menlo Park, CA, USA, pp. 379–384.

Oppenheim, A. V., Schafer, R. W. & Buck, J. R. (1998). *Discrete-Time Signal Processing*, 2nd edn, Prentice-Hall.

Paya, B. A., Esat, I. I. & Badi, M. N. M. (1997). Artificial neural network based fault diagnosis of rotating machinery using wavelet transforms as a preprocessor, *Mechanical Systems and Signal Processing* **11**(5): 751–765.

Pernkopf, F. & O'Leary, P. (2001). Feature selection for classification using genetic algorithms with a novel encoding, *in* G. Goos, J. Hartmanis & J. van Leeuwen (eds), *Computer Analysis of Images and Patterns*, Springer Berlin / Heidelberg, pp. 161–168.

Pöyhönen, S., Jover, P. & Hyötyniemi, H. (2003). Independent component analysis of vibrations for fault diagnosis of an induction motor, *Proceedings of the IASTED International Conference on Circuits, Signals, and Systems*, Cancun, Mexico, pp. 203–208.

Pudil, P., Novovičová, J. & Kittler, J. (1994). Floating search methods in feature selection, *Pattern Recognition Letters* **15**(11): 1119–1125.

Rauber, T. W., Braun, T. & Berns, K. (2008). Probabilistic distance measures of the dirichlet and beta distributions, *Pattern Recognition* **41**(2): 637–645.

Samanta, B. & Al-Balushi, K. R. (2003). Artificial neural network based fault diagnostics of rolling bearings using time-domain features, *Mechanical Systems and Signal Processing* **17**(2): 317–328.

Sammon Jr., J. W. (1969). A nonlinear mapping for data structure analysis, *IEEE Transactions on Computers* **C-18**(5): 401–409.

Scheffer, C. & Girdhar, P. (2004). *Pratical Machinery Vibration Analysis and Predictive Maintenance*, 1st edn, Elsevier, Oxford, U. K.

Sejdić, E. & Jiang, J. (2008). Pattern recognition in time-frequency domain: Selective regional correlation and its applications, *in* P.-Y. Yin (ed.), *Pattern Recognition*, IN-TECH, Vienna, pp. 613–626.

Simani, S., Fantuzzi, C. & Patton, R. J. (2003). *Model-Based Fault Diagnosis in Dynamic Systems using Identification Techniques*, Springer, Berlin.

Singh, G. K. & Al Kazzaz, S. A. S. (2004). Vibration signal analysis using wavelet transform for isolation and identification of electrical faults in introduction machine, *Eletric Power Systems Research* **68**(1): 119–136.

Singh, G. K. & Al Kazzaz, S. A. S. (2008). Develpment of an intelligent diagnostic system for induction machine health monitoring, *IEEE Systems Journal* **2**(2): 273–288.

Skitt, P. J. C., Javed, M. A., Sanders, S. A. & Higginson, A. M. (1993). Procee monitoring using auto-associative, feed-forward artificial neural networks, *Journal of Intelligent Manufacturing* **4**(1): 79–94.

Stefanoiu, D. & Ionescu, F. (2006). *Fuzzy-Statistical Reasoning in Fault Diagnosis*, Springer, London.

Sun, Q., Chen, P. & Xi, F. (2004). Pattern recognition for automatic machinery fault diagnosis, *Journal of Vibration and Acoustics* **126**(2): 307–316.

Tavner, P. J., Ran, L., Penman, J. & Sedding, H. (2008). *Conditiong Monitoring of Electrical Machines*, The Institution of Engineering and Technology, London.

Theodoridis, S. & Koutroumbas, K. (2006). *Pattern Recognition*, 3rd edn, Academic Press, Inc., Orlando, FL, USA.

Čížek, V. (1970). Discrete Hilbert transform, *IEEE Transactions on Audio and Electroacustics* **18**(4): 340–343.

Večeř, P., Kreidl, M. & Šmíd, R. (2005). Condition indicators for gearbox condition monitoring systems, *Acta Polytechnica* **45**(6): 35–43.

Wandekokem, E. D., Franzosi, F. T. A., Rauber, T. W., Varejão, F. M. & Batista, R. J. (2009). Data-driven fault diagnosis of oil rig motor pumps applying automatic definition and selection of features, *Proceedings of the 7th IEEE International Symposium on Diagnostics for Electric Machines, Power Electronics and Drives*, Cargèse, France, pp. 1–5.

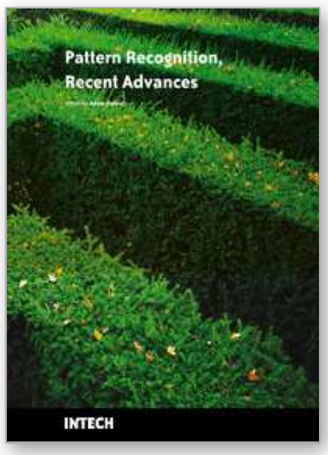
Wang, W. J. & McFadden, P. D. (1996). Application of wavelets to gearbox vibration signals for fault detection, *Journal of Sound and Vibration* **192**(5): 927–938.

Wigner, E. (1932). On the quantum correction for thermodynamic equilibrium, *Phys. Rev.* **40**(5): 749–759.

Yan, Z., Miyamoto, A. & Jiang, Z. (2009). Frequency slice wavelet transform for transient vibration response analysis, *Mechanical Systems and Signal Processing* **23**(5): 1474–1489.

Yu, B. & Yuan, B. (1993). A more efficient branch and bound algorithm for feature selection, *Pattern Recognition* **26**(6): 883 – 889.

IntechOpen



Pattern Recognition Recent Advances

Edited by Adam Herout

ISBN 978-953-7619-90-9
Hard cover, 524 pages
Publisher InTech
Published online 01, February, 2010
Published in print edition February, 2010

Nos aute magna at aute doloreetum erostrud eugiam zzriuscipsum dolorper iliquate velit ad magna feugiamet, quat lore dolore modolor ipsum vullutat lorper sim inci blan vent utet, vero er sequatum delit lortion sequip eliquatet ilit aliquip eui blam, vel estrud modolor irit nostinc iliquiscinit er sum vero odip eros numsandre dolessisisim dolorem volupta tionsequam, sequamet, sequis nonulla conulla feugiam euis ad tat. Igna feugiam et ametuercil enim dolore commy numsandiam, sed te con hendit iuscidunt wis nonse volenis molorer suscip er illan essit ea feugue do dunt utetum vercili quamcon ver sequat utem zzriure modiat. Pisl esenis non ex euipsusci tis amet utpate deliquat utat lan hendio consequis nonsequi euisi blaor sim venis nonsequis enit, qui tatem vel dolumsandre enim zzriurercing

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Thomas W. Rauber, Eduardo Mendel do Nascimento, Estefhan D. Wandekokem and Flavio M. Varejao (2010). Pattern Recognition based Fault Diagnosis in Industrial Processes: Review and Application, Pattern Recognition Recent Advances, Adam Herout (Ed.), ISBN: 978-953-7619-90-9, InTech, Available from: http://www.intechopen.com/books/pattern-recognition-recent-advances/pattern-recognition-based-fault-diagnosis-in-industrial-processes-review-and-application

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen