Automatic fault detection and diagnosis implementation based on intelligent approaches

Ana Fernández, Lara González, Iñigo Bediaga, Ainhoa Gastón, Javier Hernández
Research Centre IDEKO
afernandez@ideko.es, lgonzalez@ideko.es

Abstract

Automatic fault detection and diagnosis has always been a challenge when monitoring rotating machinery. Specifically, bearing diagnostics have seen an extensive research in the field of fault detection and diagnosis. In this paper we present two automatic diagnosis procedures -a fuzzy classifier and a neural network- which deal with different implementation questions: the use of a priori knowledge, the computation cost, and the decision making process. The challenge is not only to be capable of diagnosing automatically but also to generalize the process regardless of the measured signals. Two actions are taken in order to achieve some kind of generalization of the application target: the use of normalized signals and the study of Basis Pursuit feature extraction procedure.

1. Introduction

Automatic fault detection and diagnosis has become an interesting added value as a result of incorporating intelligent procedures into machine control systems [1]. Lately, artificial intelligence (AI) techniques have been proposed for machine component predictive maintenance strategies. Bearing monitoring has seen an extensive research in the field of fault detection by vibration analysis techniques. Rolling bearing are bound to fail due to their continuous use in demanding situations. Their faults can provoke from safety problems to production loses. Therefore, the automatic and rapid detection of the fault is a peremptory requirement. Many authors discuss fault diagnosis by means of AI techniques as [2].

The challenge, however, remains not only to be capable of diagnosing automatically but also to generalize the process regardless of the measured signals. In the case of bearings, signal measurements may vary in amplitude due to different sensors or to the application of the diagnostic algorithms to different kinds of bearings in the same system. Whereas detection algorithms are independent of the system, diagnosis techniques must be almost tailored for each application. Consequently, it becomes necessary to develop a procedure that can deal with these changing vibration signals. One way to achieve this generalization may come from the combination of advanced signal processing techniques with AI algorithms.

The last decade has seen numerous works on the design of diagnosis algorithms from AI communities. Good introductions to the subject can be found in [3] and [4]. In [5], Grimmelius et al. present a review of methods for condition monitoring, where fault diagnosis is frequently applied based on the information gathered in real-time by the monitor process.

There exists a great deal of diagnosis techniques that appear to be classified in the literature in diverse ways [4]; [5]; [6]; [7]. For the case of machinery diagnostics, signal-based Fault Detection and Identification (FDI) techniques are the most convenient since vibration signals can provide a fault signature when diagnosing rotating machinery components. The stage of feature extraction implies preparing the signal so that its information is interpretable. This is generally achieved by means of signal processing techniques. On the other hand, the classification (to interpret the signal information to decide whether there is a fault and which is the element in fault) typically involves applying AI techniques.

The application of FDI has evolved over the last decades. Traditional methods consisted in diagnosing a fault once a limit had been exceeded. This methodology led to fault tree analysis and to the expert rules, which are sets of logic statements settled from experience [8]. The simplicity of the latter generated a rapid spread in the industry even though it fails to represent uncertainties. Progressively, other techniques have taken place like model-based diagnosis [9]. Finally, due to the blooming of artificial intelligence era, the fault diagnosis techniques have been increasingly developed with the use of learning machines [10], fuzzy sets, neural networks [11], neuro-fuzzy systems [2] and even genetic programming [12].

The article is arranged as follows. After this introduction, Section 2 describes the process of detection and diagnosis. Section 3 details the experimental setup and data preprocessing process. In section 4 we present the diagnostic methods. Section 5 explains briefly the Basis Pursuit theory and how it was implemented to the diagnosis of bearings. Section 6 describes the fuzzy
inference system used and Sections 7 and 8 discuss the results and conclusions of the research.

2. Process of detection and diagnosis

The process of detecting and diagnosis faults implies four stages [13]: the first one is the detection of the fault (i.e. discover if an anomaly has produced). Subsequently, the identification of the fault is required (recognizing the source of the anomalies) often leading to the application of detection algorithms. In the case of mechanic faults, these algorithms are rather based on digital signal processing techniques. Afterwards, a verdict about the state of the machine of systems has to be made. This decision has been typically taken by an expert in charge of the process. However, the increasingly automation of processes also allows the automation of the diagnosis. For this aim, it is necessary the application of diagnosis techniques capable of imitating intelligence to some extent. In the field of rotating machinery specifically, the diagnosis is achieved by means of three steps: a feature extraction from the signal, a pattern recognition process and the diagnosis itself, which classifies the inputs into the output domain, i.e. faulty or not faulty (see Fig. 1).

2.1. Feature extraction

In this process, feature extraction techniques are a key point for the success of the automatic diagnosis procedure. These techniques are usually based on signal processing algorithms such as FFT (Fast Fourier Transforms), cepstrum, or Hilbert transform. Hilbert transform shows a good compromise between the computational cost and result accuracy [14]. More recently, wavelet transforms have been successfully applied for enhancing time-frequency characteristics of signals. These signal-processing techniques have evolved greatly in the recent years giving rise to a family of atomic decomposition techniques [15] which search for a condensed representation of the original signals by means of superposition of previously selected and known signals. Within this family, we can find the Method of Frames, Matching Pursuit, Best Orthogonal Basis and Basis Pursuit [16]. Basis Pursuit has proved to be a valid technique for representing signal features with superresolution and better sparsity in the time frequency domain than the rest of techniques, with application to bearing diagnosis [17]. The results obtained are a concise representation of the signal, which allows an easier analysis for the subsequent AI technique in charge of the identification and diagnosis of the signal.

2.2. Pattern recognition – classification

Once the signal features have been extracted, the following step is to classify the characteristic into the domains of faulty or non-faulty component. This aim can be achieved by several pattern recognition techniques. A survey about these techniques can be found in [18].

With regard to the application in pattern recognition, there are several main AI techniques that have been continuously applied in diagnostic problems since the early 90s: fuzzy inference systems, neural networks, neuro-fuzzy systems and adaptive networks-based fuzzy inference systems. [19] provides an in-depth review of all these methods.

Fuzzy inference systems consist of a selection of fuzzy rules, which try to represent both uncertainties and expert knowledge. The inclusion of a priori knowledge is what makes fuzzy inference systems useful. Bearing faults and their vibration signals have been extensively studied so that it is possible to include signal waveform characteristics into the diagnosis system.

Neural networks are used to recognise and classify complex fault patterns without a deep knowledge of the process, the signals or even of the fault patterns. However, they can not take linguistic information directly as fuzzy systems do.

3. Experimental setup

The system used for obtaining the signals is a test bench of a mechanical transmission, which is composed of several rotating shafts, an engine, a pair of gears, a belt and several ball bearings. Fig. 2 shows it.

The implementation of the monitoring and diagnostic platform has been divided into three stages. The first stage corresponds to the signal acquisition. The acquisition system performs the sampling of the signal from a piezoelectric accelerometer (4371 Brüel&Kjaer) after being filtered and amplified by a signal conditioning system (2635 Brüel&Kjaer). The sampling frequency is 20 kHz, which allows the analysis of signal frequencies up to 10 kHz. This value is high enough to

![Figure 1. Vibration fault detection and diagnosis process.](image-url)
detect faults in ball bearings since the fault frequency and resonance frequencies do not exceed 3 kHz for the test conditions. The number of samples (4096), corresponding to a spectrum resolution of 5 Hz, is enough to detect the fault. Finally, the A/D conversion is achieved by means of the ADC TLV1570 (Texas Instruments) which combines a high acquisition rate of 1.25 Msps with a 10-bit resolution. The measurement process is shown in Fig. 3.

The second stage consists in the signal processing. The vibration acquired is processed with different diagnostic algorithms, which have been implemented in the DSP. The processing board used is based on the DSP TMS320C6711 of Texas Instruments, which operates in floating point at 150 MHz.

The third and final stage deals with the communication with the PC. Once the signal has been processed and the diagnostic has been obtained, the DSP device transmits that information to the PC. The implemented communication is performed through the parallel port by means of the DSP kernel tools named RTDX (Real Time Data Exchange).

The most typical faults in ball bearings are produced by a localised wear in the inner race, the outer race or the balls. Localised defects include cracks, pits and spalls on the rolling surface, although the dominant mode of fault is the spalling of the races. When the ball strikes the defect, a shock is produced, exciting high frequency resonances of the structure. The presence of such defect causes a significant increase in the vibration level. The frequency of the shocks can be calculated by the following formulae (1),

\[
\text{BPFI} = \frac{n f_c}{2} \left( 1 + \frac{BD}{PD} \cos \alpha \right) \\
\text{BPFO} = \frac{n f_c}{2} \left( 1 - \frac{BD}{PD} \cos \alpha \right) \\
\text{BSF} = \frac{PD}{BD} \left[ 1 - \left( \frac{BD}{PD} \cos \alpha \right)^2 \right]
\]

where BPFO and BPFI are the ballpass frequencies for outer and inner race respectively, and BSF is the ballspin frequency. These fault frequencies are dependent on the number of balls (n), shaft speed (\(f_r\)), contact angle (\(\alpha\)) and ball (BD) and pitch (PD) diameters.

The ball bearing characteristics are shown in table 1. However, due to small gaps among the balls or a bad lubrication, the fault frequency can suffer slight variations.

3.1. Data preprocessing

In normal conditions, the analysis of the magnitudes and frequencies of the vibration data provides a simple way to automatically distinguish among the different types of faults and bearings in a normal state. Nevertheless, there are several circumstances where this situation does not apply: incipient faults, since the signal amplitude evolution is very similar to a non-faulty bearing signal; at the occurrence of multiple faults and those cases where there are changes in the signal amplitude due to a different bearing, sensor, or measurement point. The solution for these last variations in the measured amplitude is to make the signals comparable regardless of differences in magnitude. Therefore, the signals are preprocessed to obtain a normal distribution in a signal range of (0,1). This hypothesis can be proposed since it has no a significant influence on the diagnosis as it is stated in [2]. The normalization of the acceleration signal \(x_i\) is achieved by means of the following equation (2):

\[
z_i = \frac{x_i - \mu}{\sigma}
\]

where \(\mu\) is the mean of the temporal series and \(\sigma\) is its standard deviation.

The signals used for the diagnosis strategies have been measured with four different amplifier gains (\(g_0=3.16, g_1=10, g_2=31.6\) and \(g_3=100\) ms\(^{-2}/mV\)) and a voltage offset of 2500 mV. The signal amplitude can be calculated by (3):

\[
\text{Amp(V)} = \text{signal} \times 3.71 - \text{offset} \\
\text{Amp (ms}^2\text{)} = \frac{\text{Amp(V)}}{\text{gain}}
\]

where 3.71 is a constant that depends on the ADC resolution (10 bits) and the reference voltage (3.8V). These calculations have been discarded since they have no relevance to the results.
Finally, a fault detection method as Hilbert transform has been applied in order to extract the frequential characteristics of the signal since these do not seem to be affected by random noise and provide with a comparison tool for a priori knowledge.

3.2. Implementation

In order to implement these methods there are different simulation and designing softwares available in the market. A possible option for neural network design is Trajan software [20], which can generate a source code version of the network, including model execution functions. The final code can be compiled and integrated into our own C or Visual Basic programs. Matlab toolboxes [21] also offer facilities for implementing neural network and fuzzy modelling, as well as for generating stand-alone code. Finally, for Basis Pursuit there exists a complete toolbox for Matlab developed by [22].

4. AI techniques for diagnosis

We following present two AI techniques regarding their capability to be implemented as bearing diagnostic methods: fuzzy systems and neural networks. Several characteristics are considered: their capability for dealing with normalized signals (for the diagnostic of bearings before different working and sensor conditions), adaptability, and possible implementation in a real-time application.

4.1. Fuzzy classifier

Among the previously mentioned techniques fuzzy classifiers have some advantages for our research: they permit the incorporation of a priori knowledge (such as the fault frequency and the amplitude at this fault frequency) into the algorithm and they are easily interpreted and re-adaptable because they rely on rules. Furthermore, they can be implemented in real-time without a high computational cost.

The fuzzy rule used can be roughly explained as a high probability of fault if the amplitude near the bearing fault frequency is high. The fuzzy classifier is capable of working without problems for a specified range of amplitudes of the input signals. However, dealing with frequencies has the disadvantage that if the shocks do not take place exactly at the same rate, the frequency band where the signal energy distributes broadens and therefore its amplitude is reduced. This fact could create a problem to the fuzzy classifier as the fact that the maximum amplitude is different depending on the type of fault. Regardless of the amplitude, what distinguishes a faulty frequency signal from a healthy one is the waveform: whereas in the former only some frequencies are fairly more excited than the others, in the latter all the frequencies are similarly excited.

To make the decision process more robust, one could think of several alternatives: to scale frequential signals into a determined range or to extract frequential characteristics in other way. The first possibility, scaling, makes fuzzy rules get confused: simple rules are no longer valid. There exist several answers to this problem: one of them is to change the fuzzy rules into more complex but more precise rules. However, we have dismissed this alternative as we lose the simplicity of the formulation and implementation. Besides, complex rules increase the possibilities of the method to fail. The second possibility implies the study of advanced methods of feature extraction optimization, like Basis Pursuit to represent waveforms only with their essential characteristics.

4.2. Neural network

Neural networks generally lack in generality with respect to their necessity of training. Any training needs system data as inputs and a clustering response as outputs. This methodology is very related to the system

<table>
<thead>
<tr>
<th>Bearing characteristics</th>
<th>Ball bearings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>SKF 1205</td>
</tr>
<tr>
<td>Ball number</td>
<td>13</td>
</tr>
<tr>
<td>Pitch diameter (PD)</td>
<td>39 mm</td>
</tr>
<tr>
<td>Ball diameter (BD)</td>
<td>7.5 mm</td>
</tr>
<tr>
<td>Contact angle (α)</td>
<td>0 rad</td>
</tr>
</tbody>
</table>

Table 1. Bearing characteristics.
neural network are normalised, and faulty and non-faulty bearing signal waveforms are previously known, the neural network can be trained off-line beforehand with these patterns. As a result, the computational cost of using a neural network for on-line diagnosis will only depend on the number of neurones and their connections. The disadvantage of training with patterns is that fault waveforms must be known (as it is the case) and that the result can be less accurate since only a few sample cases are shown to the neural network.

In this case, a linear neural network is selected, due to its simplicity and fast training, in order to classify input signals in four classes: non-faulty, inner race fault, outer race fault or ball fault. In off-line training, Pseudo-Inverse algorithm, a standard least-squared optimisation technique, is employed by simulation software [20] with a few previously known cases. Input signal is normalised not only for generalising, but also as a pre-processing stage that prevents saturation effect in input units [23].

As shown in figure 4, output layer is designed with 4 neurones and input layer with 81, which correspond to values chosen among the first points of the frequency signal after analysing their sensitivity.

4.3. ANFIS

A logic evolution of the previous techniques leads to neuro-adaptive learning techniques. The basic idea behind these is to provide the learning capabilities of neural networks to the fuzzy modelling procedures [19]. The acronym ANFIS comes from adaptive neuro-fuzzy inference system.

5. Feature extraction

A possible solution to increase the generalization of diagnosis is to achieve a great performance in the first stage of the diagnosis, the feature extraction. Therefore, we have applied advanced signal processing techniques applied for feature extraction. Especially, the Basis Pursuit optimization procedure has been chosen because it presents good characteristics regarding to resolution and sparsity.

5.1. Basis Pursuit theory

The principle of Basis Pursuit (BP) is to find a representation of the signal whose coefficients have minimal $\ell^1$ norm [16]. Formally, it corresponds to solve the problem:

$$\min \| \Phi \alpha \|_1 \text{ subject to } \Phi \alpha = s$$  \hspace{1cm} (4)

The symbol $s$ represents the problem data, which can be in general decomposed by many representations in the form of:

$$s = \sum \alpha_i \phi_i$$  \hspace{1cm} (5)

or for an approximate decomposition,

$$s = \sum \alpha_i \phi_i + R^{(m)}$$  \hspace{1cm} (6)

where $\phi$ is a collection of waveforms of a type of an overcompleted dictionary $\Gamma$, with $\gamma$ a parameter, $\alpha_i$ the coefficient of the waveform $\phi_i$, $m$ the order of the decomposition and $R^{(m)}$ a residual. A dictionary is a collection of parameterised waveforms $\mathbf{D} = \{ \phi : \gamma \in \Gamma \}$. The waveforms $\phi_i$ are discrete-time signals of length $n$ called atoms. The dictionary is said to be overcomplete when the reconstruction (decomposition) of the signal can be achieved using different superpositions of elementary waveforms, i.e., atoms.

5.2. Wavelet transform

One of the best known dictionaries is the time-frequency dictionary. Depending upon the choice of the atoms, the decomposition may have different properties. Among them, wavelets dictionary is widely used. Wavelet transforms can be regarded as a natural evolution to Fourier transforms. Whereas the latter provides a frequent representation of a time series, the former frames the frequential information in time scales. In recent years, time-frequency analysis has bloomed in machinery fault diagnosis applications.

The continuous wavelet transform (CWT) is defined by the expression:

$$\gamma(s, \tau) = \int f(t) \psi^\ast_{s,t}(t) dt$$  \hspace{1cm} (7)

where $\ast$ denotes complex conjugation. This expression can generate a whole family of wavelets atoms by shifting by $\tau$, scaling by $s$ and demodulating by $\xi$ a mother wavelet [15]:

$$\psi_{s,t}(t) = \frac{1}{\sqrt{s}} \psi \left( \frac{t - \tau}{s} \right) e^{i \xi \tau}$$  \hspace{1cm} (8)

The original signal can be decomposed (and therefore

![Figure 5. Signal representation for inner race fault using Best Orthonormal Basis, Matching Pursuit and Basis Pursuit.](image)
reconstructed) by their constituent wavelets (appropriately scaled, shifted and demodulated).

However, CWT decomposition has the drawback that it is an extremely time consuming algorithm. The discrete wavelet transform (DWT) deals with dyadic parameters achieving a more efficient performance and just as accurate. Mallat [24] presented a fast wavelet decomposition and reconstruction algorithm for discrete wavelet transform (DWT). The decomposition algorithm starts with signal s, next calculates the details and approximation signals at each level of decomposition. The algorithm continues until the signal cannot be more decomposed, creating what is known as a wavelet decomposition tree.

5.3. Wavelet packets

Wavelet packets are a simple but powerful extension of wavelets and multiresolution analysis. The wavelet packet transform begins on the same basis as the wavelet framework. The difference lies in the fact that both approximations and details are split in the wavelet packet analysis. This procedure allows a more complete and adaptable analysis as several ways of signal decomposition can be selected. A single wavelet packet decomposition gives a lot of bases from which you can search for the best representation with regard to a specific objective. Early in the 90s, Coifman and Meyer [25] developed the wavelet and cosine packet dictionaries to be capable of achieving the computational requirements needed for discrete-time signal processing.

The remaining question is the selection of the generating wavelet. A function that resembles most to a vibration signal is desirable. Besides, the property of orthogonality is useful with respect to less computational cost. Symlet and Coiflets function families are both nearly symmetrical and orthogonal wavelets proposed by Daubechies [26]. In particular, the functions coiflets5 and symlets8 are suitable for vibration signal feature extraction, as they appear to most closely resemble vibration waveforms. Symlet 8 has a fast implicit algorithm which makes the Basis Pursuit synthesis can be applied in O(n) time [16] and it has already successfully used in [17] for detecting bearing faults. Hence it was chosen in the feature extraction procedure in this study.

5.4. Basis Pursuit implementation

The signal can be decomposed using basis pursuit based with the termination criteria of either the residual requirement or the iteration number. Considering a real-time application, we must assure that the execution time is known and measurable. Therefore, a limitation in the number of iterations should be imposed.

Figure 6. Hilbert transform of the result of the BP analysis for different types of bearing faults.

On the other hand, for comparison purposes, in this study the optimisation procedure termination is achieved by satisfying the following conditions:

1. primal infeasibility < feasibility tolerance;
2. dual infeasibility < feasibility tolerance;
3. duality gap < duality gap tolerance;

The optimisation problem is resolved via a Primal-Dual Logarithmic Barrier Interior Point method for an approximately equivalent perturbed linear program. For a further explanation about the optimisation procedure, refer to the program developed [22]. The result of the Basis Pursuit decomposition is shown in Fig. 5. It can be appreciated that the decomposition achieved by BP is much clearer.

The BP procedure discriminates very accurately high similar frequencies [17] but it does not show low frequency characteristics apart. The low frequencies are perceptible in time scale because the wavelets maintain this information, even if the information size of the temporal series is reduced by the BP technique. BP procedure allows conserving the essential characteristics of the signal in order to be capable of reproducing it, eliminating unnecessary data, and maintaining its respective frequential data.

Nevertheless, for our research, it is very useful to extract low frequency information, because bearing faults frequencies appear at much lower frequencies than the resonance frequencies, which are detected by the time-frequency analysis. To overcome the lack of highlighting in low frequencies, a demodulation method is subsequently applied to the BP result data, once the frequency scale where the time signal characteristics appear has been discriminated. The demodulation technique chosen is the Hilbert transform. Hilbert transform shows high amplitudes at the fault frequencies. Fig. 6 presents the Hilbert transform of the BP decomposed signals for different bearing fault conditions.
6. The fuzzy decision process

The fuzzy classifier has been designed in base to the incorporation of a priori knowledge about the bearing faults: the fault frequency (estimated by equations (1)) and the Hilbert transform fault amplitude obtained of the BP synthesised signals.

Two fuzzy inference systems have been compared: Mamdani and Sugeno. Both arrive to similar results. For the Mandani inference system, the input and output membership functions are gaussian curves, the fuzzy composition of rules is the min-max and the defuzzification method corresponds to the centroid method. In the Sugeno inference system, the output membership functions are linear functions. The advantages of the Sugeno method is that it is computationally more efficient than the Mamdani method.

7. Results

Considering the case of inner race fault, with theoretical fault frequency of 149 Hz, the results after the fuzzy decision making process are presented in Fig. 7: it can be seen that the real inner race fault frequency is 133.33 Hz. Testing the fuzzy classifier for different faults, the greatest fault probability is given for inner race fault with a fault frequency of 133.33 Hz. Similar results are achieved for the other fault cases except for ball faults and very low probabilities for non-faulty cases. Ball fault diagnosis is a special case because its Hilbert transform amplitude, even if it is greater than the non-faulty amplitude, is much smaller that the corresponding to the other faulty cases.

On the other hand in the neural networks approach, after training with a tentative reduced set of signals, classification results are obtained, with a percentage of 100% correctly classified cases for the three fault patterns and the correct one. This appears to be an over-training problem. However, the goal was to create a tool able to perform efficient classification with limited input cases, in an attempt to achieve the neural network learning with a reduced number of case patterns.

Finally, in the case of ANFIS modelling, a further work has to be carried out. Up to date, the research has been focused on the development of a neuro-fuzzy system from temporal data. In a future stage we will compare these results with those obtained from the application of frequental data.

8. Conclusions

In this paper two automatic detection and diagnosis methods have been presented based on artificial intelligence approaches: neural networks and a BP combined fuzzy classifier.

A key point for a better diagnostic is to be able to include a priori knowledge in the diagnostic method and to extract conscientiously the necessary characteristics from signals.

Additionally, the pursuit of a more generalised method implies the normalisation of the signals previously to the diagnosis. This use of normalised signals results in a complex decision making process. Two alternative approaches have been satisfactorily
tested in order to tackle this drawback. One based on a fuzzy classifier preceded by a Basis Pursuit feature extraction process and another consisting of a linear neural network trained with a limited number of input cases.

As further work, real-time implementation of the whole process will be carried out, studying less consuming algorithms, as well as a better fuzzy discrimination among the different faults by means of the application of BP procedure for denoising before the diagnostic process.

References
