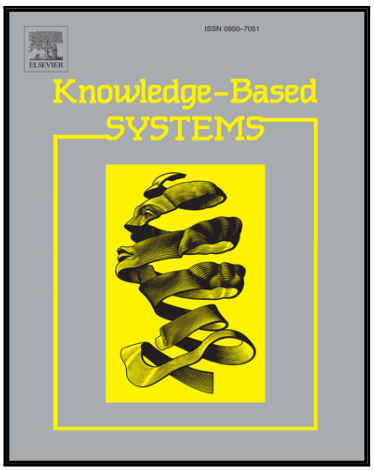


A Bayesian approach to consequent parameter estimation in probabilistic fuzzy systems and its application to bearing fault classification

Chuan Li, Luiz Ledo, Myriam Delgado, Mariela Cerrada, Fannia Pacheco, Diego Cabrera, René-Vinicio Sánchez, José Valente de Oliveira

PII: S0950-7051(17)30218-6
DOI: 10.1016/j.knosys.2017.05.007
Reference: KNOSYS 3904

To appear in: *Knowledge-Based Systems*

Received date: 11 May 2016
Revised date: 7 May 2017
Accepted date: 8 May 2017

Please cite this article as: Chuan Li, Luiz Ledo, Myriam Delgado, Mariela Cerrada, Fannia Pacheco, Diego Cabrera, René-Vinicio Sánchez, José Valente de Oliveira, A Bayesian approach to consequent parameter estimation in probabilistic fuzzy systems and its application to bearing fault classification, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.05.007

A Bayesian approach to consequent parameter estimation in probabilistic fuzzy systems and its application to bearing fault classification

Chuan Li[1], Luiz Ledo[2], Myriam Delgado[2], Mariela Cerrada[3,4], Fannia Pacheco[4], Diego Cabrera[4], René-Vinicio Sánchez[4] and José Valente de Oliveira *[1,5]

[1]National Research Base of Intelligent Manufacturing Service, Chongqing Technology and Business University, China

[2]Universidade Tecnológica Federal do Paraná, Curitiba, Brasil

[3]CEMISD, Universidad de Los Andes, Mérida, Venezuela

[4]Universidad Politécnica Salesiana, Cuenca, Ecuador

[5]On Sabbatical leaving from CEOT, Universidade do Algarve, Faro, Portugal. Email: jvo@ualg.pt

May 12, 2017

*Corresponding author

Abstract

A bearing is an essential component in rotating machinery, one of its principal cause of failure, and its health condition is directly related to the safety and effective operation of such machinery. To the best of our knowledge, it is the first time that a probabilistic fuzzy system is applied to bearing fault classification. The type of probabilistic fuzzy classifier considered is a parsimonious fuzzy rule based model where each rule can diagnose a set of faults each one with its probability. For this kind of real world application, it is desirable to develop interpretable and accurate MIMO fuzzy systems, able to deal with the dimensionality and uncertainty present in data (vibration signals). For parameter estimation we adopt a two steps sequential state-of-the-art data-driven method. First, the antecedents of the rules are estimated using an iterative supervised clustering algorithm. Based on the antecedents the consequent parameters are then estimated. For this, a new method for consequent estimation is proposed. This is based on the observation that for defining a rule consequent not all training data within the region of applicability of that rule are equality relevant. Two criteria for selecting a relevant region in the feature space for consequent parameter estimation are proposed. Results show a statistically significant improvement on the performance of probabilistic fuzzy diagnosers trained with the proposed method, independently of the criterion used for defining the relevant region, when compared with the above mentioned state-of-the-art method. Moreover, the proposed consequent parameter estimation method practically has no overhead on the overall training of the diagnoser. Results show that

an equilibrium can be found between the model level of detail and its accuracy. However, when accuracy is the sole comparison criterion, the proposed probabilistic fuzzy systems systematically matches the performance of other data-driven models like distance based methods (K-nearest neighbors), connectionists (probabilistic neural networks), or maximum margin classifiers (support vector machines).

Keywords: Clustering; Probabilistic Fuzzy systems; Vibration analysis; Bearing; Fault classification

1 Introduction

A bearing is a mechanical component used to reduce friction between other mechanical moving parts. Bearings are one of the most common components in mechanical equipment and one of the principal cause of its malfunction (Yaqub, Gondal, and Kamruzzaman, 2012). For example, in induction motors metal bearing faults account up to 40% of the faults (Siyambalapitiya and McLaren, 1990). This makes bearing fault diagnosis (i.e., detection, classification and prognosis) an economically very relevant topic. Moreover, the literature on bearing fault diagnosis is extensive which indicates that the topic is also scientific and technically challenging.

Rolling element bearings, such as ball bearings, consist of an inner, an outer race or ring, inside which there is a cage of holding the rolling elements, see Fig. 1. All these elements are prone to faults due to excessive load, lubricant failure, fatigue, corrosion, or other causes. The bearing health condition is directly related to the safety and effective operation of mechanical

Figure 1: An industrial ball bearing revealing its components: the outer and inner races together with the cage holding the rolling elements.

systems (Li et al., 2016b). For instance, should the metal engine bearings supporting a crankshaft fail the whole engine can disintegrate.

A fault can be classified according to its location and to its type. For instance, it can be a single point fault in the outer race as illustrated in Fig. 2. Often, general failures originate from incipient single point faults such as this. Thus, an early diagnosis of incipient faults is deemed necessary and this is the main motivation to focus this study on single point faults. Different faults can and do occur simultaneously and this possibility is also considered in this study. Bearing fault diagnosis involves at least the following stages:

Figure 2: An incipient single point fault in the outer race of a bearing.

data acquisition and conditioning, feature extraction and selection, and fi-

nally classification. Fuzzy formalisms have been used in all of these stages as a framework for dealing with the inherent uncertainty of the feature space. Actually, i) the number of faulty samples is much smaller than healthy samples; ii) there is no guarantee that all relevant features are fully observable; iii) the interference between different faults is not easily identified; iv) and measurement is often noisy. Therefore, the knowledge that a fault diagnoser holds about the system is necessarily incomplete and uncertain. However, another important advantage of fuzzy models when compared with other nonlinear modeling and detection techniques such as artificial neural networks, is that fuzzy models provide an insight on the linguistic relationship between the variables of the system (Valente de Oliveira, 1999); an issue that is often forgotten with some notable exceptions though (Zio and Gola, 2006; Zio and Gola, 2009).

A diversity of fuzzy formalisms has been applied to the bearing fault diagnosis, including the neuro-fuzzy approaches, e.g., (Kaplan, Kuncan, and Ertunc, 2015; Tiwari, Gupta, and Kankar, 2015; Marichal et al., 2011), clustering, e.g.,(Wang et al., 2014; Li et al., 2016a; Wei et al., 2017), application of fuzzy measures, and in particular of fuzzy entropy, e.g.,(Liu, Ma, and Mathew, 2009; Zheng et al., 2014), fuzzy support vector machines, e.g., (Zhao and Wang, 2010), possibility and Dempster-Shafer evidence theory, e.g., (Wang and Chen, 2011; Xu, Tan, and Zhan, 2014), fuzzy relations and fuzzy relation equations, e.g., (Yu and Liu, 2011), fuzzy fusion of multiple criteria, e.g., (Boutros and Liang, 2007; Liu, Ma, and Mathew, 2009), fuzzy numbers, e.g., (Juuso, Ruusunen, and Perigot, 2010), rough sets and fuzzy rough sets, e.g., (Zhao et al., 2008; Tian et al., 2012), semi-supervised

approaches, e.g., (Huang et al., 2014), ensembles of fuzzy classifiers, e.g., (Xu, Wu, and Shi, 2013), fuzzy grey-optimization methods for short-term fault prediction(Zhang, Wang, and Zhao, 2007), fuzzy similarity operators to compare two time-domain phase trajectories (Sun et al., 2012), fuzzy lattice based diagnoser (Li et al., 2012a), and health degree evaluation index based on fuzzy sets (Yang et al., 2012; Amar, Gondal, and Wilson, 2013).

In rule based bearing fault diagnosis either Mamdami or Sugeno models are used, see e.g., (Silva Vicente, Fujimoto, and Padovese, 2001; Bediaga et al., 2013; Liu, Wang, and Golnaraghi, 2010). In this work we propose an innovative approach to bearing fault classification, i.e., we propose a parsimonious type of fuzzy rule based model where each rule can diagnose a set of faults each one with an associated probability. This type of model is known as a probabilistic fuzzy system (Meghdadi and Akbarzadeh-T, 2001; Berg, Kaymak, and Almeida, 2013) and is usually composed by a set of rules combining linguistic information in the antecedents with probabilities in the consequents. Each rule can be viewed as describing a fuzzy region in the feature space where the consequent probability distribution over predicted classes is valid. That is, the j-th rule, $r^{(j)}$, can be informally viewed as:

$$
\begin{aligned}
r^{(j)} \;:=\; & \text{if the input belongs to fuzzy region}^{(j)} \\
& \text{then } \hat{y}^{(j)} = c_1 \text{ with prob. } p_1, \\
& \cdots, \\
& \hat{y}^{(j)} = c_C \text{ with prob. } p_C \; [w^{(j)}]
\end{aligned}
$$

where $\hat{y}^{(j)}$ is the rule output, $c_1, \ldots, c_C$ are class labels, and $w^{(j)}$ is a certainty

factor representing a belief in the accuracy of the rule. Although we will be using this type of probabilistic fuzzy system, it should be clear that other types are available, e.g., (Liu and Li, 2005; Aggarwal, 2016).

In real-world applications like bearing fault diagnosis there are various types of uncertainty to be handled. Uncertainties can result from partially observable dynamics, insufficient data, or coarse or noisy measurements, for instance (Zhang and Li, 2012). While stochastic modeling methods can tackle stochastic uncertainty, fuzzy systems are useful to handle incomplete and vague information. Therefore, probabilistic fuzzy systems seem particularly suitable to cope with complex, stochastic, and vague environments (Meghdadi and Akbarzadeh-T, 2001; Berg, Kaymak, and Almeida, 2013; Berg, Kaymak, and Bergh, 2004). Actually, probabilistic fuzzy systems have been successfully applied to real-world problems, e.g., in financial market analysis (Berg, Kaymak, and Bergh, 2004; Almeida and Kaymak, 2009), robotics (Liu and Li, 2005), process modeling and control (Zhang and Li, 2012), or more recently to predict the mortality of septic shock patients (Fialho et al., 2016). However, to the best of our knowledge, it is the first time that a probabilistic fuzzy system is applied to any type of bearing fault diagnosis, i.e., detection, classification or prognosis of faults. This is unfortunate as i) the problem requires a multi-input multi-output (MIMO) model with typically one output for each fault, ii) the usual Mamdami or Sugeno rule based models do not scale sufficiently well in the MIMO case, iii) the larger the model the more difficult is to interpret it, and iv) the more prone is to overfitting. The main contributions of the work can be stated as follows:

- The application of a parsimonious and accurate probabilistic fuzzy sys-

tem to the fault classification in bearing diagnostics. The employment of a fault diagnoser exhibiting reduced computational complexity is particularly relevant in this application as both the number of input variables and output classes can be large.

- A new parameter estimation method for the rule consequents. This is based on the observation that for defining a rule consequent not all training data points within its activation region are equality relevant. Criteria for selecting relevant data points are proposed revealing superior performance relatively to the currently used methods.

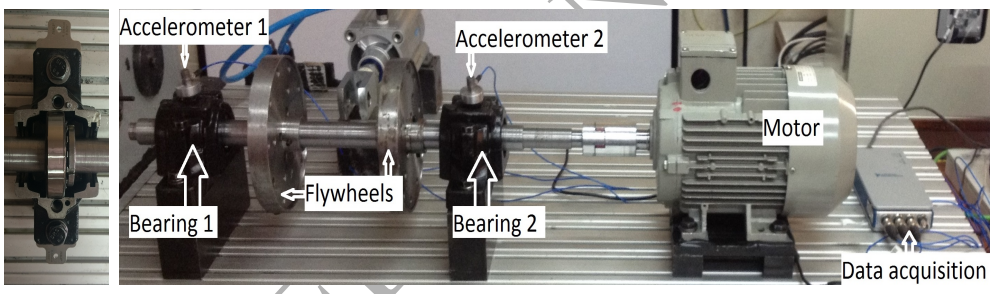
The remaining of the paper is organized as follows. Section 2 describes both the experimental apparatus and the theoretical background required in this study. Section 3 presents the proposed fuzzy probabilistic diagnoser. Section 4 presents the proposed method for consequent estimation. Results and discussion are presented in Section 5. Conclusions end the paper.

2 Material and methods

This section briefly describes the experimental apparatus used for data acquisition, the signal processing used for feature extraction, and the employed method of feature selection. Data considered in this work consist of vibration signals from which time, frequency, and time-frequency domain features are computed. A total of 1634 features are computed; see Section 2.1 for further details. For feature selection a decision tree-like entropy based criterion is used (Section 2.3).

2.1 Experimental setup

The experimental setup used to collect data is shown in Fig. 3 and can be briefly described as follows. Two bearings are mounted in their housings and installed in a shaft driven by an electric motor. Up to two flywheels can be mounted on the shaft for load purposes. One accelerometer is installed in each bearing housing for measuring the vibration signals that are collected by the data acquisition card. This setup allows us to analysis the interferences of bearing 1 faults in accelerometer 2, and the interferences of faults in bearing 2 in accelerometer 1; an usual phenomenon in industrial setups. Table 1 presents the specifications of the used equipment.



(a) (b)

Figure 3: The experimental apparatus. The sub figure (a) is a top view of a bearing installed in its housing.

An experiment is characterized by a tuple $< speed, load, bs >$ where *speed* is the shaft speed, *load* is the total load on the shaft, and *bs* stands for the bearing status. Three discrete speeds are tested: 8, 10, and 15Hz. Also, three different types of loads are essayed: zero, one, and two flywheels. Seven bearing status are considered – see Table 2. Given the low variability of the results, each experiment is only repeated 5 times. Thus, a total of $3 \times 3 \times 7 \times 5 = 315$ experiments are performed.

Table 1: Specifications of the equipment used in the experimental apparatus

Component	**Specification**
Accelerometer 1 & 2	PCB ICP 353c03
Bearing 1 & 2	SKF 1207 Ektn9/C3
Data Acquisition	NI Cdaq-9234
Housing 1 & 2	SKF Snl 507-606
Inverter	Danfoss VLT 1.5kw
Motor	Siemens 1LA7 090-4YA60 2Hp
Shaft diameter	30mm
Tachometer	Vls5/T/Laser Optical Sensor
Type of load	flywheels

Table 2: Health status of the essayed bearings

Id	**Bearing 1**	**Bearing 2**
P1	healthy	healthy
P2	inner race fault	healthy
P3	outer race fault	healthy
P4	ball fault	healthy
P5	inner race fault	outer race fault
P6	inner race fault	ball fault
P7	outer race fault	ball fault

Table 3: Used time-domain features for a signal x with mean μ and duration N.

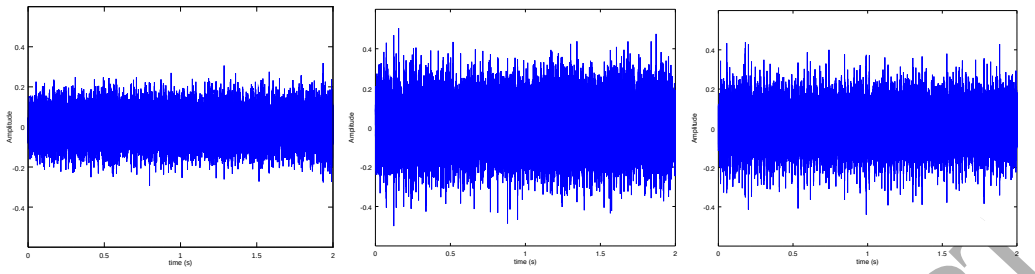
Feature	**Formula**
root mean square	$\mathrm{rms}[x] = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$
variance	$\sigma^2[x] = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
kurtosis	$\mathrm{k}[x] = \frac{N\sum_{i=1}^{N}(x_i-\mu)^4}{[\sum_{i=1}^{N}(x_i-\mu)^2]^2}$
kurtosis of the speed	$\mathrm{ks}[x] = \mathrm{k}[\frac{d}{dt}x(t)]$
kurtosis of the derivative of the acceleration	$\mathrm{kda}[x] = \mathrm{k}[\frac{d^3}{dt^3}x(t)]$
skewness	$\mathrm{s}[x] = \frac{\sum_{i=1}^{N}(x_i-\mu)^3}{N\sigma^3}$
crest factor	$\mathrm{cf}[x] = \frac{\max[x]}{\mathrm{rms}[x]}$

The sampling frequency is 50kHz, well above the requirements of the Sampling theorem, and the duration of each sample (measurement time) is 20s.

2.2 Computed features

For the signals measured in each one of the accelerometers, time, frequency, and time-frequency features are computed. The 7 considered time domain features are summarized in Table 3. A vibration time-domain signal is transformed into a frequency domain one using Fast Fourier Transform (FFT). The frequency domain signals are divided in 80 numerated 312.5Hz-bands and fea-

tures comprising mean, root mean squares, standard deviation, and kurtosis are computed for both linear and db amplitudes. Moreover, 15 octaves are considered and for each one of them: mean, standard deviation, and kurtosis are computed for both linear and db amplitudes; Five wavelets packet transforms are computed: Biorthogonal (bior6.8), Coiflets (coif4), Daubechies (db7), Symlets (sym3), and Reverse Biorthogonal (rbio6.8). These allows one to assess time-frequency information such as high frequency transients, that might be missed by FFT. Summing up, all this amounts to a total of 1634 features. Other features could have been considered, but the above ones were chosen based on our *a priori* knowledge on the their suitability for this application (Li et al., 2012b; Li and Liang, 2012; Li, Liang, and Wang, 2015; Wang et al., 2015; Li et al., 2016a). Furthermore, other type of sensors could also be considered, e.g., acoustic, electric, thermal, or oil debris signals, however vibration measurement and analysis is a well-know and widely applied technique in bearing monitoring and is also used here. Actually, comparatively to other type of signals, vibration is a cost-effective way of assessing faults and operating conditions in rotating components. As an illustration, Fig. 4 shows raw vibration signals corresponding to the healthy state under three slightly different experimental conditions, i.e., (a) $< speed = 8\text{Hz}, load = \text{none}, bs = \text{P1} >$, (b) $< speed = 10\text{Hz}, load = \text{none}, bs = \text{P1} >$, and (c) $< speed = 8\text{Hz}, load = \text{one flywheel}, bs = \text{P1} >$. From the figure, it is clear that these different operating conditions are reflected in the measured vibration signals.



(a) (b) (c)

Figure 4: Signals measured by accelerometer 1 corresponding to the healthy state under different operating conditions: (a) $< speed = 8\text{Hz}, load = \text{none}, bs = \text{P1} >$, (b) $< speed = 10\text{Hz}, load = \text{none}, bs = \text{P1} >$, and (c) $< speed = 8\text{Hz}, load = \text{one flywheel}, bs = \text{P1} >$. Clearly, these different conditions are reflected in the measured signals.

2.3 Feature selection

Feature selection is a critical step for optimizing efficiency, accuracy and for mitigating overtraining. Feature selection can be accomplished by the employment of genetic algorithms, e.g., (Lei et al., 2007), correlation-based methods such principal component analysis (PCA), e.g., (Xu et al., 2009; Vijay et al., 2013; Ben Ali et al., 2015), fuzzy measures, e.g., (Liu et al., 2008), rough sets (Zhao et al., 2008), orthogonal fuzzy neighborhood discriminant analysis, e.g., (Abed, Sharma, and Sutton, 2014), or entropy based criteria like those used for growing decision trees, e.g.,(Genuer, Poggi, and Tuleau-Malot, 2010; Li et al., 2016a) that compute the information degree contributed by each feature. In our case, this entropy-based method was found to be faster and more discriminative than the usually employed genetic algorithm based feature selection methods. In particular, the entropy-based method adopted in this work selects only 12 relevant features out of 1634; see Section 5 and (Li et al., 2016a) for details.

3 A probabilistic fuzzy system as fault diagnoser

As previously discussed, we propose the application of a probabilistic fuzzy system, motivated by the need for using an easy to interpret and accurate MIMO fuzzy system, able to deal with the type of collected data.

This section presents the proposed fuzzy probabilistic diagnoser, its inference, and the proposed method for parameter estimation.

To the best of our knowledge, it is the first time that a probabilistic fuzzy system is applied to bearing fault diagnosis. The type of probabilistic fuzzy classifier that we are interested in is a parsimonious fuzzy rule based model where each rule can diagnose a set of faults each one with its probability (Meghdadi and Akbarzadeh-T, 2001; Berg, Kaymak, and Almeida, 2013). Each rule can be viewed as describing a fuzzy region in the feature space where the probabilities in the consequents are valid. More formally, the j-th rule, $r^{(j)}, j = 1, \ldots, M$ can be specified as:

$$\begin{aligned} r^{(j)} \quad := \quad & \text{if } x_1 \text{ is } A_1^{(j)} \text{ and } \ldots \text{ and } x_d \text{ is } A_d^{(j)} \\ & \text{then } \hat{y}^{(j)} = c_1 \text{ with } p(c_1|r^{(j)}), \cdots, \hat{y}^{(j)} = c_C \text{ with } p(c_C|r^{(j)}) \ [w^{(j)}] \end{aligned} \tag{1}$$

where $\vec{x} = [x_1, \cdots, x_d]^T$ stands for a real-valued vector with d input features, $\vec{A}^{(j)} = [A_1^{(j)}, \cdots, A_d^{(j)}]^T$ are the respective membership functions representing linguistic terms; $\hat{y}^{(j)}$ is the rule output, $c_1, \ldots, c_C$ are the different label faults (or classes), $p(c_i|r^{(j)}) \in [0, 1]; i = 1, \ldots, C$ is interpreted as the conditional probability that class c_i is inferred by rule j given the occurrence of

$\vec{A}^{(j)}$, $\sum_{i=1}^{C} p(c_i|r^{(j)}) = 1$, and $w^{(j)} \in [0,1]$ is the certainty factor of the rule representing the belief in the accuracy of the rule.

3.1 Inference

Given M rules such as (1) and one input vector $\vec{x}$ the output of the model is computed as follow. First the activation strength $\beta^{(j)}(\vec{x})$ of the j-th rule is computed.

$$\beta^{(j)}(\vec{x}) = w^{(j)}\mathbf{A}^{(j)}(\vec{x}) \tag{2}$$

where $\mathbf{A}^{(j)}(\vec{x})$ represents the degree of fulfillment (or firing rate) of the j-th rule that, when the logic connective *and* is modeled by the t-norm product is:

$$\mathbf{A}^{(j)}(\vec{x}) = \prod_{l=1}^{d} A_l^{(j)}(x_l) \tag{3}$$

Second compute $p(c_i|\vec{x})$ for all $i = 1, \ldots, C$. Third, select for final output $\hat{y}(\vec{x})$ the class label c_{i*} with the highest probability, i.e.,

$$\hat{y}(\vec{x}) = c_{i*} \text{ s.t. } i* = \arg\max_{1 \le i \le C} p(c_i|\vec{x}) \tag{4}$$

For computing $p(c_i|\vec{x})$ two possibilities of aggregating the contribution of the each one of the rules are worth considering:

i) Aggregation by max

$$p(c_i|\vec{x}) = \arg\max_{1 \le j \le M} (\beta^{(j)}(\vec{x}) p(c_i|r^{(j)})) \tag{5}$$

ii) Aggregation by averaging

$$p(c_i|\vec{x}) = \frac{\sum_{j=1}^{M} \beta^{(j)}(\vec{x}) p(c_i|r^{(j)})}{\sum_{j=1}^{M} \beta^{(j)}(\vec{x})} \tag{6}$$

This is similar to the inference in Sugeno models where the rules of the fuzzy model are aggregated using a weighted average.

3.2 Parameter estimation

There are two main parameter estimation methods in fuzzy probabilistic systems. The two steps sequential conditional probability method (Berg, Kaymak, and Bergh, 2002; Abonyi and Szeifert, 2003; Berg, Kaymak, and Bergh, 2004; Lee, Park, and Bien, 2008; Melo, Lucas, and Delgado, 2012; Tang et al., 2012; Ledo, Lucas, and Delgado, 2014) and the global one based on the maximum likelihood criterion (Waltman, Kaymak, and Berg, 2005). Currently the former is much more used than the latter and will be also adopted here. The proposed method belongs to the two sequential steps method. In the first step the parameters in the antecedent of the rules are estimated. Based on these parameters, the second step computes the parameters of the consequents.

The antecedent parameters are crucial for the interpretability of the system. These can be determined with the help of experts on the problem domain or resorting to data-driven optimization methods. When the later methods are used, semantic constraints should be used to ensure that the resulting membership functions are amenable of linguistic interpretation (Valente de Oliveira, 1999). One of the simplest ways of computing the mem-

bership functions of antecedents is to resort to the so-called grid partition. In brief, for n membership functions of a given input variable, its universe is divided in $n-1$ intervals, the membership functions being distributed over each interval ensuring an user defined level of *coverage*, typically 0.5 – informally, coverage is the minimum membership degree that any datum from the universe of the variable has (Valente de Oliveira, 1999). The size of each interval can be related with the distribution of input data. A variety of more sophisticated methods exists though (Valente de Oliveira, 1995).

In this work we adopt one of the most sophisticated data-driven parameter estimation methods for antecedents. This is based on the notion of supervised clustering (Abonyi and Szeifert, 2003; Lee, Park, and Bien, 2008) and is briefly described next.

3.3 On clustering and supervised clustering for antecedent parameter estimation

The clustering problem can be stated as the problem of partitioning a finite multivariate data set $\mathbf{Z} \subset \mathbb{R}^d$ into groups (clusters) so that data in one group are similar to each other and are as dissimilar as possible from data in other groups. The (dis)similarities are evaluated through a suitable metric. In general clustering is an unsupervised process. Following (Abonyi and Szeifert, 2003; Lee, Park, and Bien, 2008) and given the existence of labeled data and the purpose of using clustering for estimating the parameters of the antecedent of the rules, one can resort to the so-called *supervised clustering* where the information on the existent class (i.e., the label) is included in the

data to be clustered. In this case, the clustering outputs are the parameters of the membership functions.

To formalize the above let $\mathbf{Z}= \{\vec{z}_1, \ldots, \vec{z}_j, \ldots, \vec{z}_N\}$ with $\vec{z}_j = [\vec{x}_j; y_j]$ be the set of observations. Associated with observation $\vec{z}_j$ there is a vector $\vec{u}_j$ such that each one of its elements represents the membership of the observation j in each one of the M clusters. Each vector's element, u_{ij}, $i = 1, \ldots, M$, $j = 1, \ldots, N$ takes values in $[0, 1]$, 0 standing for non-membership while 1 corresponds to total membership. Vectors $\vec{u}_j$ can be arrayed as the columns of a $M \times N$ partition matrix, $\mathbf{U} = [u_{ij}] \in \mathbb{R}^{M \times N}$. For the sake of correctness of matrix $\mathbf{U}$, the following conditions should be satisfied:

$$u_{ij} \in [0, 1] \text{ for all } i = 1, \ldots, M \text{ and } j = 1, \ldots, N \tag{7a}$$

$$\sum_{i=1}^{M} u_{ij} = 1 \text{ for all } j = 1, \ldots, N \tag{7b}$$

The center of the i cluster is denoted by $\vec{v}_i \in \mathbb{R}^d$. These centers can be arrayed into $\mathbf{V} = [\vec{v}_1, \ldots, \vec{v}_M]$ with $\mathbf{V} \subset \mathbb{R}^{M \times d}$. A cost functional can be defined as:

$$J = \sum_{i=1}^{M} \sum_{j=1}^{N} u_{ij}^m D_{ij}^2 \tag{8}$$

under constraints (7a) and (7b) and $m > 1$ is the so-called fuzzifier coefficient that controls how overlapped the clusters can be, $m = 1$ corresponding to no overlap at all. Moreover, D_{ij} measures the dissimilarity between the datum $\vec{z}_j$ and the center of i-th cluster $\vec{v}_i$. Different definitions of D_{ij} give rise to different clustering algorithms. For instance when D_{ij} is defined as the Euclidean distance, the resulting algorithm is the well-known fuzzy c-means

(FCM). See (Valente de Oliveira and Pedrycz, 2007) for a comprehensive treatment of this and other advanced fuzzy clustering algorithms.

Following (Abonyi and Szeifert, 2003) one can adopt a variation of the metric used by the Gath-Geva clustering algorithm. This allows one to estimate directly the parameter of the Gaussian membership functions in the antecedents of the rules. More concretely, the following metric is adopted:

$$\frac{1}{D_{jk}^2} = p(r^{(j)}) \frac{1}{|2\pi \mathbf{F}_j|^{d/2}} \exp[(\vec{x}_k - \vec{v}_j)^T (\mathbf{F}_j^{-1})(\vec{x}_k - \vec{v}_j)] p(y_k | r^{(j)}) \tag{9}$$

Once D_{jk} is specified the optimization problem of minimizing (8) under constraints (7a) and (7b) can be converted into an unconstrained one using Lagrange multipliers. The corresponding optimization algorithm can be viewed as a sequence of Picard iterations through the necessary conditions of the unconstrained cost and is given in the Appendix as Algorithm 2.

4 The proposed method for consequent estimation

Within the feature space $\mathbf{X} \subset \mathbb{R}^d$, the j-th rule has a region of influence or activation region defined by the support of the fuzzy relation

$$R^{(j)}(\vec{x}) = \{(\vec{x}, \mathbf{A}^{(j)}(\vec{x})) | \vec{x} \in \mathbf{X}\} \tag{10}$$

where $\mathbf{A}^{(j)}(\vec{x})$ is given by (3) and the support by $\text{supp}(R^{(j)}) = \{\vec{x} | \vec{x} \in \mathbf{X} \wedge \mathbf{A}^{(j)}(\vec{x}) > 0\}$.

Let $\mathbf{X}= \{\vec{x}_1, \ldots, \vec{x}_k, \ldots, \vec{x}_N\}$ be a finite (training) set of feature vectors in the $\mathbb{R}^d$ space. Associated with each feature vector $\vec{x}_k$ there is a label $y_k \in \{c_1, \ldots, c_C\}$ representing the fault type (or class) of the k-th input. Up to now, to estimate $p(c_i|r^{(j)})$ the following expression has been used, e.g., (Berg, Kaymak, and Bergh, 2002; Abonyi and Szeifert, 2003; Waltman, Kaymak, and Berg, 2005; Tang et al., 2012):

$$p(c_i|r^{(j)}) = \frac{\sum_{k=1}^{N} \mathbf{A}^{(j)}(\vec{x}_k)\chi_i(y_k)}{\sum_{k=1}^{N} \mathbf{A}^{(j)}(\vec{x}_k)} \tag{11}$$

where $\chi_i(y_k) = 1$ if $c_i = y_k$; 0 otherwise.

In our view, within the region of influence (or activation region) of a rule, different data points can have potentially different impacts on the consequent parameter estimation of that rule. For example, a *typical* point (Lesot and Kruse, 2006) in a region of influence should be *much* more relevant for consequent parameter estimation than a peripheral one that simultaneously belongs to more than one region (i.e., more than one rule). Following this observation, and informally, our proposal for the consequent parameter estimation of a rule is to consider *not* the entire region of influence of a rule, but only a *subregion* within the region of influence that best represents the rule. For characterizing this subregion, the concept of *relevant* region otherwise known as the "ideal" region in (Melo, Lucas, and Delgado, 2012) is now consolidated and generalized.

4.1 On the relevant region

The relevant region can be defined in several different and interesting ways within the region of influence. The relevant region of the j-th rule, $\mathcal{R}^{(j)}$, can be defined as the (week) α-cut of $R^{(j)}$, i.e.,

$$\mathcal{R}^{(j)} = R^{(j)}_{\alpha} = \{\vec{x} | \vec{x} \in \mathbf{X} \wedge \mathbf{A}^{(j)}(\vec{x}) \geq \alpha\} \tag{12}$$

By selecting a value of $\alpha \in (0, 1]$ we are selecting a relevant region between the support and the core of $R^{(j)}$, respectively. As we increase α we are being more and more selective on the region of the feature space that we use to estimate the consequent parameters of a given rule. For $\alpha = 1$ we only consider points in a region for which there is a complete match with the antecedent of the rule.

The above definition is crisp. However, it can be softened resorting to a *contrast intensification* operator over $R^{(j)}$, $\mathrm{CI}_p(R^{(j)})$. When applied to a fuzzy set A of membership function $\mu_A(x)$, $\mathrm{CI}_p(A)(x)$ can be given by:

$$\mathrm{CI}_p(A)(x) = \begin{cases} 2^{p-1}[\mu_A(x)]^p & \text{if } \mu_A(x) \leq 0.5 \\ 1 - [2^{p-1}(1 - \mu_A(x))^p] & \text{otherwise} \end{cases} \tag{13}$$

with $p \geq 1$. That is, for $p = 1$, nothing changes in the membership function; for $p > 1$, $\mathrm{CI}(A)$ further increases the membership values above the inflection point 0.5, while decreases the membership values below that value. Thus, an

alternative definition for the relevant region can be:

$$\mathcal{R}^{(j)} = \mathrm{CI}_p(R^{(j)}) \tag{14}$$

where the argument $R^{(j)}$ of the modifier CI is given by (10). This definition has the merit of considering all data points within the region of influence of the rule while increases the relevance of typical points and decreases the relevance of peripheral ones.

At this point one can argue that data with higher membership values already have higher relevance. Moreover, if the optimized membership values do not reflect the required high relevance then suitable requirements should be incorporated in the optimization task ensuring the required high membership values. In other words, one may ask: isn't the proposed strategy equivalent to properly tune the membership functions? The point is that as elicited from the clustering process, membership functions define the region of influence for each rule, i.e., its activation region, that there is no reason to modify. A different issue is the selection of data used to estimate the consequent parameters of each rule. For this we restrict ourselves to data within the activation region with sufficiently high membership values, i.e., the relevant region. Should we change the originally elicited membership functions to directly reflect the relevant region and we would end up with a reduced activation region, what may leave data uncovered by any rule, and thus unclassified. This is illustrated using a data set with two classes sampled from bi-variate normal distributions such that can be classified using two rules. After applying the clustering process two clusters are identified from which

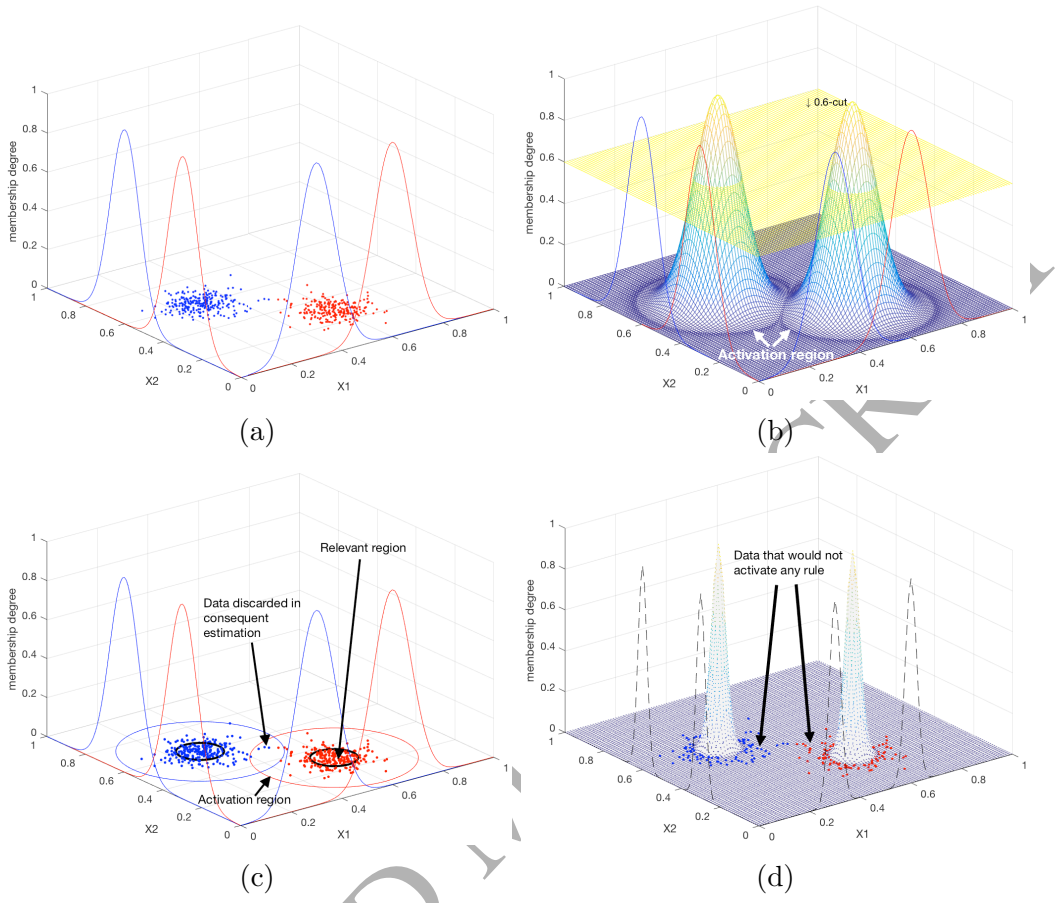


Figure 5: Example of a bi-variate data set with two classes with (a) membership functions elicited by the clustering process, (b) the corresponding rules and activation regions, and (c) relevant region of the rules selecting the data used for consequent parameter estimation. Narrowed activation regions would result if membership functions have being used to define the same relevant regions leaving uncovered unclassified data (d).

rule antecedent membership functions are elicited as represented in Fig. 5a. These define the activation region of each rule as illustrated in Fig. 5b. In this figure it is also represented a 0.6-cut used to defined the relevant region of each rule – the region from which data are used to estimate the consequent parameters of the rule. Fig. 5c shows both the activation and relevant region of the rules. Would not be possible to directly elicit other membership func-

tions that directly define the same relevant regions? Absolutely. But that would narrow the rule activation regions leaving uncovered unclassified data as per Fig. 5d.

But again what is the rationale beyond the introduction of the relevant region? After all, data with higher membership values already have higher relevance. First, consider a problem such that the resulting rules have disjoint activation regions. In this case, data used to estimate the consequent of a rule are the data that activate only that rule. It would not make any sense to use any other data to estimate the consequent of that rule. Why? Because when a test datum is presented for classification such that it has full match with a given rule (i.e., that it fully activates the rule), the derived conclusion should only reflect the training data within the region of activation of that rule. Now, consider the general case where there is some overlapping between two or more activation regions as in Fig. 5c. When a test datum exhibiting full match with a given rule (and none in any other) is presented the derived conclusion should only reflect the training data that activates that rule only. What the proposed relevant region is aiming at is to characterize the region of the feature space that activates a single rule. Only the data in that region are used for rule's consequent estimation. And what about data lying outside any relevant region? These data will not be used for estimation of the consequent but are fundamental for the definition of the antecedent membership functions as illustrated in Fig. 5. When a test datum appears for classification outside any relevant region, it will activate partially one or more rules and the corresponding final result will be given by the aggregation of the contributions of each of the partially activated rules as seen in Section

3.1.

Yet another question could be whether with this strategy we are making the model less and less fuzzy. Here two remarks are in order. At the one hand, the fuzziness of the used probabilistic fuzzy systems relies on the antecedents; not on the consequent of the rules. On the other hand, one thing is the model, another different thing is the data used to estimate the model. By using a relevant region for estimating the consequent of a rule such as (1) we are improving the probability estimate of the most likely classes within the region of influence of the rule. In contrast, by considering only plain membership values as seen in (11), the conventional approach tends to estimate more uniform probabilities over all classes within the activation region of each rule. With more uniform probabilities, a higher variance classifier is obtained, and thus comparatively lower predictive capabilities.

Typically, some features are more relevant than others. Similarly, there are some regions of the feature space which are more relevant for consequent estimation than others. The aim of defining a relevant region is twofold: To improve both the predictability and the interpretability of the model. These and other issues are further considered in Section 4.3 and following. Moreover, as no assumptions on the type of features or on the type of applications are made for the method, it is intended to be a general purpose method; not a specific one for bearing fault classification.

4.2 Consequent estimation using the relevant region

Given $\mathcal{R}^{(j)}$ the problem is now to compute $p(c_i|\mathcal{R}^{(j)})$ what can be done using a Bayesian approach:

$$p(c_i|\mathcal{R}^{(j)}) = \frac{p(\mathcal{R}^{(j)}|c_i)p(c_i)}{p(\mathcal{R}^{(j)})} \tag{15}$$

where the likelihood function

$$p(\mathcal{R}^{(j)}|c_i) = \frac{|\mathcal{R}^{(j)} \cap C_i|}{|C_i|} \tag{16}$$

computes the number of times that data of class c_i occurs within the relevant region of rule j. Notice that C_i stands for the set of data point in $\mathbf{X}$ with class label c_i, and $|C_i| = \sum_x \mu_{C_i}(x)$ is its cardinality.

Both priori $\mathcal{R}^{(j)}$ and $p(c_i)$ can be assumed as equiprobable, i.e., $p(\mathcal{R}^{(j)}) = 1/M$ and $p(c_i) = 1/C$. Alternatively, one can use a frequentist approach, i.e., *a priori* probability of $\mathcal{R}^{(j)}$ can be given by $p(\mathcal{R}^{(j)}) = \frac{|\mathcal{R}^{(j)}|}{N}$ while the *a priori* probability of class c_i can be given by $p(c_i) = \frac{|C_i|}{N}$, N being the cardinality of the data set $\mathbf{X}$. Plugging in these two latter estimates into (15) yields:

$$p(c_i|\mathcal{R}^{(j)}) = \frac{|\mathcal{R}^{(j)} \cap C_i|}{|\mathcal{R}^{(j)}|} \tag{17}$$

It is clear that (17) degenerates into (11) when $\mathcal{R}^{(j)}$ is given by (14) with $p = 1$. As we have M rules, we have M probability distribution functions

that can be arrayed into a matrix of probabilities, P, s.t.,

$$P = \begin{bmatrix} p(c_1|\mathcal{R}^{(1)}) & p(c_1|\mathcal{R}^{(2)}) & \dots & p(c_1|\mathcal{R}^{(M)}) \\ p(c_2|\mathcal{R}^{(1)}) & p(c_2|\mathcal{R}^{(2)}) & \dots & p(c_2|\mathcal{R}^{(M)}) \\ \vdots & \vdots & \ddots & \vdots \\ p(c_C|\mathcal{R}^{(1)}) & p(c_C|\mathcal{R}^{(2)}) & \dots & p(c_C|\mathcal{R}^{(M)}) \end{bmatrix} \tag{18}$$

i.e., the j-th column of P ($j = 1, \dots, M$) is the estimate of the probability distribution in the consequent of the j-th rule (1).

Now, for computing the remaining consequent parameters, i.e., the certainty factors $w^{(j)}$, we see that $p(c_i|\vec{x})$ can be viewed as a mixture of density models:

$$p(c_i|\vec{x}) = \sum_{j=1}^{M} p(r^{(j)}|\vec{x}) p(c_i|r^{(j)}) \tag{19}$$

where $p(r^{(j)}|\vec{x})$ is the *posteriori* probability of the j-th model given $\vec{x}$ that, following Bayes, can be given by

$$p(r^{(j)}|\vec{x}) = \frac{p(\vec{x}|r^{(j)}) p(r^{(j)})}{p(\vec{x})} \tag{20}$$

Moreover, let the likelihood function $p(\vec{x}|r^{(j)})$ be given by a multivariate Gaussian function of center $\vec{v}_j$ and covariance matrix, $\mathbf{F}_j$, i.e.,

$$p(\vec{x}|r^{(j)}) = \frac{1}{|2\pi\mathbf{F}_j|^{d/2}} \exp[(\vec{x} - \vec{v}_j)^T (\mathbf{F}_j^{-1})(\vec{x} - \vec{v}_j)] \tag{21}$$

From (3) and assuming also Gaussian membership functions, i.e,

$$A_l^{(j)}(x_l) = \exp[-\frac{1}{2}\frac{(x_l - v_{jl})^2}{\sigma_{jl}^2}]; (j = 1, \ldots, M; l = 1, \ldots, d) \tag{22}$$

one has

$$\begin{aligned} \beta^{(j)}(\vec{x}) &= w^{(j)} \exp[(\vec{x} - \vec{v}_j)^T (\mathbf{F}_j^{-1})(\vec{x} - \vec{v}_j)] \\ &= w^{(j)} \mathbf{A}^{(j)}(\vec{x}) \end{aligned} \tag{23}$$

That is, $\mathbf{A}^{(j)}(\vec{x})$ is now the kernel of a multivariate Gaussian with center $\vec{v}_j = [v_{j1}, \ldots, v_{jd}]^T$ and diagonal covariance matrix $\mathbf{F}_j$ with elements σ_{ij}^2. From (21):

$$\mathbf{A}^{(j)}(\vec{x}) = p(\vec{x}|r^{(j)})|2\pi\mathbf{F}_j|^{d/2} \tag{24}$$

seeing that

$$\beta^{(j)}(\vec{x}) = p(\vec{x}|r^{(j)})p(r^{(j)}) \tag{25}$$

from (23) and (24) one has:

$$w^{(j)} = \frac{p(r^{(j)})}{|2\pi\mathbf{F}_j|^{d/2}} \tag{26}$$

The outline of the proposed parameter estimation method is given in Algorithm 1.

Algorithm 1: The proposed parameter estimation algorithm

Input : Feature set: $\mathbf{X} \subset \mathbb{R}^d$, s.t., $\mathbf{X} = \{\vec{x}_1, \ldots, \vec{x}_k, \ldots, \vec{x}_N\}$; Label set, $\vec{y} = \{y_1, \ldots, y_k, \ldots, y_n\}$ where $y_k \in \{c_1, \ldots, c_i, \ldots, c_C\}$, C being the number of classes.

Output: Consequents parameters: probability distribution ($p(c_i|\mathcal{R}^{(j)})$) and certainty factor ($w^{(j)}$) for each rule j ($j = 1, \ldots, M$).

Estimate v_{jl} and σ^2_{jl} of membership functions (22) ;

Set the relevant region, $\mathcal{R}^{(j)}$, for each and every rule using either (12) or (14) ;

for *j=1* **to** M **do**

Compute the *a priori* probability of rule j:

$$p(r^{(j)}) = \frac{|r^{(j)}|}{N}$$

Compute the certainty factor:

$$w^{(j)} = p(r^{(j)}) \prod_{l=1}^{d} \frac{1}{\sqrt{2\pi\sigma^2_{jl}}}$$

for *i=1* **to** C **do**

Compute the conditional probability distribution, i.e., the elements of P in (18):

$$p(c_i|\mathcal{R}^{(j)}) = \frac{|\mathcal{R}^{(j)} \cap C_i|}{|\mathcal{R}^{(j)}|}$$

end

end

4.3 On the predictive capabilities

In a first attempt to compare the predictive capabilities of the proposed and the conventional method consider the following.

Proposition 1. *Inference using (17) generalizes (6).*

Proof.

$$\begin{aligned} p(c_i|\vec{x}) &= \sum_{j=1}^{M} p(c_i, \mathcal{R}^{(j)}, r^{(j)}|\vec{x}) \\ &= \frac{1}{p(\vec{x})} \sum_{j=1}^{M} p(c_i, \mathcal{R}^{(j)}, r^{(j)}, \vec{x}) \\ &= \sum_{j=1}^{M} p(c_i|\mathcal{R}^{(j)}, r^{(j)}, \vec{x}) p(r^{(j)}|\mathcal{R}^{(j)}, \vec{x}) p(\mathcal{R}^{(j)}|\vec{x}) \end{aligned} \tag{27}$$

as $p(\mathcal{R}^{(j)}|\vec{x}) = p(\mathcal{R}^{(j)})$ ($\mathcal{R}^{(j)}$ does not depend on input test data $\vec{x}$; it depends on training data); $p(r^{(j)}|\mathcal{R}^{(j)}, \vec{x}) = p(r^{(j)}|\vec{x})$ ($r^{(j)}$ is conditionally independent of $\mathcal{R}^{(j)}$ given input data $\vec{x}$) and $p(c_i|\mathcal{R}^{(j)}, r^{(j)}, \vec{x}) = p(c_i|\mathcal{R}^{(j)})$

$$p(c_i|\vec{x}) = \sum_{j=1}^{M} p(c_i|\mathcal{R}^{(j)}) p(r^{(j)}|\vec{x}) p(\mathcal{R}^{(j)}) \tag{28}$$

and

$$p(r^{(j)}|\vec{x}) = \frac{p(\vec{x}|r^{(j)})p(r^{(j)})}{p(\vec{x})} = \frac{p(\vec{x}|r^{(j)})p(r^{(j)})}{\sum_{r=1}^{R} p(\vec{x}|r^{(j)})p(r^{(j)})} \tag{29}$$

Plugging in this into (28) yields:

$$p(c_i|\vec{x}) = \sum_{j=1}^{M} \frac{p(\vec{x}|r^{(j)})p(r^{(j)})}{\sum_{j=1}^{M} p(\vec{x}|r^{(j)})p(r^{(j)})} p(c_i|\mathcal{R}^{(r)}) p(\mathcal{R}^{(r)}) \tag{30}$$

Taking into account (25), (30) can now be written in terms of activation degrees $\beta^{(j)}$, i.e.,

$$p(c_i|\vec{x}) = \frac{\sum_{j=1}^{M} \beta^{(j)}(\vec{x}) p(c_i|\mathcal{R}^{(j)}) p(\mathcal{R}^{(j)})}{\sum_{j=1}^{M} \beta^{(j)}(\vec{x})} \tag{31}$$

that generalizes (6) as degenerates into it under mild conditions. A sufficient condition being the *a priori* $p(\mathcal{R}^{(j)}) \sim$ unif(0,1) and $\mathcal{R}^{(j)}$ given by (14) with $p = 1$. □

This proposition tells us that it is possible to achieve exactly the same predictive performance of the conventional method when using the proposed method with a suitable chosen relevant region $\mathcal{R}^{(j)}$. The immediate question is whether superior performance can be expected. Consider the decision boundary, i.e., the region of the feature space for which the probability $p(c_i|\vec{x})$ of two or more classes is exactly the same and from which at one side of the boundary the classifier predicts one class and at the other side it predicts another. From (6) it is clear that the classifier generated by the conventional method has a decision boundary that depends only on the membership functions: i) the activation strengths $\beta^{(j)}(\vec{x})$ in (2) depend only on the membership functions (and on the training data) and ii) $p(c_i|r^{(j)})$ as given by (11) also depends only on the membership functions. As soon as the membership functions are obtained the classifier decision boundary is set. As discussed previously and illustrated in Fig. 5, it is not advisable to further change these membership functions.

From (31) it is clear that the classifier generated by the proposed method has a decision boundary that depends on the membership functions, as it also uses the activation strengths $\beta^{(j)}(\vec{x})$, but the boundary depends also on the relevant region. As previously defined this region has one hyper-parameter – α in (12) or p in (14) – that can be adjusted to change the decision boundary in such a way that the classifier performance can be improved. As shown in the illustrative example below and in the results of Section 5, by hand-tuning these hyper-parameters it is possible to obtain a statistically significant improvement in the performance of a classifier using the proposed method relatively to the conventional one, under exactly the same experimental conditions.

4.4 Illustrative example

Consider a binary classification problem whose data set has a single feature defined in the unit interval. More concretely, the data set is composed from observations sampled from two normal distributions one with 0 mean, the other with mean 1, both with the same standard deviation σ. From the possible observations sampled from these distributions the data set retains only those observations within the unit interval. Observations from the same distribution belong to the same class.

As we know that data come from two distinct distributions we use a probabilistic fuzzy classifier with two rules. Moreover, the activation region of the rules are defined by Gaussian membership functions whose parameters

are equal to the parameter of the normal distributions used to generate data. These membership functions can be thought as representing the linguistic terms *Small* and *Large*, respectively. Once the rule's antecedents are set, we can now see how the consequents can be estimated, and how this estimation influences the classifier performance and interpretability.

When $\sigma = 0.3$ the resulting dataset and membership functions are illustrated in Fig. 6a. In this figure, the relevant regions are defined by (12) with a 0.6-cut. Three different regions of the feature space are visible: the regions within which all observations activate a single rule (relevant regions), and the region where data activate both rules.

To illustrate the effect of the relevant region on the performance of the model we vary the α-cut in (12) from 0 to 1 and compare the corresponding model accuracy with a model obtained by the conventional method, using a 10-fold cross-validation plan. Fig. 6b uses boxplots for showing the accuracy distribution for each α-cut. By increasing α we are making the relevant regions smaller, i.e., we are tending for the case where for estimating the consequent of a rule, we are using data that activate only that rule. As $\alpha \to 1$ performance tends to degrade as less data tend to be available for estimation. That is, when using convex fuzzy sets as used in the example, the cardinality of the data in the relevant region decreases as α increases.

For a 0.6-cut a 10-fold cross-validation median accuracy of 100% was obtained with the following classifier:

$$r^{(1)} := \text{if } x \text{ is } Small \text{ then } \hat{y}^{(1)} = c_1 \text{ with } p(c_1|\mathcal{R}^{(1)}) = 1;\ [w^{(1)} = 0.5]$$

$$r^{(2)} := \text{if } x \text{ is } Large \text{ then } \hat{y}^{(2)} = c_2 \text{ with } p(c_2|\mathcal{R}^{(2)}) = 1;\ [w^{(2)} = 0.5]$$

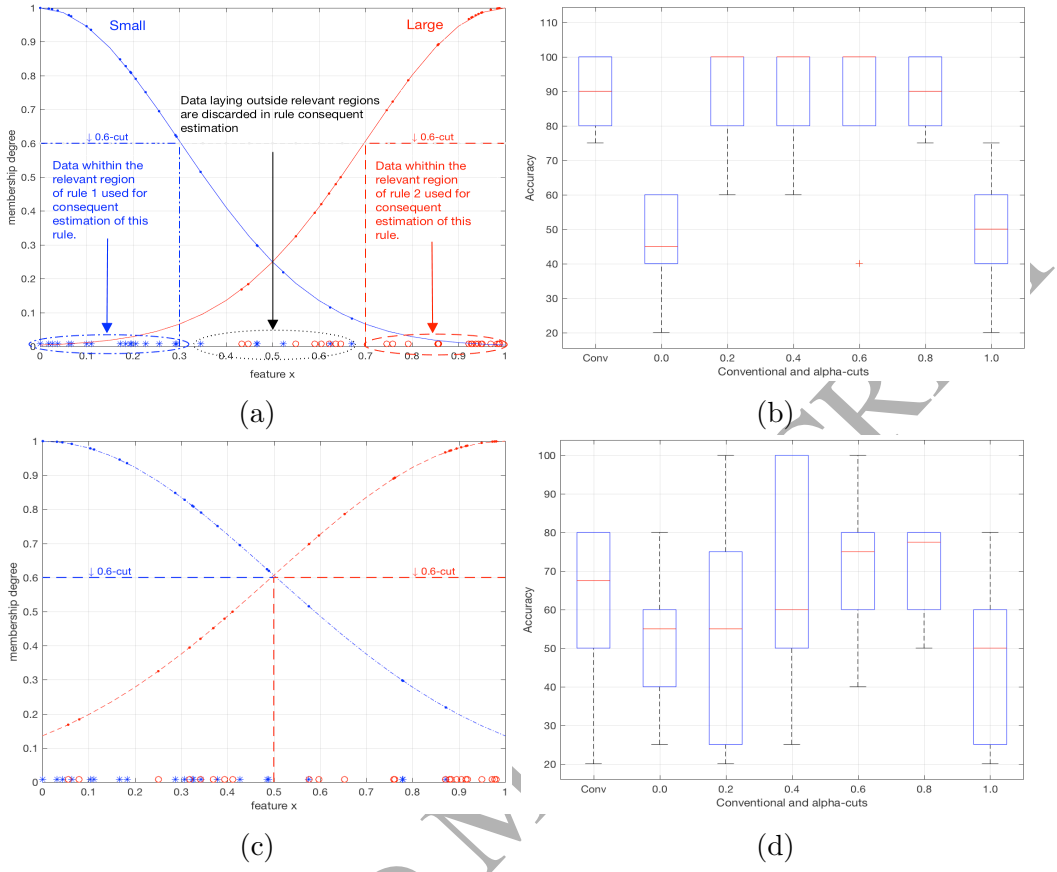


Figure 6: Data sampled from two normal distributions centered in 0 and 1, respectively, equal standard deviation σ, and restricted to [0,1]. Data, membership functions, and 0.6-cut defined relevant regions for (a) $\sigma = 0.3$, and (c) $\sigma = 0.5$. Boxplots showing accuracy distributions for the conventional method and for the proposed method under different α-cuts for (b) $\sigma = 0.3$ and (d) $\sigma = 0.5$.

The conventional method obtained a median accuracy of 90%. The maximum accuracy was obtained for the following classifier:

$$r^{(1)} := \text{if } x \text{ is } \textit{Small} \text{ then } \hat{y}^{(1)} = c_1 \text{ with } p(c_1|r^{(1)}) = 0.9017;$$
$$\hat{y}^{(2)} = c_2 \text{ with } p(c_2|r^{(2)}) = 0.0983; \ [w^{(1)} = 0.458]$$
$$r^{(2)} := \text{if } x \text{ is } \textit{Large} \text{ then } \hat{y}^{(1)} = c_1 \text{ with } p(c_1|r^{(1)}) = 0.1165;$$

$$\hat{y}^{(2)} = c_2 \text{ with } p(c_2|r^{(2)}) = 0.8835;\ [w^{(2)} = 0.5342]$$

Observe that comparatively not only accuracy has suffered. By producing a more uniform estimate of probability distribution of the two classes, the interpretability of the model obtained by the conventional method has suffered as well.

When the test input datum is either 0 or 1 (extreme values of the feature range), the proposed method yields a probability 1 for the class with full match in the rule antecedent, and 0 for the other, fully complying with the problem boundary conditions. By the contrary, the conventional method yields $p(c_1|x = 0) = 0.9017$ and $p(c_2|x = 0) = 0.0983$, and $p(c_1|x = 1) = 0.1165$ and $p(c_2|x = 1) = 0.8835$. At first, the probabilities yielded by the proposed model might seem counter-intuitive. After all, in principle, $x = 0$ might have been generated with a small probability from the normal distribution with the center in 1 (as apparently it is suggested by the conventional model). Notice however that both models were estimated using the preselected labeled data set presented in Fig. 6a where all data near $x = 0$ (represented by *) were generated from the normal distribution centered in 0; not in 1 (*mutatis mutandis* for $x = 1$). Thus none of the models (neither the proposed nor the conventional one) have any data that may indicate that $x = 0$ could have been generated from the normal distribution with center in 1.

Moreover, the decision boundary obtained by the proposed method is correctly positioned at 0.5 while the decision boundary obtained by the conventional method is erroneously positioned in 0.49, i.e., for test input data

greater or equal to 0.49, data start to be classified as class 2. As the antecedents are the same for both methods, the observed bias can only be due to the conventional method of consequent parameter estimation.

A greater performance difference is obtained for wider overlapping regions as illustrated in Figs. 6c and 6d for $\sigma = 0.5$. Notice that in this case both models have access to data examples closed to 0 that was generated by the distributions centered in 0 (represented by *) and centered in 1 (represented by o). Consequently, the proposed model reflects that in the resulting rules (for a 0.6-cut):

$$
\begin{aligned}
r^{(1)} \;:=\; & \text{if } x \text{ is } \mathit{Small} \text{ then } \hat{y}^{(1)} = c_1 \text{ with } p(c_1|r^{(1)}) = 0.6364; \\
& \qquad\qquad \hat{y}^{(2)} = c_2 \text{ with } p(c_2|r^{(2)}) = 0.3636; \;\; [w^{(1)} = 0.5238] \\
r^{(2)} \;:=\; & \text{if } x \text{ is } \mathit{Large} \text{ then } \hat{y}^{(1)} = c_1 \text{ with } p(c_1|r^{(1)}) = 0.2000; \\
& \qquad\qquad \hat{y}^{(2)} = c_2 \text{ with } p(c_2|r^{(2)}) = 0.8000; \;\; [w^{(2)} = 0.4762]
\end{aligned}
$$

In the light of reproducible research, the matlab code used for generating the presented example is publicly available from http://w3.ualg.pt/~jvo/pubs/KNOSYS-D-16-00719R4

5 Results and discussion

This section presents some experimental results and a corresponding brief discussion. Before reporting the application of the proposed methodology to the more realistic experimental apparatus of Section 2.1 where fault inter-

ferences can be studied, we report the application to a simpler but widely used benchmark dataset relative to the 6203-2RS JEM SKF deep groove ball bearing from the Case Western Reserve University (CWRU) Bearing Data Centre (Loparo, 2003). This allows the comparison of performance between our parameter estimation method with i) one of the more recent and sophisticated parameter estimation methods available for probabilistic fuzzy systems (Abonyi and Szeifert, 2003; Lee, Park, and Bien, 2008), as well as representatives of ii) distance based methods, i.e., K-nearest neighbors (KNN), iii) connectionist approaches, i.e., probabilistic neural networks (PNN), and iv) maximum margin classifiers, i.e., support vector machines (SVM).

The comparisons use accuracy (*Acc*) as a measure of classification quality and checks for statistical significant differences between the obtained results using either i) parametric statistical tests whenever their application conditions allow or ii) non-parametric ones, otherwise. In all these tests a confidence level of 95% is used. For a description of the employed statistical methods the reader is referred to (Sheskin, 2011).

5.1 Results for the CWRU 6203 SKF bearing

In the CWRU setup (Loparo, 2003) the 6202-2RS JEM SKF deep groove ball bearing is employed to support the motor shaft at the fan end side. Vibration signals acquired by accelerometers placed at 12 o'clock on the bearing housing, sampled at 12KHz, were measured under 0-load at four successive rotation speeds, i.e., 1730, 1750, 1772, and 1797 rpm. Four health conditions were observed: 0,1778 single fault in i) inner race, ii) outer race,

iii) ball, and iv) no fault. For each of the above operating conditions, 20 data acquisition experiments were performed. These are the same conditions used in (Dou and Zhou, 2016) and other works and comprise a vibration signal data set with 320 samples of 2000 points each. For facilitating the comparison of results with the previously cited work, in all subsequent experiments with this data set, 4-fold cross-validation is also adopted here.

When the proposed methodology is applied to this data set only 2 out of 805 features are selected as the most relevant ones: the time domain rms, and the linear amplitude rms of the FFT band 4.

Three different probabilistic fuzzy classifiers, hereafter referred to as *pfc1*, *pfc2*, and *pfc3*, are tested. Classifier pfc1 is a baseline classifier trained using the method proposed in (Abonyi and Szeifert, 2003; Lee, Park, and Bien, 2008). Classifiers pfc2 and pfc3 are trained with the proposed method with the relevant region computed using i) (14) with $p = 2$, and ii) (12) with $\alpha = 0.2$, respectively. There are no other differences between the tested classifiers. In particular, all of them use (6) for inference.

In addition to these classifiers, 3 others, i.e., KNN (k=5), PNN ($\sigma^2 = 0.001$), and SVM[1] were also added to the comparison. For statistical significance, 30 runs of 4-fold cross-validation are performed for each classifier.

One distinctive characteristic of rule based models such as the proposed fuzzy probabilistic one, is that it is possible to select a trade-off between the number of rules used (and thus interpretability) and accuracy. As the number of rules increases, accuracy also increases (see Table 4) but interpretability

[1]In the tests with KNN, PNN, and SVM we used the respective R function with default hyper-parameters unless stated otherwise.

suffers. This can be illustrated by the obtained membership functions for pfc3 with $M = 10$ as shown in Fig. 7. For this same model a typical rule set is:

$$
\begin{aligned}
r^{(1)} &:= \text{if } x_1 \text{ is } A_1^{(1)} \text{ and } x_2 \text{ is } A_2^{(1)} \\
&\quad \text{then } \hat{y}^{(1)} = c_4 \text{ with } p(c_4|\mathcal{R}^{(1)}) = 1;\ [w^{(1)} = 0.371] \\
r^{(2)} &:= \text{if } x_1 \text{ is } A_1^{(2)} \text{ and } x_2 \text{ is } A_2^{(2)} \text{ and} \\
&\quad \text{then } \hat{y}^{(2)} = c_3 \text{ with } p(c_3|\mathcal{R}^{(2)}) = 1;\ [w^{(3)} = 0.2592] \\
r^{(3)} &:= \text{if } x_1 \text{ is } A_1^{(3)} \text{ and } x_2 \text{ is } A_2^{(3)} \text{ and} \\
&\quad \text{then } \hat{y}^{(3)} = c_3 \text{ with } p(c_3|\mathcal{R}^{(3)}) = 1;\ [w^{(1)} = 0.2816] \\
r^{(4)} &:= \text{if } x_1 \text{ is } A_1^{(4)} \text{ and } x_2 \text{ is } A_2^{(4)} \text{ and} \\
&\quad \text{then } \hat{y}^{(4)} = c_2 \text{ with } p(c_2|\mathcal{R}^{(4)}) = 1;\ [w^{(4)} = 0.06] \\
r^{(5)} &:= \text{if } x_1 \text{ is } A_1^{(5)} \text{ and } x_2 \text{ is } A_2^{(5)} \text{ and} \\
&\quad \text{then } \hat{y}^{(5)} = c_1 \text{ with } p(c_1|\mathcal{R}^{(5)}) = 1;\ [w^{(5)} = 0.0018] \\
r^{(6)} &:= \text{if } x_1 \text{ is } A_1^{(6)} \text{ and } x_2 \text{ is } A_2^{(6)} \text{ and} \\
&\quad \text{then } \hat{y}^{(6)} = c_3 \text{ with } p(c_3|\mathcal{R}^{(6)}) = 1;\ [w^{(6)} = 0.5098] \\
r^{(7)} &:= \text{if } x_1 \text{ is } A_1^{(7)} \text{ and } x_2 \text{ is } A_2^{(7)} \text{ and} \\
&\quad \text{then } \hat{y}^{(7)} = c_2 \text{ with } p(c_2|\mathcal{R}^{(7)}) = 0.11; \hat{y}^{(7)} = c_4 \text{ with } p(c_4|\mathcal{R}^{(7)}) = 0.89 \\
&\quad [w^{(7)} = 0.2247] \\
r^{(8)} &:= \text{if } x_1 \text{ is } A_1^{(8)} \text{ and } x_2 \text{ is } A_2^{(8)} \text{ and} \\
&\quad \text{then } \hat{y}^{(8)} = c_4 \text{ with } p(c_4|\mathcal{R}^{(8)}) = 1;\ [w^{(8)} = 1] \\
r^{(9)} &:= \text{if } x_1 \text{ is } A_1^{(9)} \text{ and } x_2 \text{ is } A_2^{(9)} \text{ and} \\
&\quad \text{then } \hat{y}^{(9)} = c_3 \text{ with } p(c_3|\mathcal{R}^{(9)}) = 1;\ [w^{(9)} = 0.5332]
\end{aligned}
$$

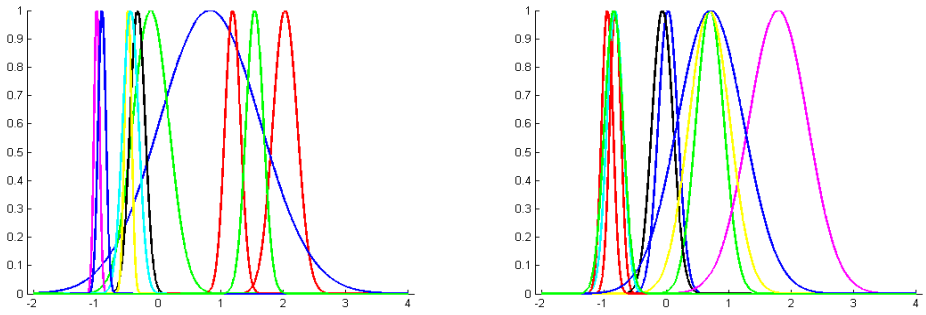
Table 4: Accuracies (mean ± standard deviation) obtained in 30 runs of 4-fold cross-validation for the probabilistic fuzzy classifiers as function of the number of rules (and clusters); the higher this number the higher the accuracy.

M	**pfc1**	**pfc2**	**pfc3**
2	36.84±10.36	50.00±0.00	53.16±3.01
3	74.81±15.05	75.00±0.00	75.01±0.05
5	98.75±3.22	98.89±0.43	98.93±0.38
10	99.03± 2.96	99.43± 0.23	99.57±0.17

$r^{(10)}$:= if x_1 is $A_1^{(10)}$ and x_2 is $A_2^{(10)}$ and

then $\hat{y}^{(10)} = c_2$ with $p(c_2|\mathcal{R}^{(10)}) = 1$; $[w^{(10)} = 0.098]$

A statistical analysis applied for each one of the three models with the same



(a) (b)

Figure 7: Typical membership functions for pfc3 trained with 10 clusters for the two selected features of the CWRU 6203 SKF data set: (a) time-domain rms and (b) linear amplitude rms of the FFT band 4. A high number of membership functions can improve accuracy but decreases interpretability.

number of rules reveals that the Analysis of Variance (ANOVA) applicability tests fail, i.e, both Bartlett's test for homoscedasticity and Shapiro-Wilk's

for residual normality yield $p_{\text{Bartlett}}, p_{\text{Shapiro-Wilk}} < 10^{-7}$. Consequently only the non-parametric Friedman test and post-hoc Tukey tests are used. These reveal that for $M = 2$, 3, and 5, pfc2 and pfc3 outperform pfc1; for $M = 10$ there is no statistically significant difference between the models.

When the issue is *only* accuracy, as in the case of non-fuzzy classifiers, it is only fair to compare these with highest accurate ($M = 10$) pfc. Fig 8 presents boxplots representing the accuracy obtained for the six considered models. Analysis show that there no statistical significant difference between pfc1, pfc2, pfc3, SVM, and 5-NN, and any of these outperforms PNN. Comparing

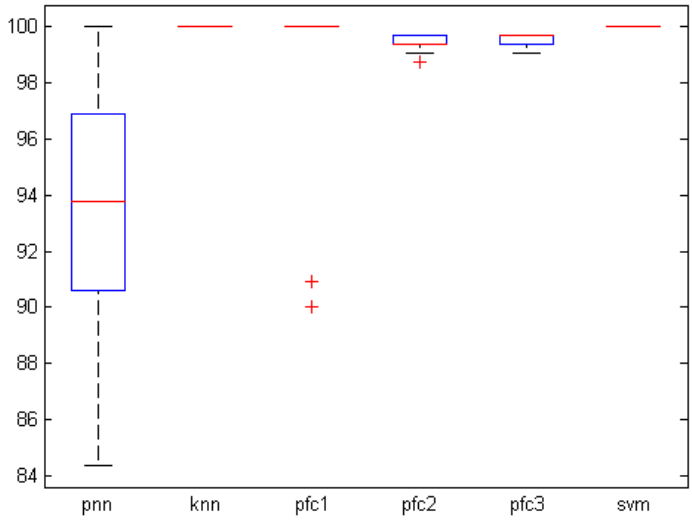


Figure 8: Boxplots representing accuracies obtained in 30 runs of 4-fold cross-validation for the different diagnosers. There is no statistically significant difference between pfc1, pfc2, pfc3, svm, and 5-NN, and any of these outperforms PNN.

the obtained results with those presented in (Dou and Zhou, 2016) for the same data – Acc of 0.9281 for k-NN ($K = 10$), 0.9438 for PNN, 0.9969 for

a PSO tuned SVM and 0.9906 for a five-rules based model (RBM) – we see that a similar level of performance (Acc) is achieved. Another immediate observation is that while our methodology was able to select only *two* relevant features, the cited study reports a selection of 11 features for the generality of the models, and 6 features for RBM. In general, for the same level of performance as is the case, a model with less features is desirable as it is more efficient, easier to understand by the user, and less prone to overfitting.

5.2 Results for the apparatus of Section 2.1

5.2.1 Features

When the proposed methodology is applied to the data acquired from the setup described in Section 2.1, 12 out of 1634 features are selected as the most relevant ones. Table 5 presents these features by descending order of their relevance. Both fault distribution among bearings (see Table 2) and fault interferences dictate that Accelerometer 1 is the most relevant being responsible for capturing 9 of the 12 selected features. Wavelets, i.e., time-frequency domain features correspond to 5 of the total number of selected features; In addition, five time domain and three frequency domain features have been selected. When we verify the distribution of the values of the above features, outliers are observed even for healthy bearings. This occurs mainly in time domain features – see Fig. 9 – but also occurs in time-frequency domain, e.g., this is the case of the feature corresponding to the energy of the leaf node 1 at the wavelet decomposition tree with Daubechies 7 wavelet (WPT (db7)–1). These outliers correspond to both mechanical and electric

Table 5: Selected features based on maximum information gain.

No.	Domain	ID	Acclr.	Obs
1	time	rms	1	
2	time	kda	1	
3	time	cf	1	
4	freq.	std of FFT band 3	1	
5	freq.	std of FFT band 9	1	
6	freq.	std of FFT band 30	1	
7	time-freq	WPT (db7)–1	1	Energy of node 1 in the db7 decomp.tree
8	time-freq	WPT (db7)–5	1	Energy of node 5
9	time-freq	WPT (sym3)–12	1	Energy of node 12
10	time	cf	2	
11	time-freq	WPT (db7)–5	2	Energy of node 5
12	time-freq	WPT (coif4)–15	2	Energy of node 15

noises that may degrade the performance of the adopted clustering method and therefore were removed using the Thompson Tau method (Dieck, 2006).

Given the number of selected features and recognizing that the employed entropy-based method can stop in local optima due to its greedy nature, it is desirable to check the possibility of further reducing the number of features using a method that involves the classifier itself. In this sense, we used the iterative process described in (Abonyi and Szeifert, 2003; Lee, Park, and Bien, 2008) where in each iteration the contribution of the current less promising feature is check against the performance of the classifier. A feature is selected for evaluation by computing an internal validity index (the Fischer index as described in the Appendix) of the obtained clustering results with and without the feature. The iterative process continues while there is no degradation on the performance of the classifier.

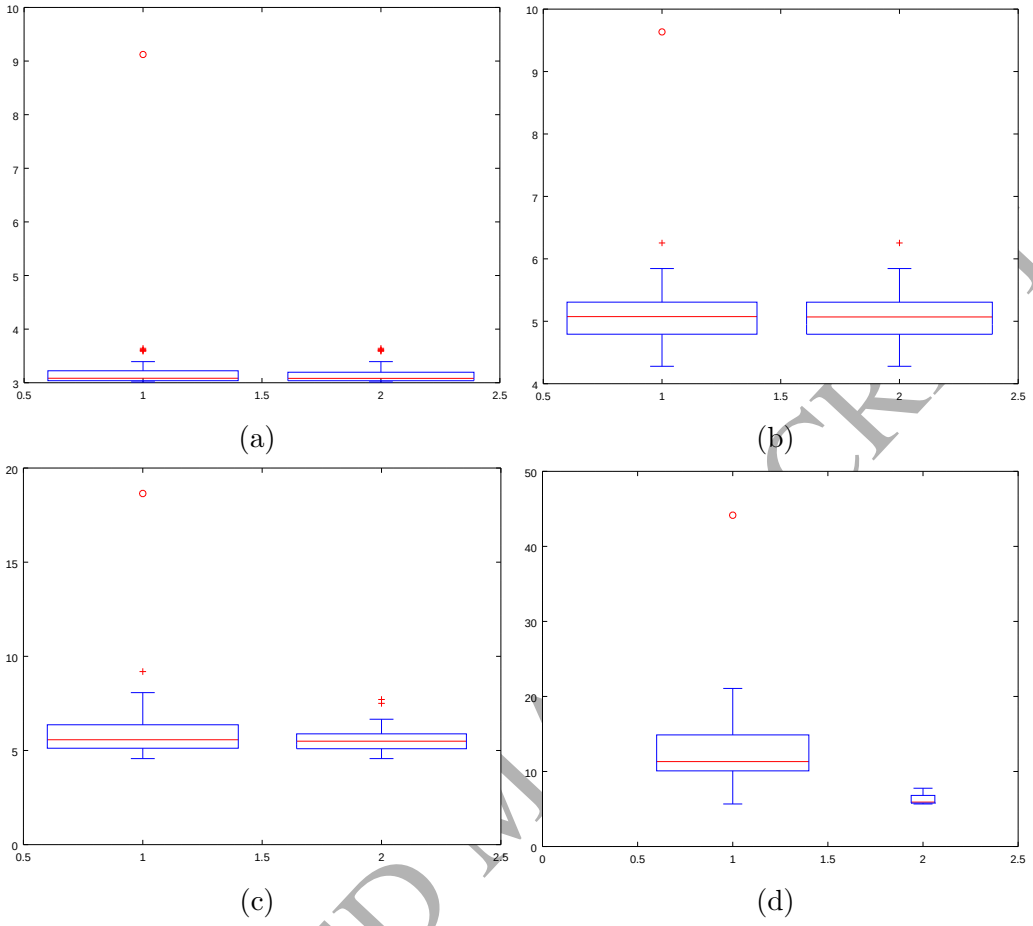
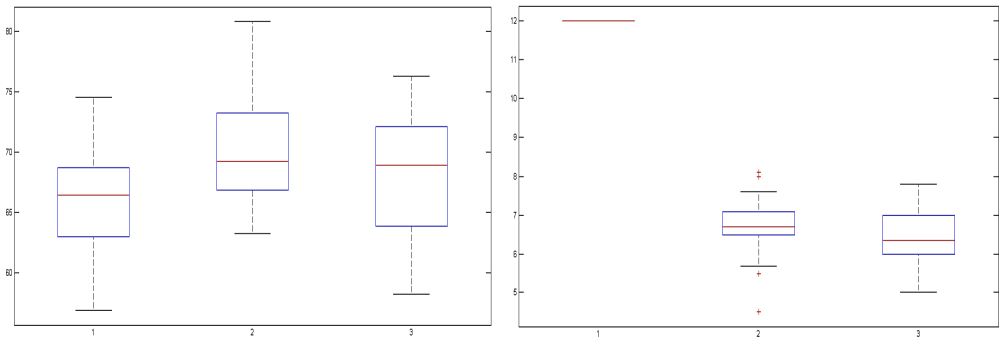


Figure 9: Boxplots exhibiting the distributions of selected features with (left) and without (right) outliers for healthy bearings: a) time domain operator kda (Acceler.1); b) time cf (Acceler.1) , c) time cf (Acceler.2), and d) WPT (db7)–1 (Acceler.1)

5.2.2 Diagnosis results

As before, pfc1 (baseline classifier trained using the method proposed in (Abonyi and Szeifert, 2003; Lee, Park, and Bien, 2008)), pfc2 (relevant region computed by (14)), and pfc3 (relevant region computed by (12)) are compared between each other for different number of rules, and compared with the KNN, PNN, and SVM.

The results involve $M = 2, 3, 5$, and 30 rules. For statistical significance, 30 runs of 10-fold cross-validation are performed for each case and each classifier. Fig 10 presents the accuracies (Acc) and the final number of features obtained by each one of the three classifiers when only two rules ($M = 2$) are used. When statistical tests are applied to these results, we see that there is a statistical significant difference in Acc, $p_{\text{ANOVA}} = 0.0042$. The conditions of applicability of the Analysis of Variance (ANOVA) are verified by the Bartlett's test for homoscedasticity ($p_{\text{Bartlett}} = 0.4726$) and Shapiro-Wilk's test for residual normality ($p_{\text{Shapiro-Wilk}} = 0.3378$). After applying post-hoc Tukey's tests we see that pfc2 outperforms pfc1, while there are no statistically significant differences among pfc1 and pfc3. In terms of the final number of features, no statistical difference between pfc2 and pfc3 is observed, and both classifiers require less features (as low as 5) than pfc1 (12 features).

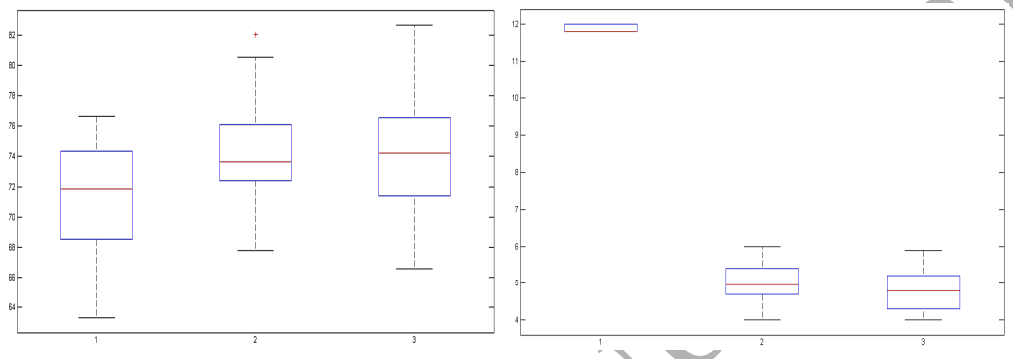


(a) (b)

Figure 10: Comparing classifiers with 2 rules: (a) Accuracy; (b) final number of features. Classifiers pfc1, pfc2, and pfc3 are identified in the x-label by 1, 2, and 3, respectively.

For three rules, there is a statistically significant difference for Acc between the three classifiers, i.e., $p_{\text{ANOVA}} = 0.0017$, $p_{\text{Bartlett}} = 0.6368$, and

$p_{\text{Shapiro-Wilk}} = 0.8535$. Subsequent Tukey's tests reveal that pfc2 and pfc3 have no statistically significant differences and outperform pfc1; see also Fig. 11. Relatively to the number of features, there is no difference among pfc2 and pfc3, and these require significantly less features than pfc1.

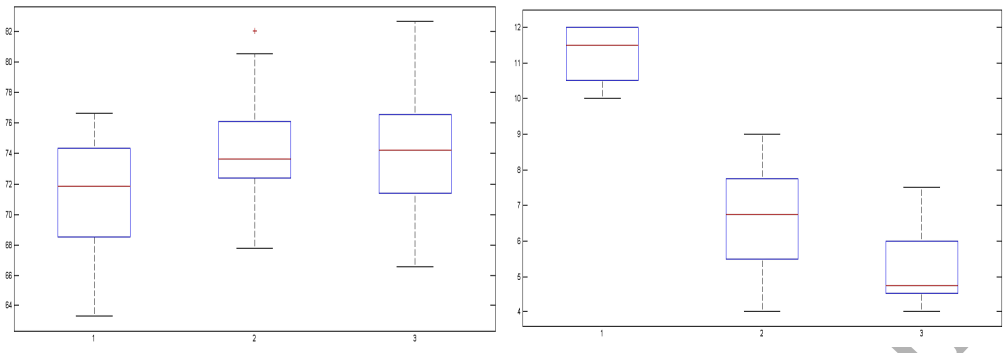


(a) (b)

Figure 11: Comparing classifiers with 3 rules: (a) Accuracy; (b) final number of features. Classifiers pfc1, pfc2, and pfc3 are identified in the x-label by 1, 2, and 3, respectively.

Fig. 12 shows the obtained results for 5 rules. ANOVA could not be used in the analysis of the Acc as $p_{\text{Bartlett}} = 0.0368$, i.e., one needs to reject the hypothesis of homoscedasticity. The non-parametric Friedman's test revealed that no statistically significant difference exists in the accuracies of the classifiers ($p_{\text{Friedman}} = 0.4966$). Again, no difference exists among pfc2 and pfc3 but these require significantly less features then pfc1.

Figs. 13, 14, and 15 show typical membership functions for some of the selected features as obtained by the supervised simplified Gath-Geva clustering algorithm (Algorithm 2 in the Appendix) for $M = 2, 3$, and 5, respectively. All the classifiers use this algorithm for antecedent parameter estimation. Similar membership functions are obtained for the other features,



(a) (b)

Figure 12: Comparing classifiers with 5 clusters: (a) Accuracy; (b) final number of features. Classifiers pfc1, pfc2, and pfc3 are identified in the x-label by 1, 2, and 3, respectively.

and are omitted for brevity.

The following is a typical rule set, one out of 300, generated obtained by pfs2 for $M = 2$ rules:

$$
\begin{aligned}
r^{(1)} \quad := \quad & \text{if } x_1 \text{ is } A_1^{(1)} \text{ and } x_2 \text{ is } A_2^{(1)} \text{ and } x_3 \text{ is } A_3^{(1)} \text{ and } x_4 \text{ is } A_4^{(1)} \text{ and } x_5 \text{ is } A_5^{(1)} \text{ and } x_6 \text{ is } A_6^{(1)} \\
& \text{then } \hat{y}^{(1)} = c_2 \text{ with } p(c_2|\mathcal{R}^{(2)}) = 0.3746; \hat{y}^{(1)} = c_3 \text{ with } p(c_3|\mathcal{R}^{(1)}) = 0.049; \\
& \hat{y}^{(1)} = c_4 \text{ with } p(c_4|\mathcal{R}^{(1)}) = 0.0576; \hat{y}^{(1)} = c_5 \text{ with } p(c_5|\mathcal{R}^{(1)}) = 0.0864; \\
& \hat{y}^{(1)} = c_6 \text{ with } p(c_6|\mathcal{R}^{(1)}) = 0.2881; \hat{y}^{(1)} = c_7 \text{ with } p(c_7|\mathcal{R}^{(1)}) = 0.1441; \ [w^{(1)} = 1] \\
r^{(2)} \quad := \quad & \text{if } x_1 \text{ is } A_1^{(2)} \text{ and } x_2 \text{ is } A_2^{(2)} \text{ and } x_3 \text{ is } A_3^{(2)} \text{ and } x_4 \text{ is } A_4^{(2)} \text{ and } x_5 \text{ is } A_5^{(2)} \text{ and } x_6 \text{ is } A_6^{(2)} \\
& \text{then } \hat{y}^{(2)} = c_1 \text{ with } p(c_1|\mathcal{R}^{(2)}) = 0.53; \hat{y}^{(2)} = c_3 \text{ with } p(c_3|\mathcal{R}^{(2)}) = 0.4543; \\
& \hat{y}^{(2)} = c_4 \text{ with } p(c_4|\mathcal{R}^{(2)}) = 0.0136; \ [w^{(2)} = 0.2487]
\end{aligned}
$$

As another example of rule set, we present again for the same type of classifier (pfc2) the following $M = 3$ rules:

$$
\begin{aligned}
r^{(1)} \quad := \quad & \text{if } x_1 \text{ is } A_1^{(1)} \text{ and } x_2 \text{ is } A_2^{(1)} \text{ and } x_3 \text{ is } A_3^{(1)} \text{ and } x_4 \text{ is } A_4^{(1)} \text{ and } x_5 \text{ is } A_5^{(1)} \text{ and } x_6 \text{ is } A_6^{(1)} \\
& \text{then } \hat{y}^{(1)} = c_2 \text{ with } p(c_2|\mathcal{R}^{(1)}) = 0.3868; \hat{y}^{(1)} = c_3 \text{ with } p(c_3|\mathcal{R}^{(1)}) = 0.0477; \\
& \hat{y}^{(1)} = c_4 \text{ with } p(c_4|\mathcal{R}^{(1)}) = 0.03; \hat{y}^{(1)} = c_5 \text{ with } p(c_5|\mathcal{R}^{(1)}) = 0.0893;
\end{aligned}
$$

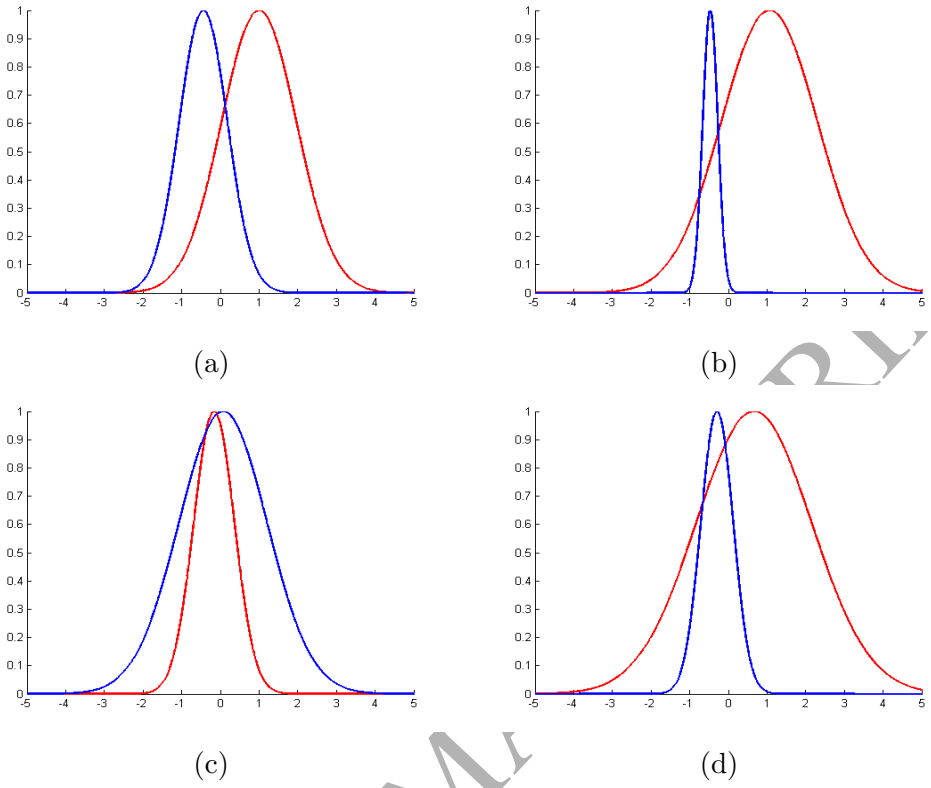


Figure 13: Typical membership functions for classifiers trained with 2 clusters for feature (a) 1; (b) 2; (c) 8, and (d) 12 in Table 5. Similar membership functions are obtained for the other features.

$$\begin{aligned}
& \hat{y}^{(1)} = c_6 \text{ with } p(c_6|\mathcal{R}^{(1)}) = 0.2975; \hat{y}^{(1)} = c_7 \text{ with } p(c_7|\mathcal{R}^{(1)}) = 0.1488;\ [w^{(1)} = 0.004] \\
r^{(2)} \quad := \quad & \text{if } x_1 \text{ is } A_1^{(2)} \text{ and } x_2 \text{ is } A_2^{(2)} \text{ and } x_3 \text{ is } A_3^{(2)} \text{ and } x_4 \text{ is } A_4^{(2)} \text{ and } x_5 \text{ is } A_5^{(2)} \text{ and } x_6 \text{ is } A_6^{(2)} \\
& \text{then } \hat{y}^{(2)} = c_3 \text{ with } p(c_3|\mathcal{R}^{(2)}) = 0.4543; \hat{y}^{(2)} = c_4 \text{ with } p(c_4|\mathcal{R}^{(2)}) = 0.0621;\ [w^{(2)} = 1] \\
r^{(3)} \quad := \quad & \text{if } x_1 \text{ is } A_1^{(3)} \text{ and } x_2 \text{ is } A_2^{(3)} \text{ and } x_3 \text{ is } A_3^{(3)} \text{ and } x_4 \text{ is } A_4^{(3)} \text{ and } x_5 \text{ is } A_5^{(3)} \text{ and } x_6 \text{ is } A_6^{(3)} \\
& \text{then } \hat{y}^{(3)} = c_1 \text{ with } p(c_1|\mathcal{R}^{(3)}) = 0.668; \hat{y}^{(3)} = c_3 \text{ with } p(c_3|\mathcal{R}^{(3)}) = 0.3149; \\
& \hat{y}^{(3)} = c_4 \text{ with } p(c_4|\mathcal{R}^{(3)}) = 0.0171;\ [w^{(3)} = 0.0267]
\end{aligned}$$

Example of a 5-rules set for pfc2:

$$\begin{aligned}
r^{(1)} \quad := \quad & \text{if } x_1 \text{ is } A_1^{(1)} \text{ and } x_2 \text{ is } A_2^{(1)} \text{ and } x_3 \text{ is } A_3^{(1)} \text{ and } x_4 \text{ is } A_4^{(1)} \text{ and } x_5 \text{ is } A_5^{(1)} \text{ and } x_6 \text{ is } A_6^{(1)} \\
& \text{then } \hat{y}^{(1)} = c_2 \text{ with } p(c_2|\mathcal{R}^{(1)}) = 0.6429; \hat{y}^{(1)} = c_7 \text{ with } p(c_7|\mathcal{R}^{(1)}) = 0.03171;\ [w^{(1)} = 0.03]
\end{aligned}$$

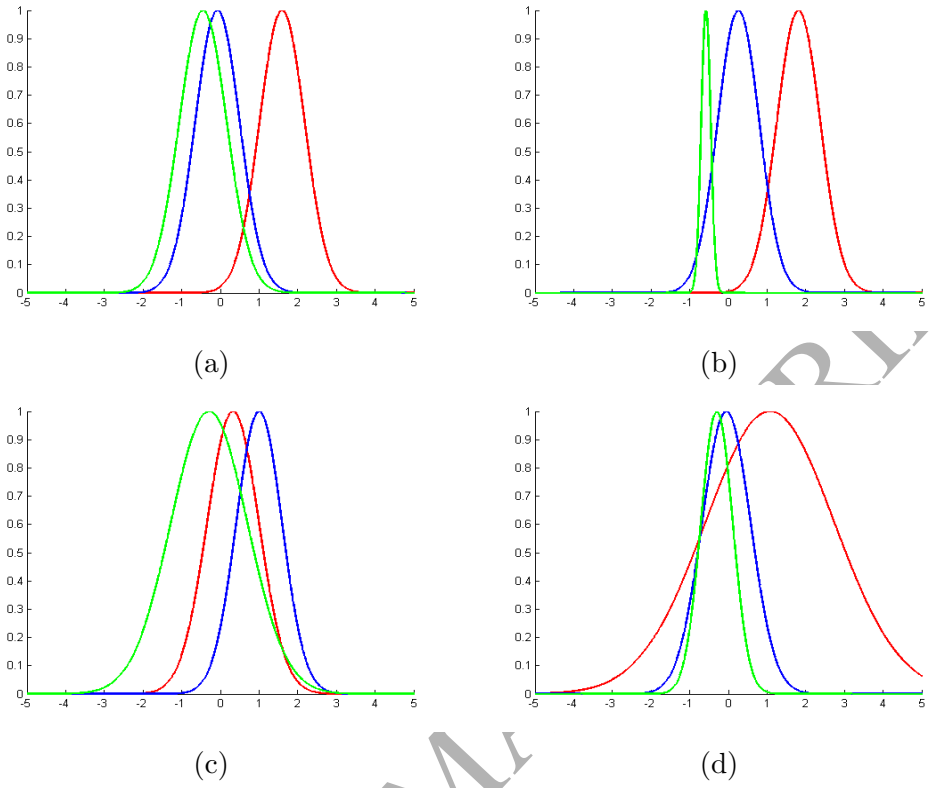


Figure 14: Typical membership functions for classifiers trained with 3 clusters for feature (a) 1; (b) 4; (c) 7, and (d) 12 in Table 5. Similar membership functions are obtained for the other features.

$r^{(2)}$:= if x_1 is $A_1^{(2)}$ and x_2 is $A_2^{(2)}$ and x_3 is $A_3^{(2)}$ and x_4 is $A_4^{(2)}$ and x_5 is $A_5^{(2)}$ and x_6 is $A_6^{(2)}$
then $\hat{y}^{(2)} = c_2$ with $p(c_2|\mathcal{R}^{(2)}) = 0.2857$; $\hat{y}^{(2)} = c_6$ with $p(c_6|\mathcal{R}^{(2)}) = 0.7143$; $[w^{(2)} = 1]$

$r^{(3)}$:= if x_1 is $A_1^{(3)}$ and x_2 is $A_2^{(3)}$ and x_3 is $A_3^{(3)}$ and x_4 is $A_4^{(3)}$ and x_5 is $A_5^{(3)}$ and x_6 is $A_6^{(3)}$
then $\hat{y}^{(3)} = c_4$ with $p(c_4|\mathcal{R}^{(3)}) = 0.4$; $\hat{y}^{(3)} = c_5$ with $p(c_5|\mathcal{R}^{(3)}) = 0.6$;
$\hat{y}^{(3)} = c_4$ with $p(c_4|\mathcal{R}^{(3)}) = 0.0171$; $[w^{(3)} = 0.5239]$

$r^{(4)}$:= if x_1 is $A_1^{(4)}$ and x_2 is $A_2^{(4)}$ and x_3 is $A_3^{(4)}$ and x_4 is $A_4^{(4)}$ and x_5 is $A_5^{(4)}$ and x_6 is $A_6^{(4)}$
then $\hat{y}^{(4)} = c_1$ with $p(c_1|\mathcal{R}^{(4)}) = 0.8662$; $\hat{y}^{(4)} = c_3$ with $p(c_3|\mathcal{R}^{(4)}) = 0.46$;
$\hat{y}^{(4)} = c_4$ with $p(c_4|\mathcal{R}^{(4)}) = 0.022$; $[w^{(3)} = 0.0154]$

$r^{(5)}$:= if x_1 is $A_1^{(5)}$ and x_2 is $A_2^{(5)}$ and x_3 is $A_3^{(5)}$ and x_4 is $A_4^{(5)}$ and x_5 is $A_5^{(5)}$ and x_6 is $A_6^{(5)}$
then $\hat{y}^{(5)} = c_3$ with $p(c_3|\mathcal{R}^{(5)}) = 1$; $[w^{(1)} = 1]$

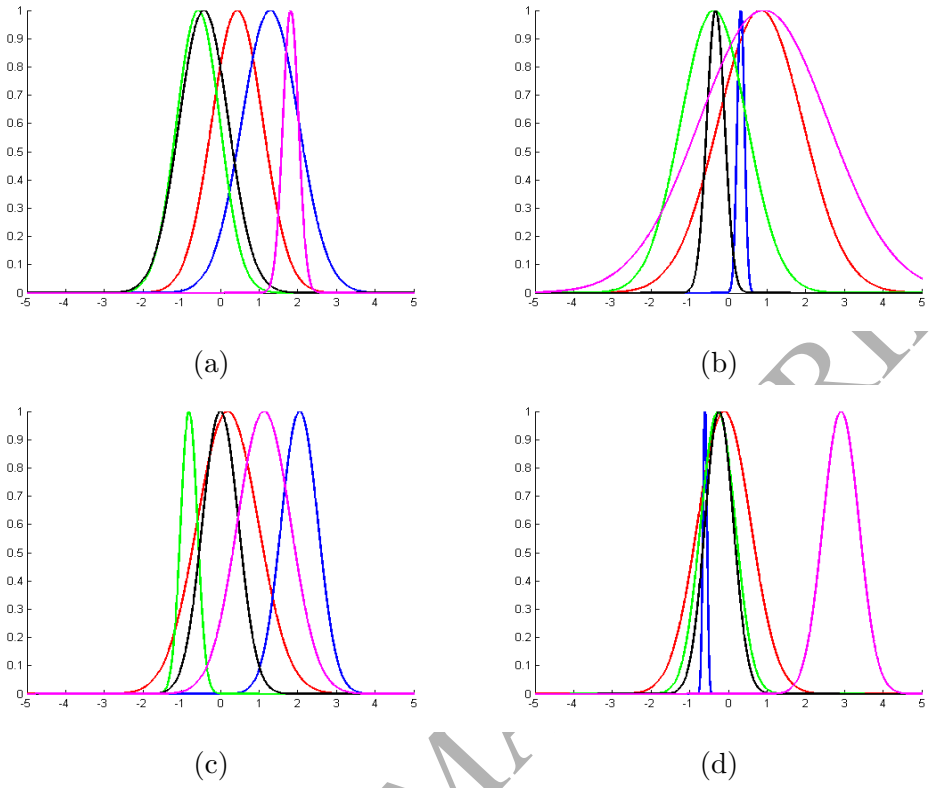


Figure 15: Typical membership functions for classifiers trained with 5 clusters for feature (a) 1; (b) 2; (c) 6, and (d) 12 in Table 5. Similar membership functions are obtained for the other features.

Again it is observed that a trade-off between Acc and interpretability exists for the above models. Again, considering Acc *only* it is only fair to compare high accurate pfc with other classifiers. Fig 16 presents box-plots representing the accuracy obtained for 30-rules pfc1, pfc2 ($\alpha = 0.2$), pfc3 ($p = 2$), KNN ($K = 5$), PNN ($\sigma^2 = 1$), and SVM. A Friedman test ($p_{\text{Friedman}} = 0.001$) followed by a post-hoc allows us to conclude that there is no statistically significant difference between pfc2, pfc3, KNN, PNN, and SVM. Moreover, any of these outperforms pfc1.

Again, for 30 rules, both pfc2 and pfc3 were able to obtained an sta-

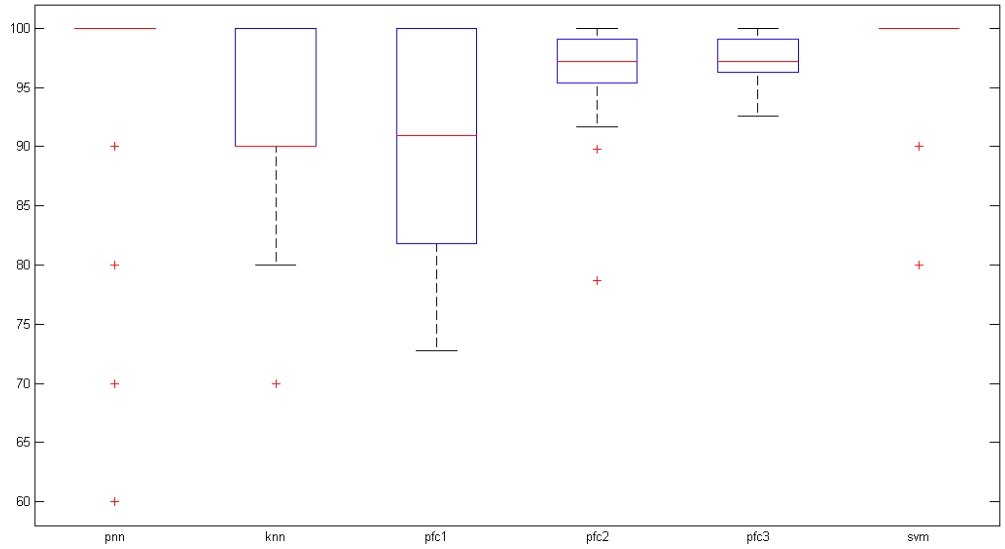


Figure 16: Boxplots representing accuracies obtained in 30 runs of 10-fold cross-validation for the different diagnosers. There is no statistically significant difference between pfc2, pfc3, KNN, PNN, and SVM, and any of these outperforms pfc1.

tistically reduction on the number of features relatively to pfc1 ($p_{\text{Friedman}} = 5.2316 \times 10^{-12}$). The former classifiers required 4 features only; See Fig. 17.

As an example, it is presented bellow a typical 4 features 30 rules set for pfc3. Similar rule sets were found also for pfc2.

$$
\begin{aligned}
r^{(1)} \quad &:= \quad \text{if } x_6 \text{ is } A_6^{(1)} \text{ and } x_9 \text{ is } A_9^{(1)} \text{ and } x_{10} \text{ is } A_{10}^{(1)} \text{ and } x_{11} \text{ is } A_{11}^{(1)} \text{ and} \\
&\quad\ \ \text{then } \hat{y}^{(1)} = c_4 \text{ with } p(c_4|\mathcal{R}^{(1)}) = 0.5; \hat{y}^{(1)} = c_6 \text{ with } p(c_6|\mathcal{R}^{(1)}) = 0.5; \ [w^{(1)} = 0.0195] \\
r^{(2)} \quad &:= \quad \text{if } x_6 \text{ is } A_6^{(2)} \text{ and } x_9 \text{ is } A_9^{(2)} \text{ and } x_{10} \text{ is } A_{10}^{(2)} \text{ and } x_{11} \text{ is } A_{11}^{(2)} \text{ and} \\
&\quad\ \ \text{then } \hat{y}^{(2)} = c_3 \text{ with } p(c_3|\mathcal{R}^{(2)}) = 1; \ [w^{(3)} = 0.001] \\
r^{(3)} \quad &:= \quad \text{if } x_6 \text{ is } A_6^{(3)} \text{ and } x_9 \text{ is } A_9^{(3)} \text{ and } x_{10} \text{ is } A_{10}^{(3)} \text{ and } x_{11} \text{ is } A_{11}^{(3)} \text{ and} \\
&\quad\ \ \text{then } \hat{y}^{(3)} = c_2 \text{ with } p(c_2|\mathcal{R}^{(3)}) = 1; \ [w^{(1)} = 0.1184] \\
r^{(4)} \quad &:= \quad \text{if } x_6 \text{ is } A_6^{(4)} \text{ and } x_9 \text{ is } A_9^{(4)} \text{ and } x_{10} \text{ is } A_{10}^{(4)} \text{ and } x_{11} \text{ is } A_{11}^{(4)} \text{ and} \\
&\quad\ \ \text{then } \hat{y}^{(4)} = c_3 \text{ with } p(c_3|\mathcal{R}^{(4)}) = 1; \ [w^{(4)} = 1] \\
r^{(5)} \quad &:= \quad \text{if } x_6 \text{ is } A_6^{(5)} \text{ and } x_9 \text{ is } A_9^{(5)} \text{ and } x_{10} \text{ is } A_{10}^{(5)} \text{ and } x_{11} \text{ is } A_{11}^{(5)} \text{ and}
\end{aligned}
$$

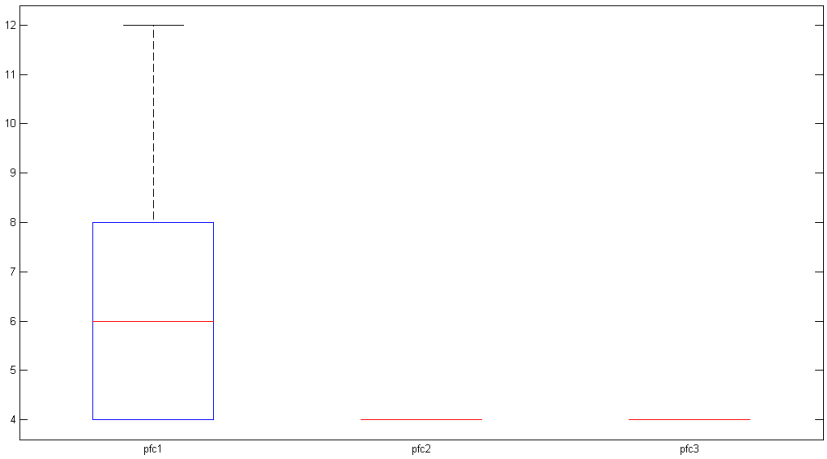


Figure 17: Boxplots representing the obtained distribution of the final number of features for 30-rules pfc1, pfc2, and pfc3. Both pfc2 and pfc3 required only 4 features.

$$
\begin{aligned}
& \text{then } \hat{y}^{(5)} = c_6 \text{ with } p(c_6|\mathcal{R}^{(5)}) = 1;\ [w^{(5)} = 0.0018] \\
r^{(6)} \quad := \quad & \text{if } x_6 \text{ is } A_6^{(6)} \text{ and } x_9 \text{ is } A_9^{(6)} \text{ and } x_{10} \text{ is } A_{10}^{(6)} \text{ and } x_{11} \text{ is } A_{11}^{(6)} \text{ and} \\
& \text{then } \hat{y}^{(6)} = c_1 \text{ with } p(c_1|\mathcal{R}^{(6)}) = 1;\ [w^{(6)} = 0.001] \\
r^{(7)} \quad := \quad & \text{if } x_6 \text{ is } A_6^{(7)} \text{ and } x_9 \text{ is } A_9^{(7)} \text{ and } x_{10} \text{ is } A_{10}^{(7)} \text{ and } x_{11} \text{ is } A_{11}^{(7)} \text{ and} \\
& \text{then } \hat{y}^{(7)} = c_1 \text{ with } p(c_1|\mathcal{R}^{(7)}) = 1;\ [w^{(7)} = 0.001] \\
r^{(8)} \quad := \quad & \text{if } x_6 \text{ is } A_6^{(8)} \text{ and } x_9 \text{ is } A_9^{(8)} \text{ and } x_{10} \text{ is } A_{10}^{(8)} \text{ and } x_{11} \text{ is } A_{11}^{(8)} \text{ and} \\
& \text{then } \hat{y}^{(8)} = c_3 \text{ with } p(c_3|\mathcal{R}^{(8)}) = 1;\ [w^{(8)} = 0.001] \\
r^{(9)} \quad := \quad & \text{if } x_6 \text{ is } A_6^{(9)} \text{ and } x_9 \text{ is } A_9^{(9)} \text{ and } x_{10} \text{ is } A_{10}^{(9)} \text{ and } x_{11} \text{ is } A_{11}^{(9)} \text{ and} \\
& \text{then } \hat{y}^{(9)} = c_7 \text{ with } p(c_7|\mathcal{R}^{(9)}) = 1;\ [w^{(9)} = 0.4219] \\
r^{(10)} \quad := \quad & \text{if } x_6 \text{ is } A_6^{(10)} \text{ and } x_9 \text{ is } A_9^{(10)} \text{ and } x_{10} \text{ is } A_{10}^{(10)} \text{ and } x_{11} \text{ is } A_{11}^{(10)} \text{ and} \\
& \text{then } \hat{y}^{(10)} = c_1 \text{ with } p(c_1|\mathcal{R}^{(10)}) = 1;\ [w^{(10)} = 0.001] \\
r^{(11)} \quad := \quad & \text{if } x_6 \text{ is } A_6^{(11)} \text{ and } x_9 \text{ is } A_9^{(11)} \text{ and } x_{10} \text{ is } A_{10}^{(11)} \text{ and } x_{11} \text{ is } A_{11}^{(11)} \text{ and} \\
& \text{then } \hat{y}^{(11)} = c_1 \text{ with } p(c_1|\mathcal{R}^{(11)}) = 1;\ [w^{(11)} = 0.99] \\
r^{(12)} \quad := \quad & \text{if } x_6 \text{ is } A_6^{(12)} \text{ and } x_9 \text{ is } A_9^{(12)} \text{ and } x_{10} \text{ is } A_{10}^{(12)} \text{ and } x_{11} \text{ is } A_{11}^{(12)} \text{ and} \\
& \text{then } \hat{y}^{(12)} = c_1 \text{ with } p(c_1|\mathcal{R}^{(12)}) = 1;\ [w^{(12)} = 0.001] \\
r^{(13)} \quad := \quad & \text{if } x_6 \text{ is } A_6^{(13)} \text{ and } x_9 \text{ is } A_9^{(13)} \text{ and } x_{10} \text{ is } A_{10}^{(13)} \text{ and } x_{11} \text{ is } A_{11}^{(13)} \text{ and} \\
& \text{then } \hat{y}^{(13)} = c_3 \text{ with } p(c_3|\mathcal{R}^{(13)}) = 1;\ [w^{(13)} = 0.001] \\
r^{(14)} \quad := \quad & \text{if } x_6 \text{ is } A_6^{(14)} \text{ and } x_9 \text{ is } A_9^{(14)} \text{ and } x_{10} \text{ is } A_{10}^{(14)} \text{ and } x_{11} \text{ is } A_{11}^{(14)} \text{ and} \\
& \text{then } \hat{y}^{(14)} = c_3 \text{ with } p(c_3|\mathcal{R}^{(14)}) = 1;\ [w^{(14)} = 0.2051] \\
r^{(15)} \quad := \quad & \text{if } x_6 \text{ is } A_6^{(15)} \text{ and } x_9 \text{ is } A_9^{(15)} \text{ and } x_{10} \text{ is } A_{10}^{(15)} \text{ and } x_{11} \text{ is } A_{11}^{(15)} \text{ and} \\
& \text{then } \hat{y}^{(15)} = c_4 \text{ with } p(c_4|\mathcal{R}^{(15)}) = 1;\ [w^{(15)} = 0.2051]
\end{aligned}
$$

$r^{(16)}$:= if x_6 is $A_6^{(16)}$ and x_9 is $A_9^{(16)}$ and x_{10} is $A_{10}^{(16)}$ and x_{11} is $A_{11}^{(16)}$ and then $\hat{y}^{(16)} = c_1$ with $p(c_1|\mathcal{R}^{(16)}) = 1$; $[w^{(16)} = 0.3828]$

$r^{(17)}$:= if x_6 is $A_6^{(17)}$ and x_9 is $A_9^{(17)}$ and x_{10} is $A_{10}^{(17)}$ and x_{11} is $A_{11}^{(17)}$ and then $\hat{y}^{(17)} = c_4$ with $p(c_4|\mathcal{R}^{(17)}) = 1$; $[w^{(17)} = 0.3828]$

$r^{(18)}$:= if x_6 is $A_6^{(18)}$ and x_9 is $A_9^{(18)}$ and x_{10} is $A_{10}^{(18)}$ and x_{11} is $A_{11}^{(18)}$ and then $\hat{y}^{(18)} = c_2$ with $p(c_2|\mathcal{R}^{(18)}) = 1$; $[w^{(18)} = 0.2122]$

$r^{(19)}$:= if x_6 is $A_6^{(19)}$ and x_9 is $A_9^{(19)}$ and x_{10} is $A_{10}^{(19)}$ and x_{11} is $A_{11}^{(19)}$ and then $\hat{y}^{(19)} = c_3$ with $p(c_3|\mathcal{R}^{(19)}) = 1$; $[w^{(19)} = 0.0238]$

$r^{(20)}$:= if x_6 is $A_6^{(20)}$ and x_9 is $A_9^{(20)}$ and x_{10} is $A_{10}^{(20)}$ and x_{11} is $A_{11}^{(20)}$ and then $\hat{y}^{(20)} = c_4$ with $p(c_4|\mathcal{R}^{(20)}) = 0.25; \hat{y}^{(1)} = c_6$ with $p(c_6|\mathcal{R}^{(20)}) = 0.75$; $[w^{(20)} = 0.0163]$

$r^{(21)}$:= if x_6 is $A_6^{(21)}$ and x_9 is $A_9^{(21)}$ and x_{10} is $A_{10}^{(21)}$ and x_{11} is $A_{11}^{(21)}$ and then $\hat{y}^{(21)} = c_1$ with $p(c_1|\mathcal{R}^{(21)}) = 1$; $[w^{(21)} = 0.3147]$

$r^{(22)}$:= if x_6 is $A_6^{(22)}$ and x_9 is $A_9^{(22)}$ and x_{10} is $A_{10}^{(22)}$ and x_{11} is $A_{11}^{(22)}$ and then $\hat{y}^{(3)} = c_2$ with $p(c_2|\mathcal{R}^{(22)}) = 1$; $[w^{(22)} = 0.6641]$

$r^{(23)}$:= if x_6 is $A_6^{(23)}$ and x_9 is $A_9^{(23)}$ and x_{10} is $A_{10}^{(23)}$ and x_{11} is $A_{11}^{(23)}$ and then $\hat{y}^{(23)} = c_3$ with $p(c_3|\mathcal{R}^{(23)}) = 1$; $[w^{(23)} = 0.0385]$

$r^{(24)}$:= if x_6 is $A_6^{(24)}$ and x_9 is $A_9^{(24)}$ and x_{10} is $A_{10}^{(24)}$ and x_{11} is $A_{11}^{(24)}$ and then $\hat{y}^{(24)} = c_1$ with $p(c_1|\mathcal{R}^{(24)}) = 1$; $[w^{(24)} = 0.0018]$

$r^{(25)}$:= if x_6 is $A_6^{(25)}$ and x_9 is $A_9^{(25)}$ and x_{10} is $A_{10}^{(25)}$ and x_{11} is $A_{11}^{(25)}$ and then $\hat{y}^{(25)} = c_1$ with $p(c_1|\mathcal{R}^{(25)}) = 1$; $[w^{(25)} = 0.4767]$

$r^{(26)}$:= if x_6 is $A_6^{(26)}$ and x_9 is $A_9^{(26)}$ and x_{10} is $A_{10}^{(26)}$ and x_{11} is $A_{11}^{(26)}$ and then $\hat{y}^{(26)} = c_2$ with $p(c_2|\mathcal{R}^{(26)}) = 1$; $[w^{(26)} = 0.1143]$

$r^{(27)}$:= if x_6 is $A_6^{(27)}$ and x_9 is $A_9^{(27)}$ and x_{10} is $A_{10}^{(27)}$ and x_{11} is $A_{11}^{(27)}$ and then $\hat{y}^{(27)} = c_2$ with $p(c_2|\mathcal{R}^{(27)}) = 1$; $[w^{(27)} = 0.0028]$

$r^{(28)}$:= if x_6 is $A_6^{(28)}$ and x_9 is $A_9^{(28)}$ and x_{10} is $A_{10}^{(28)}$ and x_{11} is $A_{11}^{(28)}$ and then $\hat{y}^{(28)} = c_5$ with $p(c_5|\mathcal{R}^{(28)}) = 1$; $[w^{(28)} = 0.3213]$

$r^{(29)}$:= if x_6 is $A_6^{(29)}$ and x_9 is $A_9^{(29)}$ and x_{10} is $A_{10}^{(29)}$ and x_{11} is $A_{11}^{(29)}$ and then $\hat{y}^{(29)} = c_3$ with $p(c_3|\mathcal{R}^{(29)}) = 1$; $[w^{(29)} = 0.0001]$

$r^{(30)}$:= if x_6 is $A_6^{(30)}$ and x_9 is $A_9^{(30)}$ and x_{10} is $A_{10}^{(30)}$ and x_{11} is $A_{11}^{(30)}$ and then $\hat{y}^{(30)} = c_3$ with $p(c_3|\mathcal{R}^{(30)}) = 1$; $[w^{(30)} = 0.001]$

These rule sets are particularly insightful as they show that to increase accuracy more classes c_j with $p(c_j|\mathcal{R}^{(i)}) = 1$ occurs. Moreover, typically about 1/3 of the rules exhibit $w^{(i)} \leq 0.001$. One might be tempted to discard

such rules. When we do such exercise in this particular case accuracy fell from 0.975 to 0.935. More systematically, Wilcox tests comparing the Acc with and without $w^{(i)} \leq 0.001$ rules shows $p_{\text{Wilcox}} = 6.5907 \times 10^{-4}$ for pfc2 and $p_{\text{Wilcox}} = 9.619 \times 10^{-4}$ for pfc3 revealing that a statistically significant Acc reduction occurs and therefore these rules are also needed for improving accuracy.

5.3 Discussion

For the considered application, the proposed method of parameter estimation outperforms the parameter estimation state-of-the-art approach in terms of accuracy, being indistinguishable in some rare cases.

The definition of a relevant region allows the proposed method to significantly reduce the number of features, what is an asset in many aspects, specially in terms of efficiency and interpretability of the classifiers.

The employed feature selection methodology contributes to build a parsimonious model as it allows the initial selection of only 2 relevant features for the CWRU data set and 12 relevant features for the proposed setup, from more than 1600. For the proposed experimental setup, and when combined with the proposed parameter estimation algorithm, a 6-features 5-rules classifier achieved averaged cross-validation accuracy rates above 80%; while a 4 features 30-rules pfc2 or pfc3 are able to obtain average accuracy of 97.5%.

Allowing the reduction of the number of features is an interesting side-effect of the proposed parameter estimation method. We are estimating the consequents of each rule using a subset of the region of influence of that rule.

This subset comprises only data with a high match (membership value) with the antecedent of the rule. This is in contrast with the common approach where the consequents were estimated using *all* data within the region of the rule no matter how low their membership values. One hypothesis for the relationship between this and the reduction of the number of features is the following: By selecting only high match data we are discarding data that otherwise would require redundant features.

Often feature selection is performed by a genetic algorithm in an outer loop of the parameter estimation process. Usually, the latter approach yields a number of relevant features as higher as several hundreds, which would make the employment of fuzzy models useless, as the interpretability of a model with such number of antecedents would be very difficult.

The proposed consequent parameter estimation method is simple to implement, practically has no overhead on the complexity of the overall training of the classifier. In general, a fuzzy definition of the relevant region yields the best results, i.e., in general pfs2 performs better or equal to pfs3. In the proposed definitions, a single parameter needs to be adjusted for different problems. It is worth stressing that to obtained these performances no fine parameter tuning was required in the definition of the relevant region. For example, pfc3 defines this region using the operator intensification of contrast with $p = 2$; which is a quite common value in this operator. Consequently, no claim is made on the optimality of the parameters used for defining the relevant region. How to optimally estimate the hyper-parameters of the relevant region (α in (12) or p in (14)) without resorting to a trial and error approach is currently an open issue that, given the empirical evidence gathered so far,

is well worth researching in the future.

The proposed estimation parameter procedure is much more efficient than the traditionally used methods for parameter estimation in fuzzy models based on gradient such in the case of (C)ANFIS, e.g., (Goode and Chow, 1994; Ertunc, Ocak, and Aliustaoglu, 2013; Zhang, Ma, and Ma, 2014).

The adopted model is much more compact than conventional Mamdani or Takagi-Sugeno-Khan models (e.g., (Lou and Loparo, 2004; Harrouche and Felkaoui, 2014)) as each rule of its rule base is able to classify different faults with different probabilities.

One distinctive characteristic of rule based models such as the proposed fuzzy probabilistic one, is that it is possible to find an equilibrium between the desirable level of detail (interpretability) and accuracy.

When accuracy is the sole comparison criterion, results have shown that the proposed model systematically matches the performance of other data-driven models like distance based methods (KNN), connectionists (PNN), and maximum margin classifiers (SVM).

Giving the no-free lunch theorems (Wolpert and Macready, 1997) no learner will outperform all the others over the set of all applications. Our classifier is not an exception. However, when compared to the conventional approach, the employment of a relevant region has shown experimentally better results over Poisson and Gaussian distributions in (Ledo, Lucas, and Delgado, 2014) and has produced statistically significant superior results here for both a physical benchmark setup and a more complex setup representative of an industrial setup for bearing fault diagnosis. It is likely that the proposed approach will prove to be effective also in other domains where the

feature space is similar to the studied ones in what concerns dimensionality and class overlapping due to interferences and noise.

6 Conclusions

Bearing fault diagnosis is both an economically relevant and a scientific and technologic challenging topic. This paper has presented a first-time application of fuzzy probabilistic classifiers to bearing fault diagnosis. These are rule-based systems where each rule can diagnose a set of faults each one of them with an associated probability. Each rule can be viewed as describing a fuzzy region in the feature space where the consequent probability distribution over predicted classes is valid.

For parameter estimation a two steps sequential state-of-the-art data-driven method was adopted. In the first step the antecedents of the rules were estimated using an iterative supervised clustering algorithm. Based on the antecedents the consequent parameters are then estimated. Another contribution of this work was the proposal of a new Bayesian parameter estimation method for the rule consequent parameters. In the proposed method, *a priori* information on the probability of each consequent is updated by a likelihood function that takes into account only the information within a relevant region of the feature space. In other words, only training data that reach high values of matching with a rule antecedent are considered as relevant *posteriori* information. Two different ways of defining the relevant region were proposed: one is based on the notion of α-cut and the other on the contrast intensification modifier.

The proposed method was applied to a simple but widely used benchmark dataset relative to the 6203-2RS JEM SKF deep groove ball bearing from the Case Western Reserve University (CWRU) Bearing Data Centre, and in a complex set of experiments exhibiting interferences among faults of different bearings and involving 1634 features and 7 output classes. Results have shown a statistically significant improvement on the performance of probabilistic fuzzy diagnosers trained with the proposed algorithm, independently of the criterion used for defining the relevant region, when compared with the original state-of-the-art method considered.

The proposed approach allows one to select the desirable trade-off between the model level of detail (interpretability) and accuracy. For example, for the proposed experimental setup, results have shown that the proposed approach is capable of achieving satisfactory average cross-validation accuracy rates (above 80%) with only five rules, six input variables. The employment of a greedy entropy based feature selection methodology (input variables) associated with an iterative feature selection method contributed to build such parsimonious fault diagnoser. When accuracy is the sole comparison criterion, results show that the proposed probabilistic fuzzy systems systematically matches the performance of other data-driven models like distance based methods (KNN), connectionists (PNN), and maximum margin classifiers (SVM).

Given the performance of the proposed model and identification procedure we can claim that this same framework would be worth exploiting in other fault diagnosis and classification problems. Moreover the recently proposed technique of observer-biased clustering (Fazendeiro and Valente de

Oliveira, 2015) is an asset as it can provide a set of reasonable clustering alternatives that allows one to build a model with different degrees of detail yielding comparable accuracy results.

Acknowledgements

The work was sponsored in part by the Prometeo Project of the Secretariat for Higher Education, Science, Technology and Innovation (SENESCYT) of the Republic of Ecuador, the National Key Research & Development Program of China (2016YFE0132200), the Chongqing Technology and Business University (CTBU) open grant number 1456027, by CNPq, Brazil, grant number 309197/2014-7, and by FCT, Portugal, grant number SFRH/BSAB/128153/2016. The experimental work was developed at the GIDTEC research group lab of UPS, Cuenca, Ecuador.

Appendix

Algorithm 2: Supervised simplified Gath-Geva clustering algorithm (Abonyi and Szeifert, 2003)

Input : Feature set: $\mathbf{X} \subset \mathbb{R}^d$, s.t., $\mathbf{X} = \{\vec{x}_1, \ldots, \vec{x}_N\}$; Label set, $\vec{y} = \{y_1, \ldots, y_k, \ldots, y_N\}$ where $y_k \in \{c_1, \ldots, c_C\}$; Number of clusters: M; fuzzification parameter: $m > 1$

Output: Prototypes: $\mathbf{V} \subset \mathbb{R}^d$, s.t., $\mathbf{V} = \{\vec{v}_1, \ldots, \vec{v}_M\}$; Partition matrix: $\mathbf{U} = [u_{ij}] \in \mathbb{R}^{M \times N}$; Diagonal elements of the covariance matrices $\mathbf{F}_i (i = 1, \ldots, M) : \sigma_{il}^2 (l = 1, \ldots, d)$.

$\mathbf{Z} = [\mathbf{X}\ \vec{y}]$;
Initialize $\mathbf{V}$;

repeat
  for *i=1* **to** M **do**
    for *j=1* **to** N **do**
      Compute the dissimilarity measure D_{ij}^2 using (9) ;
      Update the partition matrix:

$$u_{ij} = \frac{(1/D_{ij})^{2/(m-1)}}{\sum_{k=1}^{M} (1/D_{kj})^{2/(m-1)}} \tag{32}$$

    end
    for *l=1* **to** d **do**
      Compute the diagonal elements of $\mathbf{F}_i$:

$$\sigma_{il}^2 = \frac{\sum_{k=1}^{N} u_{ik}^m (x_{kl} - v_{il})^2}{\sum_{k=1}^{N} u_{ik}^m} \tag{33}$$

    end
    Update the prototypes:

$$\vec{v}_i = \frac{\sum_{k=1}^{N} u_{ik}^m \vec{x}_k}{\sum_{k=1}^{N} u_{ik}^m} \tag{34}$$

  end
until *a stop criterion is met*;

The Fischer index

The Fischer index can be viewed as an internal validity index to assess the quality of a data partition, i.e., the output of a clustering algorithm. This section briefly reviews the index as presented in (Abonyi and Szeifert, 2003). The Fischer index considers that the total covariance of data is $\mathbf{F}_T = \mathbf{F}_W + \mathbf{F}_B$ where the former is the within-class covariance matrix and the later is the between class covariance matrix which are given, respectively, by:

$$\begin{aligned}
\mathbf{F}_W &= \sum_{j=1}^{M} P(r^{(j)})\mathbf{F}_j \\
\mathbf{F}_B &= \sum_{j=1}^{M} P(r^{(j)})(\vec{v}_j - \vec{v}_0)^T(\vec{v}_j - \vec{v}_0) \\
\vec{v}_0 &= \sum_{j=1}^{M} P(r^{(j)})\vec{v}_j
\end{aligned} \tag{35}$$

The Fischer interclass separability index is then given by:

$$F = \frac{|\mathbf{F}_B|}{|\mathbf{F}_W|} \tag{36}$$

where $|.|$ is the determinant operator. As the higher the distance among groups and the lower the variability within each cluster the better, the higher F the better the partition.

References

Abed W., Sharma S., and Sutton R. (2014) "Diagnosis of bearing fault of brushless DC motor based on dynamic neural network and orthogonal fuzzy neighborhood discriminant analysis". *Control (CONTROL), 2014 UKACC International Conference on*, pp. 378–383.

Abonyi János and Szeifert Ferenc (2003) "Supervised fuzzy clustering for the identification of fuzzy classifiers". *Pattern Recognition Letters* 24.14, pp. 2195–2207.

Aggarwal M. (2016) "Probabilistic Variable Precision Fuzzy Rough Sets". *IEEE Transactions on Fuzzy Systems* 24.1, pp. 29–39.

Almeida R. J. and Kaymak U. (2009) "Probabilistic fuzzy systems in value-at-risk estimation". *Intelligent Systems in Accounting, Finance & Management* 16.1-2, pp. 49–70.

Amar M., Gondal I., and Wilson C. (2013) "Fuzzy logic inspired bearing fault-model membership estimation". *Proceedings of the 2013 IEEE 8th International Conference on Intelligent Sensors, Sensor Networks and Information Processing: Sensing the Future, ISSNIP 2013*. Vol. 1, pp. 420–425.

Bediaga I., Mendizabal X., Etxaniz I., and Munoa J. (2013) "An integrated system for machine tool spindle head ball bearing fault detection and diagnosis". *IEEE Instrumentation and Measurement Magazine* 16.2, pp. 42–47.

Ben Ali J., Saidi L., Mouelhi A., Chebel-Morello B., and Fnaiech F. (2015) "Linear feature selection and classification using PNN and SFAM neural networks for a nearly online diagnosis of bearing naturally progressing degradations". *Engineering Applications of Artificial Intelligence* 42, pp. 67–81.

Berg J. van den, Kaymak U., and Almeida R.J. (2013) "Conditional Density Estimation Using Probabilistic Fuzzy Systems". *Fuzzy Systems, IEEE Transactions on* 21.5, pp. 869–882.

Berg J. van den, Kaymak U., and Bergh W.-M. van den (2002) "Fuzzy classification using probability-based rule weighting". *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2002)*, pp. 991–996.

Berg Jan van den, Kaymak Uzay, and Bergh Willem-Max van den (2004) "Financial markets analysis by using a probabilistic fuzzy modelling approach". *International Journal of Approximate Reasoning* 35.3, pp. 291 –305.

Boutros Tony and Liang Ming (2007) "Mechanical fault detection using fuzzy index fusion". *International Journal of Machine Tools and Manufacture* 47.11, pp. 1702 –1714.

Dieck R.H. (2006) *Measurement Uncertainty: Methods and Applications*. ISA.

Dou Dongyang and Zhou Shishuai (2016) "Comparison of four direct classification methods for intelligent fault diagnosis of rotating machinery". *Applied Soft Computing* 46, pp. 459 –468.

Ertunc H.M., Ocak H., and Aliustaoglu C. (2013) "ANN- and ANFIS-based multi-staged decision algorithm for the detection and diagnosis of bearing faults". *Neural Computing and Applications* 22.SUPPL.1, pp. 435–446.

Fazendeiro P. and Valente de Oliveira J. (2015) "Observer-Biased Fuzzy Clustering". *Fuzzy Systems, IEEE Transactions on* 23.1, pp. 85–97.

Fialho A.S., Vieira S.M., Kaymak Uzay, Almeida R.J., Cismondi F., Reti S.R., Finkelstein S.N., and Sousa J.M.C. (2016) "Mortality prediction of septic shock patients using probabilistic fuzzy systems". *Applied Soft Computing* 42, pp. 194 –203.

Genuer Robin, Poggi Jean, and TuleauMalot Christine (2010) "Variable selection using Random Forests". *Pattern Recognition Letters* 14.31, pp. 2225–2236.

Goode P.V. and Chow Mo-Yuen (1994) "A hybrid fuzzy/neural system used to extract heuristic knowledge from a fault detection problem". *Fuzzy Systems, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the Third IEEE Conference on*, 1731–1736 vol.3.

Harrouche F. and Felkaoui A. (2014) "Automation of fault diagnosis of bearing by application of fuzzy inference system (FIS)". *Mechanics and Industry* 15.6, pp. 477–485.

Huang Yixiang, Gong Liang, Wang Shuangyuan, and Li Lin (2014) "A fuzzy based semi-supervised method for fault diagnosis and performance evaluation". *Advanced Intelligent Mechatronics (AIM), 2014 IEEE/ASME International Conference on*, pp. 1647–1651.

Juuso E., Ruusunen M., and Perigot G. (2010) "Linguistic equation models for failure mode identification from multisensor vibration analysis". *7th International Conference on Condition Monitoring and Machinery Failure Prevention Technologies 2010, CM 2010/MFPT 2010*. Vol. 2, pp. 1408–1420.

Kaplan K., Kuncan M., and Ertunc H.M. (2015) "Prediction of bearing fault size by using model of adaptive neuro-fuzzy inference system". *Signal Processing and Communications Applications Conference (SIU), 2015 23th*, pp. 1925–1928.

Ledo Luis, Lucas Luís Alberto, and Delgado Myriam Regattieri (2014) "Toward automatic rule-base design in probabilistic fuzzy classifies". *III Brazilian Congress on Fuzzy Systems, (III CBSF)*, pp. 69–78.

Lee Hyong-Euk, Park Kwang-Hyun, and Bien Zeungnam Zenn (2008) "Iterative Fuzzy Clustering Algorithm With Supervision to Construct Prob-

abilistic Fuzzy Rule Base From Numerical Data". *Fuzzy Systems, IEEE Transactions on* 16.1, pp. 263–277.

Lei Y., He Z., Zi Y, and Q. Hu (2007) "Fault diagnosis of rotating machinery based on multiple ANFIS combination with GAs". *Mechanical Systems and Signal Processing* 21.5, pp. 2280–2294.

Lesot M.-J. and Kruse R. (2006) "Data Summarisation by Typicality-based Clustering for Vectorial and Non Vectorial Data". *Fuzzy Systems, 2006 IEEE International Conference on*, pp. 547–554.

Li Bing, Liu Peng yuan, Hu Ren xi, Mi Shuang shan, and Fu Jian ping (2012a) "Fuzzy lattice classifier and its application to bearing fault diagnosis". *Applied Soft Computing* 12.6, pp. 1708 –1719.

Li Chuan and Liang Ming (2012) "Continuous-scale mathematical morphology-based optimal scale band demodulation of impulsive feature for bearing defect diagnosis". *Journal of Sound and Vibration* 331.26, pp. 5864 –5879.

Li Chuan, Liang Ming, and Wang Tianyang (2015) "Criterion fusion for spectral segmentation and its application to optimal demodulation of bearing vibration signals". *Mechanical Systems and Signal Processing* 64-65, pp. 132–148.

Li Chuan, Liang Ming, Zhang Yi, and Hou Shumin (2012b) "Multi-scale autocorrelation via morphological wavelet slices for rolling element bearing fault diagnosis". *Mechanical Systems and Signal Processing* 31, pp. 428 –446.

Li Chuan, Valente de Oliveira José, Cerrada Mariela, Pacheco Fannia, Cabrera Diego, Sanchez Vinicio, and Zurita Grover (2016a) "Observer-biased bearing condition monitoring: From fault detection to multi-fault classification". *Engineering Applications of Artificial Intelligence* 50, pp. 287 –301.

Li Chuan, Sanchez Vinicio, Zurita Grover, Lozada Mariela Cerrada, and Cabrera Diego (2016b) "Rolling element bearing defect detection using the generalized synchrosqueezing transform guided by time-frequency ridge enhancement". *ISA Transactions* 60, pp. 274 –284.

Liu J., Wang W., and Golnaraghi F. (2010) "An enhanced diagnostic scheme for bearing condition monitoring". *IEEE Transactions on Instrumentation and Measurement* 59.2, pp. 309–321.

Liu X., Ma L., Zhang S., and Mathew J. (2008) "Feature group optimisation for machinery fault diagnosis based on fuzzy measures". *Australian Journal of Mechanical Engineering* 5.2, pp. 191–197.

Liu Xiaofeng, Ma Lin, and Mathew Joseph (2009) "Machinery fault diagnosis based on fuzzy measure and fuzzy integral data fusion techniques". *Mechanical Systems and Signal Processing* 23.3, pp. 690 –700.

Liu Zhi and Li Han-Xiong (2005) "A Probabilistic Fuzzy Logic System for Modeling and Control". *IEEE Transactions on Fuzzy Systems* 13.6, pp. 848 –859.

Loparo K.A. (2003) *Bearings vibration data set. The Case Western Reserve University Bearing Data Center.* http://www.eecs.cwru.edu/laboratory/bearing/download.htm. [Online; accessed 2011].

Lou X. and Loparo K.A. (2004) "Bearing fault diagnosis based on wavelet transform and fuzzy inference". *Mechanical Systems and Signal Processing* 18.5, pp. 1077–1095.

Marichal G.N., Arts Mariano, Prada J.C. Garca, and Casanova O. (2011) "Extraction of rules for faulty bearing classification by a Neuro-Fuzzy approach". *Mechanical Systems and Signal Processing* 25.6. Interdisciplinary Aspects of Vehicle Dynamics, pp. 2073 –2082.

Meghdadi A.H. and Akbarzadeh-T M.-R. (2001) "Probabilistic fuzzy logic and probabilistic fuzzy systems". *Fuzzy Systems, 2001. The 10th IEEE International Conference on.* Vol. 3, pp. 1127–1130.

Melo Leonardo, Lucas Luís Alberto, and Delgado Myriam Regattieri (2012) "Rule-base Design using Probabilistic Weights: a Preliminary Analysis of Uncertainty Aspects". *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based System*, pp. 655–664.

Sheskin D.J. (2011) *Handbook of Parametric and Nonparametric Statistical Procedures, Fifth Edition.* Taylor & Francis.

Silva Vicente S.A. da, Fujimoto R.Y., and Padovese L.R. (2001) "Rolling bearing fault diagnostic system using fuzzy logic". *Fuzzy Systems, 2001. The 10th IEEE International Conference on.* Vol. 2, 816–819 vol.3.

Siyambalapitiya D.J.Tilak and McLaren P.G. (1990) "Reliability improvement and economic benefits of online monitoring systems for large induction machines". *Industry Applications, IEEE Transactions on* 26.6, pp. 1018–1025.

Sun X., Xia X., Liu Y., and Gao L. (2012) "Evaluation of rolling bearing vibration using fuzzy set theory and chaos theory". *Advanced Materials Research* 424-425, pp. 338–341.

Tang Min, Chen Xia, Hu Weidong, and Yu Wenxian (2012) "Generation of a probabilistic fuzzy rule base by learning from examples". *Information Sciences* 217, pp. 21–30.

Tian Hao, Kang Xiao-Yong, Zhang Jun-Nuo, and Han Shan-Shan (2012) "Application of fuzzy rough sets in patterns recognition of bearing". *Quality, Reliability, Risk, Maintenance, and Safety Engineering (ICQR2MSE), 2012 International Conference on*, pp. 731–734.

Tiwari R., Gupta V.K., and Kankar P.K. (2015) "Bearing fault diagnosis based on multi-scale permutation entropy and adaptive neuro fuzzy classifier". *JVC/Journal of Vibration and Control* 21.3, pp. 461–467.

Valente de Oliveira José (1995) "A design methodology for fuzzy system interfaces". *Fuzzy Systems, IEEE Transactions on* 3.4, pp. 404–414.

— (1999) "Semantic constraints for membership function optimization". *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 29.1, pp. 128–138.

Valente de Oliveira José and Pedrycz Witold (2007) *Advances in Fuzzy Clustering and Its Applications*. John Wiley & Sons, Inc.

Vijay G.S., Pai S.P., Sriram N.S., and Rao R.B.K.N. (2013) "Radial basis function neural network based comparison of dimensionality reduction techniques for effective bearing diagnostics". *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology* 227.6, pp. 640–653.

Waltman L., Kaymak U., and Berg J. van den (2005) "Maximum likelihood parameter estimation in probabilistic fuzzy classifiers". *Fuzzy Systems, 2005. FUZZ '05. The 14th IEEE International Conference on*, pp. 1098–1103.

Wang C.J., Li H.Y., Xiang W., and Zhao D. (2014) "A new signal classification method based on EEMD and FCM and its application in bearing fault diagnosis". *Applied Mechanics and Materials* 602-605, pp. 1803–1806.

Wang Huaqing and Chen Peng (2011) "Intelligent diagnosis method for rolling element bearing faults using possibility theory and neural network". *Computers & Industrial Engineering* 60.4, pp. 511 –518.

Wang Tianyang, Liang Ming, Li Jianyong, Cheng Weidong, and Li Chuan (2015) "Bearing fault diagnosis under unknown variable speed via gear noise cancellation and rotational order sideband identification". *Mechanical Systems and Signal Processing* 62–63, pp. 30 –53.

Wei Zexian, Wang Yanxue, He Shuilong, and Bao Jiading (2017) "A novel intelligent method for bearing fault diagnosis based on affinity propagation clustering and adaptive feature selection". *Knowledge-Based Systems* 116, pp. 1 –12.

Wolpert D. H. and Macready W. G. (1997) "No Free Lunch Theorems for Optimization". *Trans. Evol. Comp* 1.1, pp. 67–82.

Xu W.-X., Tan J.-W., and Zhan H. (2014) "Research and application of the improved dst new method based on fuzzy consistent matrix and the weighted average". *Advanced Materials Research* 1030-1032, pp. 1764–1768.

Xu Z., Xuan J., Shi T., Wu B., and Hu Y. (2009) "A novel fault diagnosis method using pca and art-similarity classifier based on yu's norm". *Key Engineering Materials* 413-414, pp. 569–574.

Xu Z.B., Wu H., and Shi T.L. (2013) "Fault diagnosis of bearing based on selective ensemble of multiple fuzzy ARTMAP neural networks". *Applied Mechanics and Materials* 423-426, pp. 2480–2485.

Yang Y., Zhang L., Zhang L., Cai X., and Zhang S. (2012) "The health degree evaluation of rolling element bearings using an improved BP neural network". *Journal of Information and Computational Science* 9.14, pp. 4217–4227.

Yaqub M.F., Gondal I., and Kamruzzaman J. (2012) "Inchoate Fault Detection Framework: Adaptive Selection of Wavelet Nodes and Cumulant Orders". *Instrumentation and Measurement, IEEE Transactions on* 61.3, pp. 685–695.

Yu X.G. and Liu J. (2011) "Research on fault diagnosis and application for rolling element bearing based on fuzzy analysis". *Advanced Materials Research* 189-193, pp. 1358–1361.

Zhang Geng and Li Han-Xiong (2012) "An Efficient Configuration for Probabilistic Fuzzy Logic System". *IEEE Transactions on Fuzzy Systems* 20.5, pp. 898–909.

Zhang J., Ma W., and Ma L. (2014) "A fault diagnosis method based on ANFIS and bearing fault diagnosis". *Proceedings - 2014 International Conference on Information Science, Electronics and Electrical Engineering, ISEEE 2014*. Vol. 2, pp. 1274–1278.

Zhang Laibin, Wang Zhaohui, and Zhao Shangxin (2007) "Short-term fault prediction of mechanical rotating parts on the basis of fuzzy-grey opti-

mising method". *Mechanical Systems and Signal Processing* 21.2, pp. 856 –865.

Zhao W. and Wang L. (2010) "Rolling bearing fault diagnosis based on wavelet packet feature entropy-MFSVM". *Advanced Materials Research* 121-122, pp. 813–818.

Zhao Xinze, Zhao Chunhua, Gao Hongliang, and Wu Gang (2008) "Knowledge mining for fault diagnosis based on rough sets theory". *FUZZ-IEEE 2008, IEEE International Conference on Fuzzy Systems*, pp. 744–749.

Zheng Jinde, Cheng Junsheng, Yang Yu, and Luo Songrong (2014) "A rolling bearing fault diagnosis method based on multi-scale fuzzy entropy and variable predictive model-based class discrimination". *Mechanism and Machine Theory* 78, pp. 187 –200.

Zio Enrico and Gola Giulio (2006) "A neuro-fuzzy technique for diagnosing faults in rotating machinery". *Proceedings of the European Safety and Reliability Conference 2006, ESREL 2006 - Safety and Reliability for Managing Risk*. Vol. 1, pp. 247–254.

— (2009) "A neuro-fuzzy technique for fault diagnosis and its application to rotating machinery". *Reliability Engineering & System Safety* 94.1, pp. 78 –88.