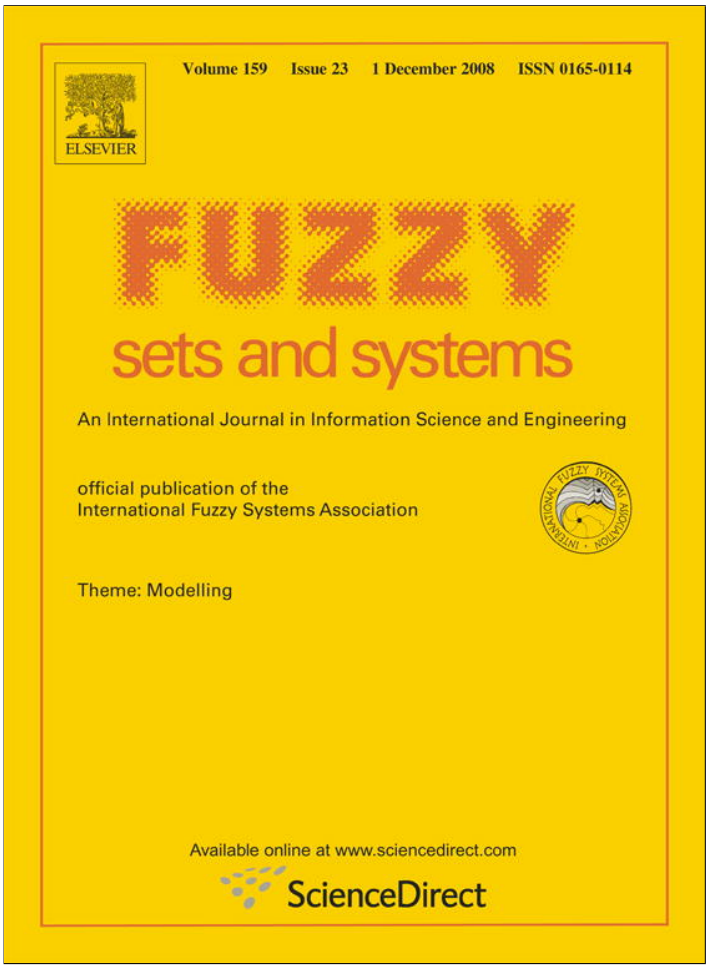
2008-12

Low-level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modelling

Zhou, Shang-Ming

10.1016/j.fss.2008.05.016
Fuzzy Sets and Systems
Elsevier BV



Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Sets and Systems 159 (2008) 3091–3131
FUZZY sets and systems
www.elsevier.com/locate/fss

Low-level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modelling

Shang-Ming Zhou[a,*], John Q. Gan[b]

[a] *Centre for Computational Intelligence, School of Computing, De Montfort University, Leicester LE1 9BH, UK*
[b] *Department of Computing and Electronic Systems, University of Essex, Colchester CO4 3SQ, UK*

Received 28 September 2007; received in revised form 10 May 2008; accepted 26 May 2008
Available online 20 June 2008

Abstract

This paper aims at providing an in-depth overview of designing interpretable fuzzy inference models from data within a unified framework. The objective of complex system modelling is to develop reliable and understandable models for human being to get insights into complex real-world systems whose first-principle models are unknown. Because system behaviour can be described naturally as a series of linguistic rules, data-driven fuzzy modelling becomes an attractive and widely used paradigm for this purpose. However, fuzzy models constructed from data by adaptive learning algorithms usually suffer from the loss of model interpretability. Model accuracy and interpretability are two conflicting objectives, so interpretation preservation during adaptation in data-driven fuzzy system modelling is a challenging task, which has received much attention in fuzzy system modelling community. In order to clearly discriminate the different roles of fuzzy sets, input variables, and other components in achieving an interpretable fuzzy model, a taxonomy of fuzzy model interpretability is first proposed in terms of *low-level interpretability* and *high-level interpretability* in this paper. The low-level interpretability of fuzzy models refers to fuzzy model interpretability achieved by optimizing the membership functions in terms of semantic criteria on fuzzy set level, while the high-level interpretability refers to fuzzy model interpretability obtained by dealing with the coverage, completeness, and consistency of the rules in terms of the criteria on fuzzy rule level. Some criteria for low-level interpretability and high-level interpretability are identified, respectively. Different data-driven fuzzy modelling techniques in the literature focusing on the interpretability issues are reviewed and discussed from the perspective of low-level interpretability and high-level interpretability. Furthermore, some open problems about interpretable fuzzy models are identified and some potential new research directions on fuzzy model interpretability are also suggested.

Keywords: Data-driven fuzzy systems; Interpretable; Fuzzy models; Interpretability; Transparency; Criteria; Parsimony; Distinguishability; Low-level interpretability; High-level interpretability.

1. Introduction

As the main objective in system modelling, development of reliable and understandable models is crucial for human being to understand real-world systems or natural phenomena. Generally there are three different strategies for system modelling. The first is white-box modelling, in which the parameters characterizing the systems have clear and interpretable physical meanings (one typical example is the Newton's universal gravitation law). However, white-box

* Corresponding author.
E-mail address: smzhou@ieee.org (S.-M. Zhou).

0165-0114/$ - see front matter
doi:10.1016/j.fss.2008.05.016

modelling usually becomes impractical when complex systems are considered. The second strategy is black-box modelling without using prior knowledge, in which relationships between inputs and outputs are established fully based on observational data. Black-box modelling can simulate a real-world system reliably and precisely, but the model structure and parameters usually give no explicit explanation about the system behaviours. For example, in medical domains, although black-box models, such as bagged decision trees and neural networks, can achieve good fit performance, the resulting decisions received much suspicion [1,186]. The third strategy, grey-box modelling [41,96], can be referred to as an eclecticism between precision and interpretability, in which the prior knowledge about the system is considered and unknown parts of the system are identified by black-box modelling approaches. As one of the most successful tools to develop grey-box models [112], fuzzy modelling describes systems by establishing relationships between input and output variables in terms of a fuzzy logic-based descriptive language [114–116,191,193].

Compared to black-box modelling, fuzzy modelling formulates the system knowledge with rules in a transparent way to interpretation and analysis (to a certain degree) so as to gain insights into the system being modelled. This has opened up a brand-new approach to modelling complex systems in system identification and control, fault diagnosis, classification and intelligent data analysis, etc. Fuzzy rules can be generated based on human expert knowledge or heuristics, which offers a good high-level semantic generalization capability. However, for complex systems, the interactions between system behaviours are very difficult to be grasped so that the system model just based on expert knowledge may suffer from a loss of accuracy [68] and combinatorial rule explosion. On the other hand, due to the advanced development of modern information technology, the ever-increasing quantity of data is available from complex industrial and commercial processes (systems). In the past decade or so, data-driven fuzzy rule generation has been widely investigated and shown to be very successful. Compared to heuristic fuzzy rules, fuzzy rules generated from data are able to extract more specific knowledge for more complex systems.

It is usually assumed that the interpretability of fuzzy models is automatically given just due to using linguistic rules. However, it is not true when adaptive learning techniques are used to optimize the fuzzy inference processes for complex systems. In accuracy-oriented adaptive learning processes, interpretation preservation during adaptation cannot always be guaranteed due to the conflicting objectives of accuracy and interpretation [8,24,28,67,68,86,121,132,160,180], so model interpretability is usually lost, which has been referred to as an incentive for Mamdani, the founder of fuzzy control, to recommend against the use of adaptation of fuzzy inference parameters [115]. Particularly, because one of the important incentives of introducing fuzzy sets for modelling complex systems is that they can formulate the knowledge extracted from data or supplied by experts with a more transparent way to gain insights into the complex systems, it is often desirable and sometimes even essential to be able to interpret the final model structure. Rule base legibility is an important condition to take full advantage of fuzzy models. Hence, model interpretability improvement is regarded as one of the most important issues in data-driven fuzzy modelling [5,25,46,83,91,94,93,113,156,174,179,190].

However, global model accuracy and interpretability are two conflicting modelling objectives, as improving interpretability of fuzzy models generally degrades the global model performance of fuzzy models, and vice versa. Hence one challenging problem is how to construct a fuzzy model with not only good global performance but also good model interpretability, particularly in coping with high-dimensional input space, that is to say, how to achieve a fuzzy model with good performance trade-off between global model accuracy and model interpretability. An interpretable fuzzy model is expected to provide high numeric precision while incurring as little a loss of linguistic descriptive power as possible [5,45,83,139,190,198].

It is noteworthy that as we address the interpretability issues in fuzzy modelling, transparency and interpretability issues in traditional statistical system modelling have already received much attention from the aspect of complexity reduction [33,55,73,75,97,164,177,178]. The objective of complexity reduction-based statistical system modelling is to provide a trade-off between how well the model fits the training data and model complexity. Model complexity depends on the number of independent and adjustable parameters, also called degrees of freedom, to be adapted during the learning process. Traditionally, the well-known complexity reduction techniques for achieving transparent and interpretable system models include input feature selection [196], orthogonal least square (OLS) [30], basis pursuit [29,31], sparse regression [170], support vector machines (SVM) [175,176], etc. More importantly, these complexity reduction techniques possess strong potentials of improving fuzzy model interpretability when applied to fuzzy modelling. A natural question arises: what is the difference between the interpretability achieved by considering fuzzy systems' own idiosyncrasies and the one obtained by introducing the traditional complexity reduction techniques into fuzzy modelling? Guillaume has provided a good overview of interpretability-oriented fuzzy inference systems designed from data [67]. However, the above question was not considered and addressed in [67]. The purpose of this paper is

not only to provide an overview of the state-of-the-art of data-driven interpretable fuzzy modelling techniques, but also to offer an overall view of this domain in a framework by syncretizing the traditional parsimonious statistical system modelling with interpretable fuzzy system modelling, so that one may form a systematic picture of the field, instead of focusing on specific parts, dealing with particular methods and techniques. To the best of our knowledge, this type of effort has not been reported yet in literature.

As a matter of fact, in the traditional statistical modelling techniques, only the influences of input variables on system outputs are taken into account to construct transparent and interpretable system models. While in fuzzy system modelling, due to its own idiosyncrasies, additional influences on system outputs should also be taken into account to achieve a transparent and interpretable fuzzy model, such as the number of membership functions (MFs), coverage and distinguishability of MFs, normality of MFs. In order to clearly discriminate the different roles of fuzzy sets, input variables and other components in achieving a transparent fuzzy model, a framework is suggested in this paper to categorize fuzzy model interpretability into *low-level interpretability* and *high-level interpretability*, and the criteria are defined for low-level interpretability and high-level interpretability, respectively. Low-level interpretability of fuzzy models is achieved on fuzzy set level by optimizing MFs in terms of the semantic criteria on MFs, whilst high-level interpretability is obtained on fuzzy rule level by conducting overall complexity reduction in terms of some criteria, such as a moderate number of variables and rules, completeness and consistency of rules. The complexity reduction techniques used in traditional system modelling, if exploited in fuzzy system modelling, can serve as fuzzy rule optimization in nature, which corresponds to aiming at the parsimony of the fuzzy rule base, one of the main high-level interpretability criteria of fuzzy systems. This clarification is helpful, as there are plentiful traditional system modelling methods on complexity reduction with great potentials of being used to induce compact rule base in fuzzy system modelling.

This paper is organized as follows. Section 2 briefly presents the currently used fuzzy rule structures. Various definitions of fuzzy model interpretability are given in Section 3, followed by criteria for fuzzy model interpretability in Section 4. Constructive techniques for fuzzy model interpretability are analysed and reviewed in Section 5. Section 6 addresses some open problems and potential new research topics on fuzzy model interpretability, followed by conclusions in Section 7.

2. Rule structures of fuzzy rule-based systems

Basically, in fuzzy modelling the relationships between input and output variables are established in terms of a descriptive language based on if–then rules:

$$\text{If } \textit{antecedent proposition} \text{ then } \textit{consequent proposition} \tag{1}$$

Depending on the consequent part there exist different rule structures with different capabilities of description and approximation for rule-based fuzzy modelling.

Linguistic fuzzy model, also known as Mamdani type fuzzy model [114–116,193], is represented by linguistic rules with the following structure:

$$\textit{Rule}_i: \text{ If } x_1 \text{ is } A_{i,1} \textit{ and } \cdots \textit{ } x_n \textit{ is } A_{i,n} \text{ then } y \textit{ is } B_i \quad (i = 1, \ldots, L) \tag{2}$$

where $\textit{Rule}_i$ denotes the ith rule, L is the number of rules in the rule base, $x = (x_1, \ldots, x_n)^{\mathrm{T}}$ and y are the input and output linguistic variables, respectively, $A_{i,j}$ and B_i are the linguistic labels expressed as fuzzy sets with specific semantic meanings regarding the behaviours of the system being modelled. These fuzzy sets are characterized by MFs generated by expert knowledge or fully from data. An outstanding advantage of the linguistic fuzzy model lies in that it may offer a high semantic level with good interpretability and a good generalization capability.

In order to enhance the representation power of a fuzzy system, Takagi and Sugeno proposed a fuzzy model (TS fuzzy model) that differs from the linguistic one by using the following different consequent structure [168]:

$$\begin{aligned}&\textit{Rule}_i: \text{If } x_1 \textit{ is } A_{i,1} \textit{ and } \cdots \textit{ } x_n \textit{ is } A_{i,n} \text{ then}\\ &\quad y_i = a_{0i} + a_{1i}x_1 + \cdots + a_{ni}x_n \quad (i = 1, \ldots, L)\end{aligned} \tag{3}$$

where x_i are input variables and y_i are local output variables that determines local linear input–output relations by means of the real-valued coefficients a_{ji}. The output of a fuzzy system with a knowledge base composed of L TS rules

is computed as the weighted average of the individual rule outputs y_i, $i = 1, \ldots, L$:

$$y = \sum_{i=1}^{L} r_i(x) y_i \Big/ \sum_{i=1}^{L} r_i(x) \tag{4}$$

with $r_i(x) = \mathrm{T}_j \mu_{i,j}(x_j)$ being the matching degree between the antecedent part of the ith rule and the current system inputs $x_1, \ldots, x_n$, where T is a t-norm and $\mu_{i,j}(\cdot)$ the MF of fuzzy set $A_{i,j}$.

Compared to a linguistic model, the rule consequent part in the TS model is an affine linear function of input variables. As such, each rule can be considered as a *local linear model* that is fused together by means of aggregation to produce an overall output$y$. A special case of the affine function with offset $a_{0i} = 0$ ($i = 1, \ldots, L$) results in a homogeneous TS system. Another interesting special case is that the consequent is a constant, i.e., $y_i = a_{0i}$, which is called 0-order TS system. Actually the 0-order TS model retains certain linguistic interpretability in the manner of a linguistic model while possessing the attractive properties of the TS system, particularly the automatic determination of model parameters from data [157]. Some researchers have made efforts to modify the form of the consequent polynomials of TS models, which allows specific interpretation to the systems being modelled [7,18,54].

However, either the Mamdani fuzzy rule in the form of (2) or the TS rule in the form of (3) shares the same fuzzy sets of each input variable with other rules, i.e., the Mamdani rule (2) and the TS rule (3) are based on the global grid, which offers good linguistic readability [67]. The disadvantage is that the fuzzy systems built based on (2) or (3) suffer from the curse of dimensionality in high-dimensional input space: the problem will increase exponentially in volume associated with adding extra dimensions to the input space. The curse of dimensionality is a significant obstacle to solving machine learning problems that involve learning a "state-of-nature'' from a finite number of data samples in a high-dimensional space. One efficient way of evading this problem for fuzzy systems is to generate the input space via scatter-partition (prototype) rather than grid-partition. The Mamdani and TS fuzzy rules generated by scatter-partition have the following forms respectively:

$$\mathit{Rule}_i : \text{If } x \text{ is } A_i \text{ then } y \text{ is } B_i \tag{5}$$

and

$$\mathit{Rule}_i : \text{If } x \text{ is } A_i \text{ then } y_i = a_{0i} + a_{1i}x_1 + \cdots + a_{ni}x_n \tag{6}$$

where $x = t(x_1, x_2, \ldots, x_n) \in \Re^n$ is an n-dimensional input vector and $A_i \in F(\Re^n)$ are the fuzzy sets defined on $\Re^n$ with multi-dimensional MFs. Very interestingly, the fuzzy systems consisting of multi-dimensional Mamdani rules (5) are totally equivalent to fuzzy graphs introduced by Zadeh [192,195]. A fuzzy graph is build by a collection of fuzzy points corresponding to fuzzy rules, which represents a functional dependency f^* of Y on X:

$$f^* = A_1 \times B_1 + \cdots + A_n \times B_n \tag{7}$$

Some researchers have made efforts to construct fuzzy graphs from data [6,13]. More details about how to generate the prototype-based fuzzy systems and their pros and cons will be reviewed in Section 5.2.4) of this paper.

Interestingly, different from the above rule structures Gegov proposed to use Boolean matrices and binary relations to express fuzzy models [62]. For example, given the following fuzzy rule base:

Rule_1 : If x_1 *is Small and* x_2 *is Small* then y *is Negative*

Rule_2 : If x_1 *is Small and* x_2 *is Big* then y *is Positive*

Rule_3 : If x_1 *is Big and* x_2*is Small* then y *is Negative*

Rule_4 : If x_1 *is Big and* x_2 *is Big* then y *is Zero*

If *Small* $= 1$, *Big* $= 2$, *Negative* $= 1$, *Zero* $= 2$, *Positive* $= 3$, then a Boolean relation and a binary relation that equivalently express the above fuzzy system individually are indicated in Table 1 and (8).

$$\{[\mathit{and}\ (11, 1)]\ \mathit{or}\ [\mathit{and}\ (12, 3)]\ \mathit{or}\ [\mathit{and}\ (21, 1)]\mathit{or}[\mathit{and}\ (22, 2)]\} \tag{8}$$

Moreover, for the sake of inducing compact rule base a formal approach to manipulation of rule bases is suggested by performing operations on Boolean matrices and binary relations [62]. Because Boolean transformations of rule base

Table 1
Boolean relation for a fuzzy system

Inputs	Outputs		
	1	2	3
11	1	0	0
12	0	0	1
21	1	0	0
22	0	1	0

can lead to more compact representations, Klose and Nurnberger presented an approach to building more expressive rules by performing Boolean transforming during and after learning from data [103].

3. Definitions of fuzzy model interpretability

Although interpretability issues in fuzzy system modelling and statistical system modelling have received much attention in recent years, there is no well-established definition about model interpretability. In this section, we review the existing efforts made by the researchers on addressing the connotations of model interpretability and propose a new framework for fuzzy model interpretability.

3.1. Transparency and interpretability

Model *transparency* is defined as *a property that enables us to understand and analyse the influence of each system parameter on the system output* [22,74,156]. In connection with transparency in fuzzy modelling, another symbiotic term is *interpretability* mentioned by Jang et al. [86], in which some basic ideas were discussed about how to constrain the optimization of the adaptive-network-based fuzzy inference system (ANFIS) [87] to preserve the interpretability.

In Riid's opinion [142], transparency and interpretability do not match when used to characterize fuzzy systems. Interpretability is a property of fuzzy systems which exists by default, being established with linguistic rules and fuzzy sets associated with these rules. Transparency, on the other hand, is not a default property of fuzzy systems, but a measure of how valid or how reliable the linguistic interpretation of the system is. However, the meanings of these two concepts possess so much synonymity in describing system modelling. Particularly, in fuzzy modelling, it may not be reasonable to distinguish them, because the two terms had been used in parallel in traditional statistical system modelling [30,33,55,73,164,170], long before they were used in fuzzy system modelling [3,11,19,20,26]. Transparency and interpretability share the same connotations according to the two definitions in practice. In this paper, the two terms are used in parallel with the same meaning in fuzzy modelling, unless otherwise stated.

3.2. Formalized definitions

Some researchers have tried to give formalized definitions for interpretability or transparency of fuzzy models.

Definition 1 (*Bodenhofer and Bauer [19,20]*). Consider a linguistic variable $V = (N, T, X, G, S)$ and an index set I, where N is the name of the linguistic variable V, G is a grammar, T is the so-called *term set*, i.e., the set of linguistic expressions resulting from G, X is the universe of discourse, and S is a $T \to F(X)$ mapping which defines the semantics, i.e., a fuzzy set on X, of each linguistic expression in T. Let $RE = (RE_i)_{i\in I}$ be a family of relations on the set of verbal values T, where each relation RE_i has a finite entry a_i. Assume that, for every relation RE_i, there exists a relation Q_i on the fuzzy power set $F(X)$ with the same entry. [1] Let Q represent the family $(Q_i)_{i\in I}$. Then the linguistic variable V is called *R-Q-interpretable* if and only if the following holds for all $i \in I$ and all $x_1, \ldots, x_{a_i} \in T$:

$$RE_i(x_1, \ldots, x_{a_i}) \Rightarrow Q_i(s(x_1), \ldots, s(x_{a_i})) \tag{9}$$

where $\Rightarrow$ represent an implication relation.

[1] Q_i is associated with the "semantic counterpart'' of RE_i, i.e., the relation that models RE_i on the semantic level.

Riid and Rustern gave a formalized definition for transparency of linguistic fuzzy model as follows.

Definition 2 (*Riid [142], Riid et al. [143]*). The ith rule of the linguistic fuzzy system (2) is transparent if its activation degree satisfies

$$r_i(x) = T_j \mu_{i,j}(x_j) = 1 \tag{10}$$

and results in system output as $y = y_i$, where y_i is the centre of the output MF $B_i(y)$ associated with the activated rule.

In such a way, to define the transparency of 1-order TS fuzzy system (3) and (4) is much more complicated [142–144]. Riid et al. suggested that the first order TS fuzzy system is transparent if the global output y can be derived directly on the basis of its local output y_i, and they proposed a measure for transparency evaluation, but this measure results in further problems [143], i.e., the minimum transparency error is obtained for systems that are non-fuzzy, and there is no trivial way for preserving this transparency for the first order TS fuzzy system.

Unlike the above two formalized definitions, Nauck suggested an index to measure the interpretability of fuzzy rule bases for classification problems in terms of complexity, the degree of coverage of fuzzy partition over the domain, and a partition index that penalizes partitions with a high granularity [124]. However, this definition only considers special circumstances of model construction for classification instead of prediction. More importantly, it ignores many other aspects of interpretable fuzzy models, such as distinguishability, rule base simplicity, etc.

In interpretable fuzzy modelling, it might be not feasible to give a formalized definition for transparency of fuzzy models in practice [28], because the interpretability of fuzzy models heavily depends on human's prior knowledge [58,59,198]. To decide whether a specific fuzzy model is interpretable or not is a highly subjective task and it is controversial sometimes [38,121]. Some researchers have become aware of this matter and proposed some constraint criteria that fuzzy models should meet to assure good interpretability during model adaptation or optimization [45,67,135–139].

3.3. A framework for fuzzy model interpretability

In order to clearly discriminate the different roles of fuzzy sets, input variables and other components in achieving an interpretable fuzzy model, we propose in this paper a taxonomy of fuzzy model interpretability: *low-level interpretability* and *high-level interpretability*. First let us review fuzzy models from the perspective of statistical system modelling.

3.3.1. Review of fuzzy models from the perspective of traditional statistical system modelling

At first sight, fuzzy modelling algorithms for the purpose of improving interpretability may seem rather strange and hardly related to the existing methods of traditional statistical parsimonious system modelling techniques. Once a fuzzy model is cast into a more standard mathematical notation, we will observe its connections to the traditional statistical system modelling methods. In the following, we consider the TS fuzzy models. However, depending on the defuzzification method, the Mamdani models can also be expressed in the similar manner, i.e., as *linear-in-parameters* models for regression problems. Eq. (4) can be rewritten in a slightly different form,

$$\widehat{y} = \sum_{i=1}^{L} p_i(x)(a_{0i} + a_{1i}x_1 + \cdots + a_{ni}x_n) \tag{11}$$

where

$$p_i(x) = r_i(x) \Big/ \sum_{i=1}^{L} r_i(x) \tag{12}$$

is the normalized firing strength of the ith rule. The great advantage of the fuzzy model is its representative power, that is, to describe a highly nonlinear system by using simple local models (rules).

On the other hand, Eq. (11) can also be viewed as a *linear-in-parameters* model,

$$\widehat{y}(x, \theta) = \sum_{i=1}^{L} w_i^{\mathrm{T}}(x)\theta_i \tag{13}$$

where $w_i(x) = [p_i(x), p_i(x)x^{\mathrm{T}}]^{\mathrm{T}} \in R^{n+1}$ and $\theta_i = [a_{0i}, a_{1i}, \ldots, a_{ni}]^{\mathrm{T}} \in R^{n+1}$. If basis functions $w_i(x)$ are fixed, parameterization of (13) becomes linear with respect to parameters θ_i. Given N input–output data sequences $\{x^{(k)}, y^{(k)}\}$, $k = 1, \ldots, N$, (13) can be expressed in the matrix form,

$$Y = W\theta + \varepsilon \tag{14}$$

where $Y = [y^{(1)}, y^{(2)}, \ldots, y^{(N)}]^{\mathrm{T}} \in R^N$, $W = [w^{(1)}, w^{(2)}, \ldots, w^{(L)}] \in R^{N\times(n+1)L}$, with $w^{(i)} = [w_i(x^{(1)}), w_i(x^{(2)}), \ldots, w_i(x^{(N)})]^{\mathrm{T}} \in R^{N\times(n+1)}$, which is called *firing strength matrix* with each column corresponding to one of the fuzzy rules, $\theta = [\theta_1, \theta_2, \ldots, \theta_{\mathrm{L}}]^{\mathrm{T}} \in R^{(n+1)L}$, and ε represents model error. Thus, the linear-in-parameters model can be addressed further from the angle of statistical regression analysis.

According to the taxonomy developed in [33] to categorize methods for estimating continuous-valued functions from noisy samples, (13) is a *dictionary representation* in nature, comparing to *kernel representation* in which a continuous-valued function estimator is expressed as a distance-weighted combinations of observed output values. The set of $w_i(x)$ is called a *dictionary*. The goal of a predictive learning system that employs (13) for function regression is to adjust the degrees of freedom in (13) based on training samples, in such a way, the approximation function provides minimum prediction risk. In case the transparency of the fuzzy model with (13) and (14) is considered, the problem is reduced to the automatic selection of the dictionary from observed data accounting for the interpretability of the rule base. Thereupon, many statistical regression techniques, such as regularization and sparse regression, can be used or extended to attack this problem. However, one fundamental problem is that as the number of predictors in the regression function increases the curse of dimensionality applies.

On the other hand, when viewing regression model (13) as a function estimation problem, one attractive approach for ameliorating the curse of dimensionality is to model the regression function as an additive function of the predictors [164,177]. This approach has been popularized by Hastie and Tibshirani [75], who emphasized the use of back-fitting together with a one-dimensional smoother to fit the additive models to data as follows:

$$f(x_1, \ldots, x_n) = f_0 + \sum_{i=1}^{n} f_i(x_i) \tag{15}$$

where f_i are "smooth'' functions obtained by some smoothing processes, including the use of cubic smoothing splines. More general models for f, which allow the explicit modelling and visualization of possible interactions between variables, are expressed via functional analysis of variance decompositions, that is, the output can be represented by the ANalysis Of VAriance (ANOVA) decomposition [177,178]. An often used special ANOVA decomposition is expressed as follows:

$$f(x) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{i=1}^{n}\sum_{j=i+1}^{n} f_{i,j}(x_i, x_j) + \cdots + f_{1,2,\ldots,n}(x) \tag{16}$$

where f_0 is a constant, f_i are the x_i main effects, $f_{i,j}$ are the effects from interactions between x_i and x_j, and so on. The regression function is thus represented as additive components by the subset of the terms from this expansion, which shows the attraction of decomposing the model into simpler and more transparent pieces that can be easily interpreted.

Another statistical system modelling technique that can produce a parsimonious model structure is the method of orthogonal-least squares (OLS) [30] when applied to a linear-in-parameters model (14). The OLS method transforms the columns of the coefficient matrix W into a set of orthogonal basis vectors. The Gram–Schmidt orthogonalization procedure is used to perform this decomposition, i.e., $W = UA$, where U is an orthogonal matrix such that $U^{\mathrm{T}}U = I$ (I is the unity matrix) and A is an upper-triangular matrix with unity diagonal elements. Substituting $W = UA$ into (14), we have

$$Y = UA\theta + \varepsilon = Uz + \varepsilon \tag{17}$$

where $z = A\theta$. Let u_i be the ith column of U, so the OLS solution of the system is

$$\hat{z}_i = \frac{u_i^{\mathrm{T}}Y}{u_i^{\mathrm{T}}u_i} \quad (i = 1, \ldots, L) \tag{18}$$

Then, the optimal $\hat{\theta}$ can be computed from triangular system $\hat{z} = A\theta$, in which $\hat{z} = [\hat{z}_1, \ldots, \hat{z}_L]^T$. Because vectors u_i are orthogonal, there is no covariance. Hence the column individual contributions are additive as indicated by the sum of squares of Y:

$$Y^T Y = \sum_{i=1}^{L} z_i^2 u_i^T u_i + \varepsilon^T \varepsilon \tag{19}$$

The part of the output variance $Y^T Y/N$ explained by the regressors is $\sum_{i=1}^{L} z_i^2 u_i^T u_i/N$. Thus, an error reduction ratio [30] due to vector u_i can be defined as

$$[err]^i = \frac{z_i^2 u_i^T u_i}{Y^T Y} \tag{20}$$

At each step the OLS algorithm selects vector u_i that maximizes the explained variance of the observed output in terms of criterion (20). The selection stops when the cumulated explained variance is satisfactory, i.e., the output is reconstructed well enough, which occurs at step r when

$$1 - \sum_{i=1}^{r} [err]^i < \varepsilon \tag{21}$$

with ε being a threshold value.

3.3.2. The framework of low-level interpretability and high-level interpretability of fuzzy models

Generally speaking, the identification of fuzzy models from observational data consists of two parts: (i) structure identification for rule induction and (ii) parameter estimation for identifying the number of rules M, the antecedent fuzzy sets $A_{i,j}$, and the consequent parameters a_{ji} (for the Mamdani fuzzy model, the identification of consequent fuzzy sets B_i corresponds to the identification of the consequent parameters a_{0i}). Hence, in nonlinear fuzzy modelling the interpretability of the model hails from two aspects: the structure of rule base and the expression of induced fuzzy sets. However, the traditional statistical parsimonious system modelling techniques, when applied to fuzzy system modelling, aim at identifying the structure of rule base to produce compact rule bases of fuzzy models in nature, in which model transparency only depends on the regression function representation without considering the expressions of fuzzy sets as an issue of interpretability improvement.

A taxonomy for fuzzy model interpretability is suggested in this paper, by which fuzzy model interpretability is categorized into *low-level interpretability* and *high-level interpretability*. Low-level interpretability is achieved by optimizing MFs on fuzzy set level. Specifically speaking, as shown in Fig. 1, low-level interpretability hails from the improvement on interpretability by introducing semantic constraint criteria into fuzzy modelling, which focus on the modifications of MFs. Some helpful semantic criteria include distinguishability of MFs, moderate number of MFs, natural zero positioning, normality and coverage [45,136,135,139]. High-level interpretability is obtained by performing operations on fuzzy rules to generate a compact and consistent rule base. The criteria, including moderate number of variables and rules, single-rule readability, completeness and consistency of rules, are useful in achieving a high-level interpretable fuzzy model. Thus, in terms of this taxonomy, the traditional statistical parsimonious system modelling techniques, if employed to fuzzy system modelling, are to improve the high-level interpretability of fuzzy models in nature.

This clarification is helpful, because according to this taxonomy, the plentiful traditional system modelling methods on complexity reduction have the great potentials of being used to improve high-level interpretability in fuzzy modelling. However, it should be noted that the main role of traditional statistical parsimonious system modelling techniques playing in fuzzy system modelling is to produce a parsimonious rule base. Parsimony of a fuzzy rule base is just one of the high-level interpretability issues, it not only depends on the number of rules or the avoidance of redundant rules, but also relies on the number of fuzzy sets used in the rule premise parts. Many other high-level interpretability issues, such as readability of single rule, consistency of rules, completeness of rules, etc., have not been involved in traditional statistical parsimonious system modelling techniques.

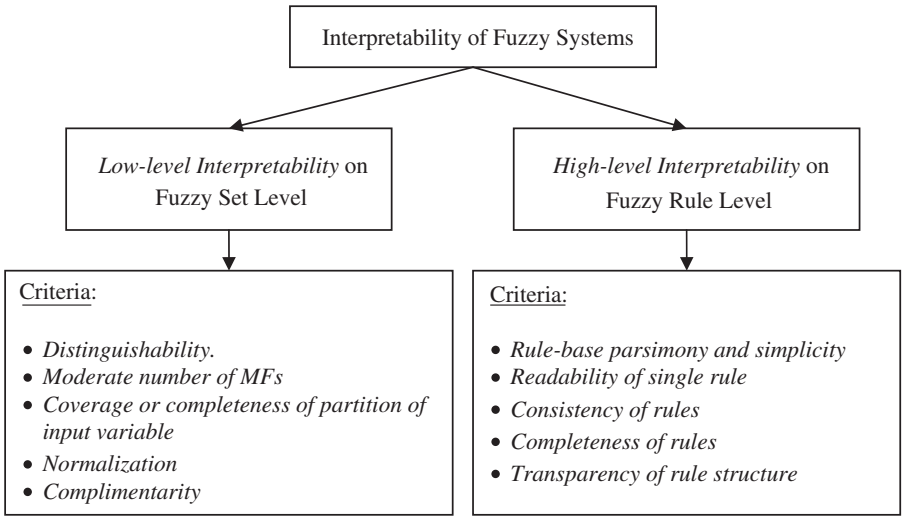


Fig. 1. A taxonomy of interpretability of fuzzy systems.

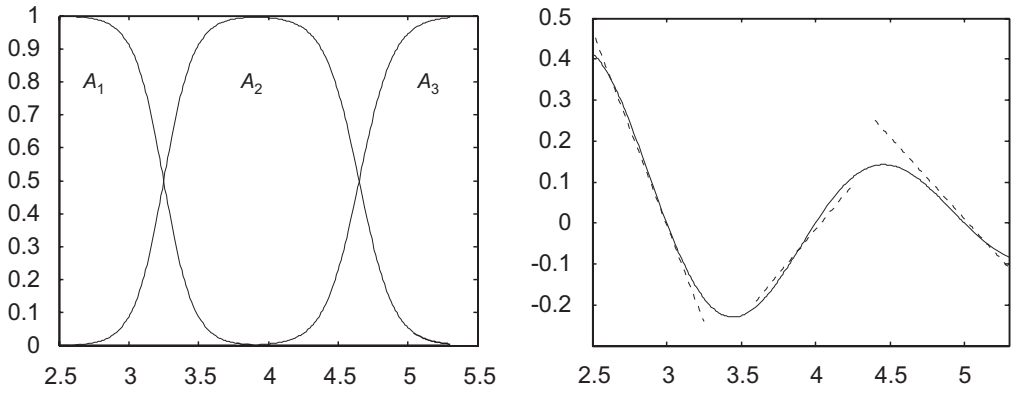


Fig. 2. TS model with good low-level interpretability and high-level interpretability: MFs for input space partition (left) and TS local models (right)—solid line represents the global model, dotted lines represent the local models.

Fig. 2 shows a TS fuzzy model that possesses good low-level interpretability and high-level interpretability. This TS model consists of the following three fuzzy rules:

$$Rule_1: \text{If } x \text{ is } A_1 \text{ then } y_1 = -0.949x + 2.842$$

$$Rule_2: \text{If } x \text{ is } A_2 \text{ then } y_2 = 0.437x - 1.762$$

$$Rule_3: \text{If } x \text{ is } A_3 \text{ then } y_3 = -0.400x + 2.01$$

where A_1, A_2 and A_3 are the fuzzy sets used to partition the input space. The three fuzzy sets with normal MFs show good distinguishability and coverage of the input space. The TS local linear models show good interaction with the global model: they tend to represent the system behaviours in their corresponding subareas. The TS fuzzy model interpretability due to its rule structure, which is different from the Mamdani model, will be addressed in Section 4.2.1).

4. Criteria for fuzzy model interpretability

As addressed above, it might be not feasible to use a formalized definition about fuzzy model interpretability in practice [28]. Some researchers instead handled the interpretability issues by proposing some constraint criteria that fuzzy models should meet to assure good interpretability during model adaptation or optimization. We categorize these

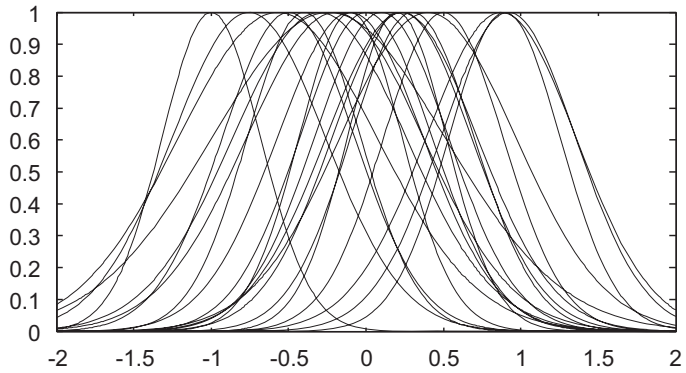


Fig. 3. Fuzzy sets without distinguishability.

criteria into the semantic criteria for *low-level fuzzy model interpretability* and the criteria for *high-level fuzzy model interpretability*.

4.1. Criteria for low-level fuzzy model interpretability

The following semantic criteria describe a set of conditions which, if satisfied by the MFs of a fuzzy variable, are of great benefit to assigning linguistic terms [44,45,135–138]. These criteria defined by de Oliveira [44] do not focus on the absolute meaning of a single linguistic term, but consider the meaning of the ensemble of labels. In this paper, we refer to them as semantic conditions that a fuzzy model should meet to achieve low-level interpretability.

4.1.1. Definitions of semantic criteria

These criteria include distinguishability of MFs, normalization of MFs, moderate number of linguistic terms per variable, and coverage of the universe of discourse.

4.1.1.1. Distinguishability Distinguishability is an essential and basic criterion, because, in input space partitioning for interpretable fuzzy modelling, fuzzy sets should clearly define the distinctive ranges in the universe of discourse of a variable, and each MF should be distinct enough from each other so as to represent a linguistic term with a clear semantic meaning. Fig. 3 shows that a certain distinguishability of the fuzzy sets may be lost, as a result, it is difficult to assign distinct linguistic labels and semantic meaning to these fuzzy sets. For example, in data-driven neuro-fuzzy modelling, an accuracy-oriented adaptive learning algorithm is implemented in a fuzzy inference process, and undistinguishable fuzzy sets are often generated due to its accuracy-oriented nature. These undistinguishable fuzzy sets may be beneficial to improving the training performance, but will usually deteriorate the generalization quality and interpretability of the fuzzy system.

4.1.1.2. Moderate number of MFs The number of MFs of a variable should not be arbitrary, but should be compatible with the number of conceptual entities a human being can efficiently handle and apply during inferential activities. According to a well-known rule of thumb in cognitive psychology, the number of different entities efficiently stored at the short-term memory should not exceed the limit of 7 ± 2. This observation on the human information processing limitations has guided most of the strategies for segmenting, factoring, or decomposing problems in computer science and other fields [45,135].

4.1.1.3. Coverage or completeness of fuzzy partitioning The entire universe of discourse of a variable should be covered by the MFs generated, and every data point should belong to at least one of the fuzzy sets and have a linguistic representation, that is, it is required that membership value should not be zero for at least one of the linguistic labels.

4.1.1.4. Normalization In interpretability-oriented fuzzy modelling, each MF of a variable is expected to represent a linguistic label with clear semantic meaning, thus, at least one data point in the universe of discourse should have a membership value equal to one, that is, MFs of a variable should be normal.

4.1.1.5. Complementarity For each element of the universe of discourse, the sum of all its membership values should be equal to one. This guarantees uniform distribution of meaning among the elements.

However, the *complementarity* requirement is only suitable for probability fuzzy systems (i.e., the sum of all the membership values for every input vector is 1) to guarantee uniform distribution of meaning among the elements so that a sufficient overlapping of MFs is obtained. But a possibility fuzzy system (i.e., the sum of all the membership values for every input vector is between 0 and 1) does not consider this requirement. In practice most neuro-fuzzy systems are possibility fuzzy systems in which sufficient overlapping is necessary for model accuracy but not for model interpretability.

4.1.2. Formalized expressions for semantic criteria in improving low-level fuzzy model interpretability

These semantic criteria have been shown to be very helpful in improving fuzzy model low-level interpretability during parameter optimization or rule generation from data [45,136,135]. To develop an interpretable controller, Lotfi et al. [113] and Anderson [8], from the perspective of system control, illustrated the importance of interpretability and defined a set of constraints on the parameters of fuzzy inference systems for interpretation preservation by a method of loosely limiting the location of the MFs during learning. As a result, the completeness of the fuzzy partitioning will be kept and the distinguishability of different fuzzy sets will be preserved. de Oliveira proposed the following constraint expression for distinguishability and used it to enforce the interpretability of fuzzy systems during MFs optimization [45]:

$$J_2 = \frac{1}{2}\sum_{k=1}^{N}(M_p(x^{(k)}) - 1)^2 I(M_p(x^{(k)}) - 1) \tag{22}$$

where $x^{(k)}$ is the kth sample, $M_p(\cdot)$ is the sigma-count measure of the internal representation of membership degrees $\{\mu_i(x^{(k)})\}_{i=1}^{n}$, defined by

$$M_p(x^{(k)}) = \sqrt[p]{\sum_{i=1}^{n_i}(\mu_i(x^{(k)}))^p} \tag{23}$$

and

$$I(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leqslant 0 \end{cases} \tag{24}$$

where $\mu_i(\cdot)$ is the MF of fuzzy set A_i, p is used to specify the strength of the distinguishability requirement imposed on the linguistic terms. The situation where $p = 1$ exhibits a strong constraint satisfaction, whereas $p = \infty$ describes a weak constraint. The rationality of (22) lies in that if two fuzzy sets, A_1 and A_2, are very close to each other, there will exist a point x for which its internal representation will have components $\mu_1(x)$ and $\mu_2(x)$ with about the same *high* value, and on the other hand, if A_1 and A_2 are far enough, there will be no such a point in the domain whose membership degrees are simultaneously high. This reasoning can be formalized in the following constraint:

$$M_p(x^{(k)}) - 1 \leqslant 0 \quad (k = 1, \ldots, N) \tag{25}$$

which gives rise to the operational expression (22).

As a matter of fact, most existing algorithms for generating distinguishable fuzzy sets focus on merging similar fuzzy sets by using similarity measures between fuzzy sets such as the following commonly used one [48,200] to quantify the distinguishability:

$$sim(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{26}$$

where A and B are two fuzzy sets, $|\cdot|$ and $\cap$ represent the cardinality of a set and the intersection, respectively. In this scheme, the two most similar fuzzy sets are merged into a new one and then the rule base is updated. This process is repeated until the similarity measure of the compatible fuzzy sets generated is greater than a given threshold. Finally,

sets that are close to being universal are removed. Various similarity measures can be found in the literature. A good comparative analysis about many similarity measures among fuzzy sets was made in [200].

Castellano et al. suggested a possibility measure to quantify distinguishability by replacing the similarity measure, aiming at reducing computational load during the fuzzy set merging process by similarity measure [119]. Furthermore, the possibility measure is justified as a way of evaluating distinguishability in [118] by proving that minimizing possibility also minimizes similarity and hence improves distinguishability. The possibility measure between two fuzzy sets A and B is defined as follows:

$$\Pi(A, B) = \sup_x(\min(\mu_A(x), \mu_B(x))) \tag{27}$$

This possibility measure evaluates the degree of applicability of the fuzzy constraint (A is B) [48] by quantifying the extent to which A and B overlap. Interestingly, Hefny derived two useful generalized formulas in [77] for both similarity and similarity/possibility relationships of Gaussian fuzzy sets with different widths.

By combining the De Luca and Termini's fuzzy entropy [42] with a deviation measure based on analogy of relative entropy, Suzuki et al. proposed a measure called *conciseness* of fuzzy models to distinguish the shapes of MFs [58,59], which is closely related to the distinguishability of MFs. The fuzzy entropy is defined by

$$d(A) = \int_{x_1}^{x_2} (-\mu_A(x) \ln \mu_A(x) - \mu_A(1-x) \ln \mu_A(1-x))\, \mathrm{d}x \tag{28}$$

where $\mu_A(x)$ is the MF of fuzzy set A, x_1 and x_2 are the left and right points of the support of fuzzy set A, respectively. The deviation measure is to evaluate the discrepancy of an MF from symmetry, which is defined by

$$r(A) = \int_{x_1}^{x_2} \left(\mu_C(x) \ln \frac{\mu_A(x)}{\mu_C(x)}\right) \mathrm{d}x \tag{29}$$

where $\mu_C(x)$ is the symmetrical MF of fuzzy set C, which has the same support as fuzzy set A. The following measure is used to measure the *conciseness* of fuzzy models by evaluating the shapes and allocations of fuzzy sets A_i ($i = 1, \ldots, S$):

$$\mathrm{d}r_{avr} = \frac{1}{N-2} \sum_{i=2}^{S-1} dr(A_i) \tag{30}$$

where

$$\mathrm{d}r(A) = \int_{x_1}^{x_2} (\mu_C(x) \ln \mu_A(x))\, \mathrm{d}x \tag{31}$$

In [167], some conflicting relationships between the conciseness and the accuracy of fuzzy models are further discussed.

However, most existing measures for quantifying the distinguishability of fuzzy sets used in fuzzy system modelling are only based on input data without using the information contained in the output data. In [198], Zhou and Gan proposed a so-called "local'' entropy to measure the distinguishability of fuzzy sets by taking into account the information provided by input–output sample pairs to optimize input space partitioning:

$$LE_K = -\sum_{k=1}^{K} \sum_{m=1}^{N} \widehat{U}_{km} \cdot MC(\beta_m) \log(\widehat{U}_{km} \cdot MC(\beta_m)) \tag{32}$$

where the data space $\{\beta_m | m = 1, \ldots, N\}$ is partitioned into ψ_k, $k = 1, 2, \ldots, K$,

$$U_{km} = \begin{cases} 1 & \text{if } \beta_m \in \psi_k \\ 0 & \text{otherwise} \end{cases} \tag{33}$$

and

$$\widehat{U}_{km} = U_{km} \Bigg/ \left(\sum_{m=1}^{N} U_{km} \cdot MC(\beta_m)\right) \tag{34}$$

where $MC(\beta_m)$ is the compactness index of fuzzy set A_m with core centre β_m. By maximizing this entropy measure the optimal number of merged fuzzy sets with good distinguishability can be obtained, which leads to a parsimonious input space partitioning while preserving the information of the original fuzzy sets as much as possible.

In [146], Rojas et al. proposed a method to decide the number of MFs in antecedent in a dynamic way based on the following controversy index for 0-order TS models, which is called the sum of controversies *CI* associated with a given MF:

$$SCMF(X_v^j) = \sum_{i \in R} CI(i) \tag{35}$$

where R represents the set of rules whose antecedent in variable X_v refers to the ith MF, $CI(i)$ defined at the rule level suggests the difference between the rule conclusion and the observed output for the data points that activate the corresponding rule. $CI(i)$ is evaluated as follows for the ith rule:

$$CI(i) = \sum_{k=1}^{N} ([(y_k - R_i) r_i(k)]^2)^{1/2} \tag{36}$$

where R_i is the ith rule conclusion, y_k is the kth sample observed output, $r_i(k)$ is the ith rule fire strength for the kth sample. Note that the criterion used here is not evaluated in an input space region, but at the rule level. We review this method here only because the method can dynamically choose the number of MFs during the fuzzy sets generation. Some methods performed on the rule level will be reviewed in the next section.

In data-driven fuzzy modelling, overfitting of the MFs may result in the incompleteness of fuzzy partitioning during parameter optimization. In order to enforce the interpretability originating from the coverage of space partition during the MFs optimization, de Oliveira proposed the following constraint expression for coverage in [45],

$$J_1 = \frac{1}{2} \sum_{k=1}^{N} (x^{(k)} - \widehat{x}^{(k)})^2 \tag{37}$$

where $x^{(k)}$ is the kth numeric sample and $\widehat{x}^{(k)}$ is calculated by any differentiable defuzzification method such as

$$\widehat{x}^{(k)} = \frac{\sum_i \mu_{A_i}(x^{(k)}) \beta_i}{\sum_i \mu_{A_i}(x^{(k)})} \tag{38}$$

where β_i is the core centre of the MF of fuzzy set A_i. The detailed analysis on how the expression (37) is allowed to implement the constraint of coverage can be found in [45].

However, the current efforts on formalizing the semantic constraints for fuzzy partitions mainly focus on the individual criteria for designing MFs. In interpretable fuzzy system modelling, a desirable fuzzy partition should be the one that satisfies all the above semantic criteria, which is called the standardized fuzzy partition.

Definition. A fuzzy partition in input space is called a standardized partition if it satisfies all the semantic criteria.

The objective of low-level interpretable fuzzy modelling is to find a good trade-off between fuzzy partition interpretability and global model performance, in which a standardized fuzzy partition is generated while keeping the global system performance at a satisfied level. Currently, in order to meet all the semantic criteria, how to partition the input space effectively is still a challenging problem.

4.2. Criteria for high-level fuzzy model interpretability

The interpretability of each rule and the whole rule base plays a key role in understanding a fuzzy system. Some researchers have suggested several individual syntactic constraints for fuzzy rules [49,65,67,91,92,139]. If these conditions are satisfied by the fuzzy rules, the interpretability of a fuzzy system would be enforced. We refer to them as the criteria for high-level interpretability of fuzzy systems.

4.2.1. Definitions of the criteria

4.2.1.1. Rule base parsimony and simplicity According to the principle of *Occam's razor* (the best model is the simplest one fitting the system behaviours well), the set of fuzzy rules must be as small as possible under the condition that the model performance is preserved at a satisfied level. Large rule base would lead to a lack of global understanding of the system. A parsimonious fuzzy system is very desirable when the number of input variables is large, especially for TS fuzzy systems with general consequent forms [92] and the parameters trained by an adaptive learning algorithm. The traditional statistical parsimonious system modelling techniques possess great potentials of being used to produce compact rule base, which will be reviewed in more detail in the next section.

4.2.1.2. Readability of single rule To improve readability, the number of conditions in the premise part of the rule should not exceed the limit of 7 ± 2 distinct conditions, which is the number of conceptual entities a human being can efficiently handle [139].

4.2.1.3. Consistency Rule base consistency means the absence of contradictory rules in rule base in the sense that rules with similar premise parts should have similar consequent parts [49,67,92].

4.2.1.4. Completeness For any possible input vector, at least one rule should be fired to prevent the fuzzy system from breaking inference [67]. However, in practice an input space partition with very low rule activation such as 0.00001 may deteriorate fuzzy model interpretability. A good choice is to set up a tolerance threshold for rule activation during fuzzy sets and rule generation to prevent rule base from being activated at a very low level.

4.2.1.5. Transparency of rule structure The criteria for high-level interpretability of fuzzy models evaluate the structure of fuzzy rules and their constitutions. A fuzzy rule should characterize human knowledge or system behaviours in a clear way.

Currently, two most widely used fuzzy rule structures are Mamdani rule and TS rule. As indicated above, the Mamdani rule and TS rule share the same premise structures, the only difference lies in their consequent parts. The consequent structure of a Mamdani rule is a fuzzy set, while the one of a TS rule is a linear real function. Because fuzzy sets can be used to express human perception knowledge, so it is accepted that Mamdani model interpretability due to its rule structure is a default property. However, a consequent variable expressed in terms of a real function does not exhibit clear physical meanings [92], so it seems that the Mamdani fuzzy system offers a more comprehensible way of characterizing system behaviours than the TS system, which may be the reason why most existing interpretable fuzzy system modelling techniques focus on Mamdani systems in different domains.

However, we argue that TS fuzzy rules characterize system behaviours in a different way, in which each rule represents a local linear model in nature, as shown in Fig. 4. Hence, TS fuzzy model interpretability due to its rule structure should be studied from the perspective of the interaction between global model and its local linear models, which is different from Mamdani model interpretability in this aspect. Indeed, some researchers have proposed to achieve the interpretable TS local models in the sense of the following definition [93,190].

Definition. The local models (rules) of a TS model are considered to be interpretable if they fit the global model well in their local regions, and result in fuzzy rule consequents that are local linearizations of the non1inear system.

According to this definition, interpretable local models of a TS model should dominate the system behaviours separately in their local regions, hence the TS local models in Fig. 4 (right) possess better local model interpretability than the ones in Fig. 4 (left). And the TS local models in Fig. 2 (right) show good local model interpretability as well.

4.2.2. Formalized expressions for the criteria in improving high-level fuzzy model interpretability

Some researchers have investigated the formalized expressions of the criteria in order to achieve a high-level interpretable fuzzy model.

The problem of consistency of fuzzy rules is usually thought to be trivial, which is believed to be insensitive to the inconsistency of fuzzy rules to a certain degree, if the rules are extracted from expert knowledge. However, in data-driven fuzzy modelling, this problem should not be ignored, as the rules may be automatically generated from noisy data. Seriously inconsistent fuzzy rules will undoubtedly result in performance deterioration and make the fuzzy

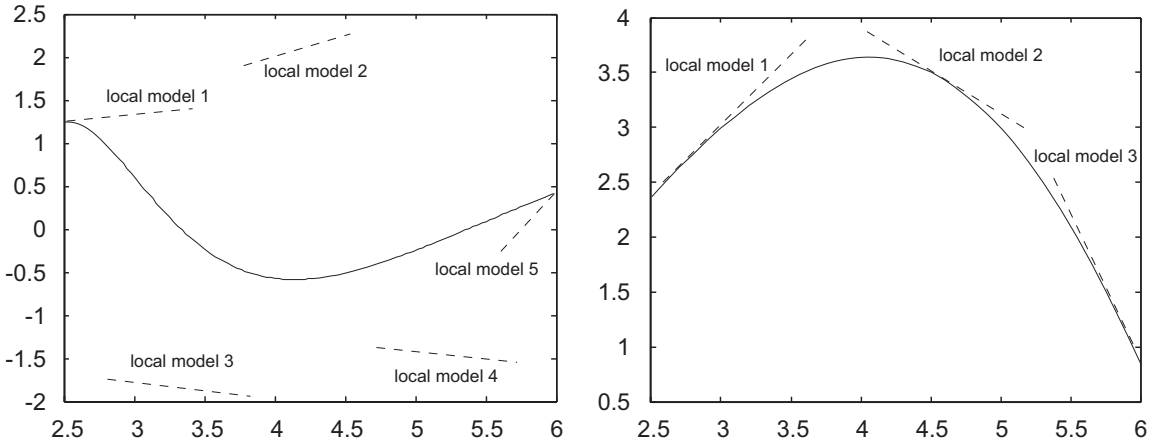


Fig. 4. TS models with uninterpretable local models (left) and interpretable local models (right): solid line represents the global model, dotted lines represent the local models.

system inexplicable. Therefore, the consistency of fuzzy rules has received increasing attention [49,65,91,92]. In [92], not only the consistency among fuzzy rules in the generated rule base is considered, but also the consistency of rules with the intuition and common sense of human beings is dealt with. And a proper definition of consistency is provided by Jin et al. in [92] to characterize a certain degree of inconsistency of fuzzy rules for Mamdani fuzzy models in the following senses:

- the rules have *very similar* premise parts, but possess *rather different consequents* and
- the rules conflict with the expert knowledge or heuristics.

The consistency of rules embracing these considerations is defined as follows:

$$Cons(R_i, R_k) = \exp\left\{ -\frac{\left(\frac{SRP(R_i, R_k)}{SRC(R_i, R_k)} - 1\right)^2}{\left(\frac{1}{SRP(R_i, R_k)}\right)^2} \right\} \tag{39}$$

where *SRC*, standing for *similarity of rule consequent*, is defined as

$$SRC(R_i, R_k) = S(B^i, B^k) \tag{40}$$

which is the similarity measure $S(\cdot, \cdot)$ of two fuzzy sets B^i and B^k in the following two rules:

$$R_i: \text{ If } x_1 \text{ is } A_1^i \text{ and } \cdots x_n \text{ is } A_n^i \text{ then } y \text{ is } B^i$$

$$R_k: \text{ If } x_1 \text{ is } A_1^k \text{ and } \cdots x_n \text{ is } A_n^k \text{ then } y \text{ is } B^k$$

and *SRP*, standing for *similarity of rule premise*, is defined as

$$SRP(R_i, R_k) = \min_{j=1}^{n} S(A_j^i, A_j^k) \tag{41}$$

In terms of definition (40), the degree of consistency tends to be high when *SRP* and *SRC* of two rules are in proportion, provided that the SRP of the two rules is high. Particularly, if the rules have the same premise and the same consequent, the degree of consistency reaches its highest value of 1. If the premises are the same but the consequents are different, then the consistency ranges from 0 to 1.0. Additionally, the degree of consistency is always high if the *SRP* of two rules is very low, no matter how the relation of *SRP* and *SRC* changes, which is concordant with the assumption that two rules will always be considered to be consistent if they have very different premises.

Obviously this definition of consistency of fuzzy rules is generally suitable for Mamdani fuzzy model and 0-order TS type fuzzy model. However, for general TS-type fuzzy models this definition is difficult to evaluate the consistency of two fuzzy rules.

Interestingly, in [190] Yen et al. proposed local error function as a way of evaluating the degree of TS local model dominating the behaviours of the global system. This local error function defined as follows can be used as a measure of the interaction between TS local models and global model:

$$J_{\mathrm{L}} = \sum_{i=1}^{L} \sum_{K}^{N} w_i(k) \left[y^{(k)} - y_i(k) \right]^2 \tag{42}$$

where $w_i(k)$ is the normalized firing strength of the ith rule given the kth sample. Because $w_i(k)$ has non-zero values only in a small region of the input space, as a result each fuzzy rule acts like an independent submodel that is only related to a subset of training data.

All in all, the criteria presented in this section intend to define a number of constraints on the fuzzy parameters and rule base for the sake of securing the low-level and high-level interpretability of fuzzy systems at fuzzy set level and fuzzy rule level separately. Semantic criteria limit the choice of MFs on fuzzy set level, while high-level interpretability criteria bind the fuzzy rule base on the rule level. It is pinpointed that among the semantic constraints for achieving low-level interpretability, distinguishability is essential for semantic integrity. As a basic criterion, distinguishability is the first concern in interpretable fuzzy modelling [45,68,132,135,156,180]. Correspondingly, parsimony is a basic criterion for achieving high-level interpretability due to the *Occam's razor* principle. In fact, the widely investigated transparent and interpretable models in traditional statistical system modelling also originate from this motivation. Currently, most interpretable fuzzy models were constructed in terms of individual semantic criterion or high-level interpretability criterion, and the distinguishability and parsimony are the most attention-getting criteria for achieving low-level interpretability and high-level interpretability separately.

5. Constructive techniques for fuzzy model interpretability

In this section, we review the state-of-the-art about how to achieve an interpretable fuzzy model from data with emphasis on the principles and common characteristics used behind individual methods.

5.1. Constructive techniques for low-level interpretability of fuzzy models

Data-driven transparent fuzzy modelling involves construction techniques for low-level interpretability of fuzzy sets, in connection with the optimization of MFs, and for high-level interpretability of fuzzy rules, which concentrates on the overall rule base optimization.

In this subsection, we review some techniques to obtain interpretable fuzzy models in terms of semantic criteria on input space partitioning. Four schemes for achieving interpretable MFs are summarized and reviewed: (1) antecedent parameter regularization by introducing semantic criteria in objective functions; (2) multi-objective optimization for antecedent parameter estimation; (3) fuzzy set merging and (4) user-oriented interactive modelling techniques.

5.1.1. Regularization techniques for parameter estimation

Regularization is a natural strategy to attack ill-posed problems of function approximation in data-driven modelling [51], in which a prior knowledge in the form of penalty functional is included. Depending on the penalty terms we classify the regularization techniques into two parts: one includes the classic approaches with smoothness constraints as penalty terms, and the other involves the newly developed methods imposing non-smoothness constraints in objective functions.

The concept of regularization was introduced by Tikhonov [171,172] in the study of Fredholm integral equations of the first kind. The main idea underlying Tikhonov regularization method is that the solution of an ill-posed problem can be obtained from a variational principle, which contains both approximation error and prior smoothness information. The Tikhonov-type regularization techniques with the main penalty on smoothness of the approximated functions have been broadly employed to solve ill-posed problems, particularly in signal and image processing. They may also improve the robustness of the construction algorithm in interpretability-oriented fuzzy modelling, eventually leading to more relevant (interpretable) parameter estimates [11].

Once the semantic constraints or requirements that characterize the semantic criteria, as addressed above for low-level transparency of fuzzy models, are formalized mathematically, there are potentials of penalizing them in objective

functions. de Oliveira suggested a formalized constraint expression for distinguishability and a mathematical expression for coverage separately, and then optimized the MFs by penalizing the two expressions (37) and (22) in objective functions as follows, aiming to generate highly distinguishable antecedent MFs and ensure the completeness of the fuzzy partition individually [45].

$$J_{\text{overall}} = J_{\text{G}} + \lambda_1 J_1 + \lambda_2 J_2 \tag{43}$$

where λ_1 and λ_2 are the regularization parameters and J_{G} is the commonly used global learning objective function

$$J_{\text{G}} = \frac{1}{2}\sum_{K}^{N} \| y^{(k)} - \widehat{y}^{(k)} \|^2 \tag{44}$$

in which $\widehat{y}^{(k)}$ is the output of fuzzy model, $y^{(k)}$is the desired output.

In [91], Jin proposed the following penalty term in the objective learning function to merge similar fuzzy MFs during adaptation of MFs' parameters in order to generate distinguishable fuzzy sets:

$$\Omega = \frac{1}{2}\left(\sum_{i=1}^{n}\sum_{k}^{m_i}\sum_{A_{ij}\in U_{ik}} (a_{ij} - \bar{a}_{ik})^2 + \sum_{i=1}^{n}\sum_{k}^{m_i}\sum_{A_{ij}\in U_{ik}} (b_{ij} - \bar{b}_{ik})^2 \right) \tag{45}$$

where $\bar{a}_{ik}$, $\bar{b}_{ik}$ are the two parameters in the Gaussian function shared by all the fuzzy subsets in group U_{ik}, so that the overall objective learning function becomes

$$J_{\text{overall}} = J_{\text{G}} + \lambda\Omega \tag{46}$$

in which λ is the regularization parameter.

However, the current efforts of using regularization techniques to improve low-level interpretability of fuzzy models only focus on the objectives of distinguishability and completeness of input space partitions and the global system performance. In practice, given the formalized constraint expressions for all the other low-level interpretability criteria, one can also employ the regularization technique to pursue the good trade-off between global model performance and low-level fuzzy interpretability, so that the standardized fuzzy partition can be induced.

5.1.2. Multi-objective optimization for antecedent parameter estimation

In addition to the regularization techniques for antecedent parameter learning, another possibility of considering semantic criteria in fuzzy modelling is to combine semantic constraints and model accuracy measures in a multi-criteria function. Roubos and Setnes proposed a method of genetic multi-criteria optimization for building a compact and transparent fuzzy model [147]. To reduce the model complexity, the accuracy objective is combined with a similarity measure of fuzzy sets in a genetic algorithm (GA) objective function as follows:

$$J_{\text{GA}} = (1 + \lambda S*)J* \tag{47}$$

where $J*$ is to measure the model accuracy either for system approximation or for pattern classification. $S*$ is the average of the maximum pair-wise similarity that is present in each input, defined by

$$S* = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\max(S(A_{ij}, A_{ik}))}{\eta_i - 1}\right) \quad (j, k \in \{1, 2, \ldots, \eta_i\},\ j \neq k) \tag{48}$$

where n is the number of inputs and η_i the number of fuzzy sets A_{ij} for each input variable. Similarity is rewarded during the iterative process, and the redundancy is used to remove unnecessary fuzzy sets in the next iteration. Finally, fine-tuning is combined with a penalty for similar fuzzy sets in order to obtain a distinguishable term for linguistic interpretation. In [146], Rojas et al. used the controversy index *SCMF* defined in (35) to decide in a dynamic way

where it is necessary to assign a larger number or density of rules and to increase the density of MFs in a specific input variable for an accurate approximation of the target function. As a result, a fuzzy system is constructed with a good compromise between the accuracy of the approximation and the complexity of the rule set. A new MF is added for variables whose controversy index variance is high [146]. By using the triangular partitioning, the new centre location is computed as

$$c_v^* = \frac{\sum_{j=1}^{n_v} c_v^j SCMF(X_v^j)}{\sum_{j=1}^{n_v} SCMF(X_v^j)} \tag{49}$$

In [180], Wang et al. proposed a scheme of extracting interpretable rule-based knowledge based on multi-objective hierarchical GA, in which the genes of the chromosome are arranged into control genes and parameter genes, and the similarity measure (26) based fuzzy set merging method was used to remove the redundancy of input space partition. To develop interpretable fuzzy models, Anderson [8], Paiva and Dourado [132], Lotfi et al. [113], Kumar et al. [108] defined a set of constraints on the parameters of fuzzy inference systems for interpretation preservation by methods of loosely limiting the location of the MFs during learning. As a result, the completeness of the fuzzy partitioning will be kept and the distinguishability of different fuzzy sets will be preserved.

However, the current efforts of using multi-objective optimization techniques to improve low-level interpretability of fuzzy models only focus on the objectives of distinguishability of input space partitions and the global system performance. As a matter of fact, given the formalized constraint expressions for all the other low-level interpretability criteria, the multi-objective optimization technique is a natural choice for one to get the standardized fuzzy partition by seeking the good trade-off between global model performance and low-level fuzzy interpretability. The Pareto-optimal solutions of the underlying multi-objective optimization problem are used to essentially determine the trade-off between the possibly conflicting objectives of global model accuracy and local model interpretation.

5.1.3. Fuzzy set merging techniques

In order to produce distinguishable input space partitioning for fuzzy modelling, to merge fuzzy sets is a natural choice and has been well applied. Setnes et al. proposed a method to add or merge fuzzy subsets based on similarity measures indicating the degree to which two fuzzy sets are equal [156]. In [52], Espinosa and Vandewalle proposed an algorithm, named FuZion, to merge MFs whose cores are "too close'' to each other. Castellano et al. suggested an agglomerative approach-based double-clustering technique to generate distinguishable fuzzy granules [27]. Jimenez et al. used a merging–splitting process to improve the transparency and compactness of fuzzy models [90]. Based on a multi-objective optimization method, Wang et al. presented a fuzzy set agent-based evolutionary approach to extract interpretable fuzzy rules, in which an interpretability-based regulation scheme for merging similar fuzzy sets was used [179].

However, most existing merging procedures to generate distinguishable input space partitioning are only based on input data without using the information contained in the output data. It is known that for system identification or modelling, the output data contains useful information for input space partitioning. Hence, a more attractive and reasonable merging method should take into account the information provided by input–output sample data pairs during the fuzzy set merging process, which is the motivation of the research work in [198].

5.1.4. User-oriented interactive modelling techniques

It is known that how to interpret a fuzzy model heavily depends on human's prior knowledge [58,59,198], so it is a highly subjective task. For the sake of inducing an interpretable fuzzy model from data, one sensible way may be to allow users to get involved in the model generation process interactively so that the final system is obtained in terms of subjective human evaluation. However, the majority of current efforts on fuzzy modelling do not allow users to get involved in the generation process, they produce only one result no matter whether the users like or not.

Interestingly, Nauck and Kruse took a view of neuro-fuzzy models as a way of heuristically learning parameters of fuzzy systems from data, and proposed two neuro-fuzzy models NEFCLASS [125,127] and NEFPROX [126], which are not automatic "fuzzy system creators''. The users are allowed to supervise and interpret the learning procedure in all stages. The NEFCLASS model is used for classification, whilst the NEFPROX model for function approximation and regression.

Table 2
Techniques for low-level interpretability of fuzzy models

Techniques	Low-level interpretability criteria					Comments
	Distinguishability	Moderate number of MFs	Completeness	Normalization	Complementarity	
Regularization	Yes	Yes	Yes	Yes	Yes	• Difficult to select optimal regularization parameter • High computing overhead
Multi-objective optimization	Yes	Yes	Yes	Yes	Yes	• Difficult to select optimal multi-objective coefficients • High computing overhead
Fuzzy set merging	Yes	No	No	No	No	• High computing overhead
User-oriented interactive technique	Yes	Yes	No	No	No	• MFs are limited • May take long time to get an acceptable model

The main features of NEFCLASS and NEFPROX models are the shared weights on some of connections of 3-layer networks used, which guarantee that for each linguistic value there is only one representation as a fuzzy set. To keep fuzzy model interpretability, NEFCLASS allows to impose different restrictions on learning process, such as MFs must not pass its neighbours, must stay symmetrical, have a certain degree of overlapping, etc. Among the pruning strategies offered by NEFCLASS, the users can interactively delete antecedents if the degree of fulfilment of the fuzzy sets in antecedent is not identical to the rule's degree of fulfilment at least z percent (the user can specify the value of z), and delete fuzzy sets if they cover more percent of domain than the percentage value specified by the users. However, the NEFCLASS procedures of deleting antecedents and fuzzy sets can lead to inconsistent rule base [125,127]. More issues arising in NEFCLASS/NEFPROX models include: the learning algorithm is only suitable for triangular MFs including trapezoid functions [126]; Because the users are offered many choices of manipulating the system in all stages, they may take much more time to yield a fuzzy model with acceptable model accuracy and interpretability by trying different strategies and combinations and picking a "best'' result.

The characteristics of the existing construction techniques for low-level interpretability of fuzzy models are summarized in Table 2, in which some drawbacks of these techniques are indicated, whilst the advantages lie in their abilities to achieve low-level interpretable fuzzy models by considering corresponding criteria in modelling process.

5.2. Constructive techniques for high-level interpretability of fuzzy models

In this subsection, techniques for securing high-level interpretability are to be reviewed, some of which directly originate from traditional statistical system modelling.

5.2.1. ANOVA decomposition

In traditional statistical system modelling, one attractive way of ameliorating the curse of dimensionality is to model the regression function as an additive function of predictors [164,177,178]. Technically, the ANOVA decomposition can be constructed explicitly by using the tensor product spline technique based on the construction of tensor product reproducing kernel Hilbert spaces [9].

When applied to fuzzy modelling, the individual subfunctions of the ANOVA representation can be modelled by the conventional fuzzy model, which makes a fuzzy system composed of a series of simpler submodels. The outputs of these submodels are fused together to produce the overall output. Accordingly (13) can be reformed

as follows:

$$\widehat{y} = y_0 + \sum_{s=1}^{S} \widehat{y}_s(x^s) \tag{50}$$

$$\widehat{y}(x^s) = \sum_{k=1}^{K_s} w_k^{\mathrm{T}}(x^s)\theta_k^s \tag{51}$$

where x^s represents a subspace, S is the number of submodels and K_s is the number of rules in the s-th submodel. As a matter of fact, the ANOVA decomposition of a function provides a kind of so-called "*overcomplete*'' bases (also called *overcomplete dictionaries* in signal processing community). Unlike the case of a complete basis, where signal decomposition is well defined and unique, finding the "best'' representation in terms of an overcomplete basis is a challenging problem [110]. Nevertheless, this kind of decomposition offers some outstanding advantages. One is that there is greater flexibility in capturing structure in the data and it can form more compact representation, because each basis function can describe a significant amount of structure in the data. An additional advantage is that the stability of the representation increases in response to small perturbations in presence of noise. Specifically, the great advantage of the additive representation for fuzzy modelling is that the resulting rules within the submodels possess the antecedents with a smaller subsets of the input variables, making them simpler and more interpretable, in contrast to the majority of neural network models, providing the modeller with structural knowledge that can be used for both validation and model interpretation [21]. Thus, the interpretation preservation during parameter adaptation in the original TS model is reduced to the problems of subset (basis functions) selection and least square estimation in its ANOVA representation (50) and (51). It is worthwhile to note that this interpretation preservation is performed on a fuzzy rule level, since as stated above each column of W corresponds to one of the fuzzy rules such that the operations on parameter θ aim at fuzzy rules rather than individual MFs.

In order to produce parsimonious fuzzy models, the ASMOD (adaptive spline modelling of observational data) algorithm was developed [100] for constructing 0-order TS fuzzy models with ANOVA decomposition, and the MASMOD algorithm, a modified version of ASMOD, was proposed in [60] to construct 1-order TS fuzzy models based on ANOVA decomposition. Furthermore, in [61], Gan and Harris suggested a hybrid learning scheme combining the expectation-maximization (EM) algorithm, which is a technique for maximum likelihood or maximum *a posteriori* estimation [43], and the MASMOD algorithm for fuzzy local linearization modelling. The ANOVA decomposition-based construction algorithms produce parsimonious models by adding submodels on lower-dimensional spaces into the model which starts from an empty model. This submodel addition process is controlled by some complexity penalty criteria such as structural risk minimization [159]. However, although deploying additive submodels shows promising performances in advanced flexible non-linear modelling methods, particularly for ameliorating the curse of dimensionality [75,97,163,164,177], a problem may happen in some cases for interpretable fuzzy modelling, i.e., they cannot provide a transparent solution if the object being modelled contains high-dimensional interactions [69].

5.2.2. Feature selection

The goal of feature selection in statistical model identification is to find an optimal trade-off between fitness and complexity, thus, to produce a compact and concise model structure with good generalization and interpretation. Because fuzzy system modelling possesses its own idiosyncrasy, particularly in the aspect of interpretability, variable selection should receive special attention in fuzzy system modelling. An important part of model interpretation lies in the evaluation of the effectiveness of the input variables (features) on the decision process. Parsimony of a fuzzy rule base is not only related to the number of rules or the avoidance of redundant rules, but also concerned with the number of fuzzy sets used in the rule premise parts. The latter issue depends on the number of input variables used in a fuzzy model. As a matter of fact, in fuzzy system modelling, variable selection impacting the number of fuzzy rules in rule base can be achieved in a global or a local way. In the global case, all rules cannot use a removed variable. In the local case, the variable selection is conducted based on the contribution of each rule, which may lead to incomplete rules. Currently some researchers have either extended the classical variable information criteria to fuzzy model construction or developed some special criteria.

5.2.2.1. AIC, BIC and HQ statistical criteria Among the model selection criteria, statistical information criteria provide a simple method to choose from a range of competing models. In statistical system modelling, the goodness of the model can be evaluated by the "similarity'' between the selected model with the associated density function $f(Y|\theta)$and the true (unknown) model with the density function $f_0(Y)$that generate the data, where $Y = (Y_1, \ldots, Y_N)$ is a random vector, and θ is a finite-dimensional vector of unknown parameters. One method of evaluating such a similarity is to use *Kullback–Leibler distance* [107],

$$\begin{aligned} I(f, f_0) &= \int \log[f_0(y)/f(y|\theta)]\, f_0(y)\, \mathrm{d}y \\ &= E\{\log[f_0(y)/f(y|\theta)]\} \\ &= E\{\log[f_0(y)]\} - E\{\log[f(y|\theta)]\} \end{aligned} \tag{52}$$

where $E\{\log[f(y|\theta)]\}$, which is called *expected log-likelihood function*, is usually unknown and needs to be estimated from a sample of N observations. The so-called information criterion is an "estimator'' of $E\{\log[f(y|\theta)]\}$, hence the different information criteria take the following general form:

$$\widehat{IC}(m) = -\frac{1}{N}\log(f(y|\theta)) + \alpha(N)\, m \tag{53}$$

where m is the dimension of θ, $\alpha(N)$ is a positive decreasing function of the sample size and satisfies the condition $\lim_{N\to\infty} \alpha(N) = 0$. The optimal m can be obtained by minimizing $\widehat{IC}$. In the case where $f(Y|\theta)$ is a normal density, $\widehat{IC}$ takes the following equivalent expression:

$$IC(m) = \log(\widehat{\sigma}^2) + m\alpha(N) \tag{54}$$

where $\widehat{\sigma}$ is the estimate variance of model residuals. As a matter of fact, different criteria are obtained by selecting different functions $\alpha(N)$. The most important criteria for optimal model construction to estimate the unknown parameters are: *Akaike information criterion* (*AIC*) [2], *Bayesian information criterion* (*BIC*) [145,153], and *HQ criterion* [72].

In AIC, the function $\alpha(N)$ is chosen as follows:

$$\alpha(N) = \frac{2}{N} \tag{55}$$

Clearly, AIC reflects a balance between the goodness of fit and the complexity of a model; it ingeniously incorporates two information resources: the variance of the residuals and the number of parameters of the model. It is theoretically shown that AIC gives an asymptotic unbiased estimate of the generalization error. However, it does not work well with small samples and has a substantial probability of overestimating the true order of a model, that is, it tends to identify m with a large value [17,72,182]. Bhansali and Downham suggested to use a varying parameter instead of a constant 2 in AICs $\alpha(N)$ form [17], for example,

$$\alpha(N) = \frac{\beta}{N} \tag{56}$$

The BIC chooses the $\alpha(N)$ as following form:

$$\alpha(N) = \frac{\log(N)}{N} \tag{57}$$

It is possible to show that BIC can produce a more parsimonious model than AIC [169].

The HQ criterion uses a more complex $\alpha(N)$ as follows:

$$\alpha(N) = \frac{c\log(\log(N))}{N} \quad (c > 2) \tag{58}$$

Many other information criteria that are similar to AIC have also been developed [63,71]. A common feature of these criteria is that they all pursue a balance between the goodness of fit and the model complexity, and all these information criteria can be examined under a general framework in (53) and (54).

In [189], Yen and Wang extended the above statistical information criteria to fuzzy model construction and proposed an approach to exploring the fitness-complexity trade-off by using these optimality criteria together with a fuzzy model reduction technique based on singular value decomposition (SVD). The advantage lies in that the number of rules necessary for a compact rule base can be determined automatically.

5.2.2.2. Regularity criteria Many researchers have also applied traditional variable selection method to fuzzy model construction. Sugeno and Yasukawa extended the classic cross validation procedure to fuzzy modelling for variable selection [166]. The criterion is defined as

$$RC = \frac{1}{2}\left(\frac{1}{K_A}\sum_{i=1}^{K_A}(y_i^A - y_i^{AB})^2 + \frac{1}{K_B}\sum_{i=1}^{K_B}(y_i^B - y_i^{BA})^2\right) \tag{59}$$

where A and B are two groups into which the training set is divided randomly, y_i^A (respectively, y_i^B) is the observed output for pair i of group A(respectively, B), and y_i^{AB} (respectively, y_i^{BA}) is the inferred output, for pair i of group A (respectively, B) after training using group B (respectively, A) sample.

5.2.2.3. Other statistical criteria Takagi and Sugeno [168] proposed a heuristic method to search the number of fuzzy partitions and fuzzy rules in which the process termination criteria comprises either the *mean-squared error* (MSE) (if the MSE becomes less than a threshold) or the threshold for the number of fuzzy rules (if the number exceeds the threshold). However, this method is difficult to find an optimal solution. In [187], Yager and Filev proposed an *entropy criterion* to produce compact model structures. An *unbiasedness criterion* used for fuzzy modelling by Sugeno and Kang [165] was employed in this entropy criterion to determine the number of fuzzy rules. The unbiasedness criterion can effectively find a model with a low error by cross-validating two sets of training data, but it needs to consider enough different model structures [189]. Another entropy-based variable evaluation index is proposed by Pal to attack classification problems [133]. In the method, a variable which separates one class from all the others will not be assigned a high entropy value, and the most discriminant variable for the two classes is the one that minimizes the variable evaluation index.

In order to quantify the discriminative power of the input features (variables) in a fuzzy model, Silipo and Berthold suggested a measure of the information available in the system based on the separability among all the rules of a fuzzy model [161]. Feature selection can be conducted in terms of the information gains of different variables. In [101], Kim et al. proposed a method of rule selection by considering the correlation among components of input data based on principal component analysis (PCA) technique. By the PCA method, input data matrix of (possibly highly) correlated variables is transformed to an orthogonal system based on the eigenvectors of the feature covariance matrix, as a result, some of the transformed features are selected from the most significant axes in order. The rules are generated in eigenvector space in which each eigenvector is a linear combination of the input variables. One issue arising in PCA-based method is that the fuzzy rules generated in eigenvector space may suffer from loss of direct linguistic interpretability of the system.

5.2.3. Grid partitioning scheme

Grid partitioning is to decompose the input space of a fuzzy system based on axis-orthogonal splits in a heuristic way. It defines a number of fuzzy sets for each variable, which are interpreted as linguistic labels and shared by all the rules. Generally, a grid partitioning is produced by dividing each variable domain into a given number of intervals without consideration of a data density repartition function, and there is no need for these intervals to have any physical meaning. A training procedure is used to optimize the grid structure, as well as the rule consequences, according to given data samples [67]. In [128–130], an iterative heuristic search algorithm, called local linear model tree (LOLIMOT), is presented and used to decompose the input space into hyperrectangles by introducing a flexible k-d tree space structure. In each iteration, a local error measure is used to check every region, and the worst one is partitioned into two halves. Finally, the split with the highest performance improvement is chosen, in which the structural risk minimization metric is used to determine the most appropriate model structure and the number of required splits automatically. An advantage of LOLIMOT is its search effectiveness, particularly for high-dimensional data. However, the tree building procedure in LOLIMOT is suboptimal (greedy). As ANFIS does [87], LOLIMOT makes no attempt to eliminate redundant rules.

Hence, the number of rules generated can be quite large, and the outcome is that the rules often have a lot of non-uniform overlapping MFs to which it is hard to assign understandable linguistic terms.

In general, grid partitioning approaches have the advantage that they can easily yield a higher degree of transparency in the sense that all the fuzzy rules are expressed with a relatively small number of linguistic terms and these linguistic terms are shared in all rules, which is helpful to the readability of single rule. The problem is that in the k^n possible antecedents (where each input variable is characterized by k linguistic terms in n-dimensional space), there exist only $k \times n$ linguistic terms that are needed to produce. Hence, grid partitioning approaches may produce a large number of small identical rule antecedents, thus the comprehensibility of the fuzzy system would be lost at some point. It becomes even more serious for high-dimensional data [56]. Indeed the curse of dimensionality prevents the use of the methods which generate all the possible rules in high-dimensional space.

5.2.4. Prototype/scatter-partitioning scheme

One mechanism that can eliminate the problems associated with grid-partitioning is the decomposition of the input space in scatter partitioning way to get a set of points in feature space, also called prototypes, which results in finding only the relevant rules. Each prototype is associated with one fuzzy rule. The prototypes may be vectors of the same dimension as the data points, but they can also be "higher-level'' geometrical objects, such as linear or nonlinear subspaces or functions. Unlike grid-partitioning, scatter-partitioning positions small patches (corresponding to the antecedents of fuzzy rules) at locations which are unknown *a priori*. Thus, in scatter partitioning, the antecedent of the fuzzy rules is allowed to be positioned at arbitrary locations of the input space.

One outstanding advantage of scatter-partitioning fuzzy systems over grid-based fuzzy ones lies in that the scatter-based fuzzy systems can overcome the curse of dimensionality, leading to highly compact rule bases. So scatter-based fuzzy systems possess good high-level interpretability in terms of parsimony of rule bases. There are three fundamental issues needed to be addressed in designing a scatter-based system [109]: (i) How many prototypes are to be generated; (ii) How to generate the prototypes and (iii) How to use the prototypes to design a system model. Currently, in most efforts made to design prototype-based fuzzy systems, these three issues are addressed independently and separately. For example, unsupervised clustering algorithms such as fuzzy c-means (FCM) [14,50] and fuzzy learning vector quantization (FLVQ) [16], are widely used to generate prototypes, but most of the clustering algorithms require the number of clusters (prototypes) to be supplied externally or to be determined by using some cluster validity indices. Once the prototypes are generated, there are different ways of using the prototypes to design the classifier. One commonly used strategy is that these currently generated prototypes are used as initial fuzzy partitions, and an adaptive learning algorithm such as neural network learning algorithm is then applied to update these prototypes. Finally, based on training data set, the adaptive prototype-based classifier is trained optimally with good generalization performance on test data set. Recently, new efforts have been made to develop prototype-based models by integrating the above issues into one modelling process. Subtractive clustering [35] is a fast one-pass algorithm for estimating the number of clusters and the cluster centres in a set of data given parameter setting. Promisingly, the SVM-based fuzzy models [32,199] can fulfil the integration of all the three issues together in one modelling process. On the other hand, in order to overcome the curse of dimensionality in neuro-fuzzy system modelling, some researchers [57,79] proposed new ways of partitioning input space based on computational geometry theory, such as Voronoi tessellation and Delaunay triangulations, leading to parsimonious rule bases.

However, one common disadvantage of prototype-based fuzzy models is that because each prototype is characterized by a multi-dimensional fuzzy set, it is hard to render the multi-dimensional MFs meaningful linguistic terms, so fuzzy rules with multi-dimensional fuzzy sets in premise parts possess poor readability. One usually obtains the MFs of each input variable by projecting the multi-dimensional fuzzy sets on that input variable space, but every MF obtained in this way is now only used by a single rule so that the number of different MFs for one input variable may be as high as the total number of fuzzy rules. Therefore, assigning meaningful linguistic labels to these MFs is impossible in all but trivial cases. A problem to be solved with scatter partitions is to find a suitable number of rules and suitable positions and width of the rule patches in input space. In summary, although prototype-based fuzzy models can achieve high-level interpretability in terms of parsimony of rule bases and escape the curse of dimensionality, they may suffer from poor rule readability.

5.2.4.1. FCM clustering FCM unsupervised clustering is the most popular method of generating the prototypes for fuzzy systems. This prototype-based fuzzy algorithm is to partition a given finite data set $X = \{x_1, \ldots, x_N\}$ into

c fuzzy prototypes by minimizing the following objective function:

$$J_m(U, V) = \sum_{k=1}^{c} \sum_{j=1}^{N} (u_{kj})^m \|x_j - V_k\|^2 \tag{60}$$

s.t.

$$0 \leqslant u_{kj} \leqslant 1; \quad 0 < \sum_{j=1}^{N} u_{kj} < N, \ \forall k; \quad \sum_{k=1}^{c} u_{kj} = 1, \ \forall j \tag{61}$$

where U is a fuzzy c-partition of the given data samples, V_k are the cluster centres (prototypes), m (>1) is a weighting exponent. This constraint optimization problem leads to the following solution:

$$\left.\begin{aligned} u_{kj} &= \frac{(d_{kj}^2)^{1/(1-m)}}{\sum_{k=1}^{c}(d_{kj}^2)^{1/(1-m)}} \text{ if } I_j = \emptyset \\ \left.\begin{aligned} u_{kj} &= 0, \quad k \notin I_j \\ \sum_{k \in I_j}^{u_{kj}} &= 1, \ k \in I_j \end{aligned}\right\} &\text{ if } I_j \neq \emptyset \end{aligned}\right\} \tag{62}$$

where $d_{kj}^2 = \|x_j - V_k\|^2$, $I_j = \{k | 1 \leqslant k \leqslant c, d_{kj}^2 = 0\}$. One drawback of the prototype-based FCM algorithm is that it is highly sensitive to noise [40]. One may think that if the membership values of noise points in all prototypes are small then their influence can be reduced. However, the MF derived in (62) indicates that the membership degrees generated are relative numbers due to the constraint of the partition unity. As a result, noise points or outliers will also get high membership values, thus they can severely impact the estimation of the prototypes. Another issue arising in the FCM is that it tends to generate clusters with hyper-spherical shape and approximately same size regardless of the natural cluster structures within the date set due to Euclidian distance used. Some efforts have been made to avoid implicit assumptions on cluster shapes, such as by using the Mahalanobis distance in clustering [14,70], kernel method [197], defining new cluster shapes [15,78] or using probability relaxation procedure [89]. Hence, the common and reasonable way of applying FCM technique to fuzzy modelling is to initially partition the input space, then other fuzzy modelling methods are used to optimize this input space partitioning with various objectives.

5.2.4.2. FLVQ clustering FLVQ [16] is a popular batch clustering algorithm. The design of FLVQ is to improve performance and usability of Kohonen's on-line vector quantization and SOM algorithms [106] by combining Kohonen's on-line weight adaptation rule with the fuzzy set MF proposed in the FCM algorithm.

The classic learning vector quantization (LVQ) aims at minimizing an objective function that places all of its emphasis on the winning prototype for each data point. However, information due to data point x is carried by *all* of the c distances $\{\|x_j - V_k\|\}$. FLVQ allows all c quantizers to be updated from data during each updating epoch, thereby eliminates the need to define an update neighbourhood in LVQ. Unlike SOM using topological neighbours belonging to an output lattice, FLVQ employs metrical neighbours in the input space. In the adaptation step, FLVQ uses the absolute distances of reference vectors with respect to the current input, while SOM employs the neighbourhood ranking of the processing units within the external lattice.

Descending FLVQ employs a weighting exponent $m = m(t)$ monotonically decreasing with processing time t, where t is the number of processing epochs. It takes into account winner and non-winner information to compute the membership degree of an input pattern belonging to a cluster centre:

$$u_{kj}(t) = \left(\frac{1}{d(x_j(t), V_k(t))^2}\right)^{1/(m(t)-1)} \Bigg/ \sum_{i=1}^{c} \left(\frac{1}{d(x_j(t), V_i(t))^2}\right)^{1/(m(t)-1)} \tag{63}$$

where $k = 1, \ldots, c, k = 1, \ldots, N, x_j(t)$ is the input pattern to be processed by FLVQ at time t, and $V_k(t)$ represents the cluster centre generated at time t. It can be seen that the fuzzy subset with the MFs defined by (63) are a fuzz-partition

of input space. FLVQ updates the cluster templates as follows:

$$\begin{aligned} V_i(t+1) &= V_i(t) + \eta_i(t) \sum_{j=1}^{N} w_{i,j}(t)(x_j(t) - V_i(t)) \\ &= V_i(t) + \sum_{j=1}^{N} \alpha_{i,j}(t)(x_j(t) - V_i(t)) \\ &= \sum_{j=1}^{N} \alpha_{i,j}(t) x_j(t) \end{aligned} \tag{64}$$

where $i = 1, \ldots, c$,

$$w_{i,j}(t) = \left(u_{i,j}(t)\right)^{m(t)} \tag{65}$$

$$\eta_i(t) = \frac{1}{\sum_{j=1}^{N} w_{i,j}(t)} \tag{66}$$

$$\alpha_{i,j}(t) = \eta_i(t) w_{i,j}(t) \tag{67}$$

It can be observed that if $d(x_j(t), V_k(t)) \to 0$, then $u_{kj} \to 1/c$, while the MFs defined by (63) are relative numbers or probabilistic function. As a result, FLVQ share the same problem of high sensitivity to noise points and outliers with FCM in estimating the prototypes. In fact, according to Eq. (64), if the weighting exponent $m = m(t)$ is fixed, then FLVQ is equivalent to FCM. Because the general FLVQ does not minimize any known objective function, termination is not based on optimizing any model of the process or its data. However, the advantage of FLVQ is that, owing to its soft competitive implementation, FLVQ is expected to be less likely trapped in local minima and less likely to generate dead units than hard competitive alternatives. Moreover, FLVQ employs a smaller set of user defined parameters than SOM, and the batch learning scheme makes the final weight vectors generated by FLVQ not affected by the order of the input sequence when its traditional termination criterion is removed.

5.2.4.3. Subtractive clustering Different from FCM and FLVQ clustering algorithms, subtractive clustering is a potential function-based method. The potential function approach is a classical approach to design of automatic pattern classification systems [173]. In this approach, the feature points (i.e., potential centres) are interpreted as energy sources. A potential function is defined for each feature point which requires that the potential function has a peak value at the location of the feature point and decreases rapidly at any point away from the feature point. Yager and Filev [188] have proposed an unsupervised clustering method based on potential function (it is called "mountain function'' in [188]). However, this mountain clustering method requires setting up a grid in feature space for searching for the peaks of the mountain function directly. It becomes impractical for higher dimensions due to curse of dimensionality. Chiu proposed an improved implementation through a simple but significant change, which led to the subtractive clustering approach [35,36]. In this approach the centre candidates are the data samples themselves, rough estimates of the local population densities are used to determine cluster centres in a serial fashion. Subtractive clustering computes the total potential $P(x_i)$ at a data point x_i as

$$P(x_i) = \sum_{j=1}^{N} \exp(-\alpha \|x_i - x_j\|^2) \tag{68}$$

where α is a user defined resolution parameter. After the potential for each input point is computed, the data point with the highest potential is considered to be the first cluster centre. Let x_1^* be the first prototype and $P(x_1^*)$ denote the corresponding potential of x_1^*. In the next step, the potential of all the remaining points is reduced by removing the contribution of this cluster centre from the data set according to

$$P(x_i) \leftarrow P(x_i) - P(x_1^*) \exp(-\beta \|x_i - x_1^*\|^2) \tag{69}$$

where β is another user defined resolution parameter. At this step, the current highest density location is found. This procedure of centre selection and potential reduction is repeated until a certain termination criterion is met.

For a small size of data set, subtractive clustering is very fast compared with FCM, but it is no longer the case as the number of data samples increases [40]. Because the potential function used is defined as accumulation potentials of individual points, so noise points will not seriously affect the peak locations of the potential function. As a result, subtractive clustering is quite robust against noise points for roughly hyper-spherical clusters, if suitable α and β are used. Another outstanding advantage of this approach over other unsupervised clustering methods is that it does not need user to specify the number of clusters in advance as this approach estimates the clusters and the number of clusters in one model structure. However, the clustering result by the subtractive method is highly sensitive to the choices of resolution parameters α and β.

5.2.4.4. Growing neural gas Based on the scatter-partitioning mechanism, growing radial basis function networks (RBFNs) have been proposed in [56,99], in which localized BFs are selectively positioned at some locations in input space. Because the overall architecture of a 0-order TS fuzzy system with Gaussian MFs is equivalent to an RBFN [85], Fritzke extended it to neuro-fuzzy system modelling and presented the scatter-partitioning fuzzy systems with supervised growing neural network [57]. The goal of this method is to minimize the mean quantization error for a given data set $D_U = \{x_i \in \Re^n\}_{i=1}^N$ during distributing a limited number of centres w_i in the n-dimensional input space. This method can be described as follows:

- A soft competitive learning scheme is used to adapt centres on-line, in which for each presented input point the *winner* (the centre nearest to the input pattern) is moved towards the input pattern:

$$V_s := V_s + \varepsilon(x - V_s) \tag{70}$$

  To a smaller degree also the direct neighbours within the network topology are moved towards the input signal.
- At each adaptation step the squared distance between current input pattern and winner is added to a local error variable of the winner:

$$E_s := E_s + \|x - V_s\|^2 \tag{71}$$

- All error variables undergo slow exponential decay to eventually forget earlier accumulated errors.
- Among the centres a topology consisting of neighbourhood edges is created through *competitive Hebbian learning*. The principle of this method is to always insert an edge between the winner and the second nearest centre for the current input signal. If such an edge does already exist, its age parameter is set to zero. Edges with an age larger than a maximum value are removed again.
- A new centre is always inserted after a fixed number of input signals have been processed. It is positioned at the centre of one of the edges emanating from the centre with the largest accumulated error. The error values are reduced and redistributed locally after each insertion.

It is known that competitive Hebbian learning is an effective method of generating a subgraph of the Voronoi tessellation (corresponding to the current set of centres), and the Voronoi tessellation is the best for function interpolation among all triangulations [131], hence the growing neural gas method can generate Voronoi input space partition. As FLVQ does, owing to its soft competitive implementation, the growing neural gas method is less likely to be trapped in local minima and to generate dead units than hard competitive alternatives. In particular, due to its insertion strategy, this method is robust against noise by using accumulated errors. However, it is not very easy to use in practice, because many user defined parameters are required in this method, while the meanings of these parameters are not always straightforward.

5.2.4.5. Bezier–Bernstein neurofuzzy model Hong and Harris [79] proposed a neurofuzzy model construction algorithm for nonlinear dynamic systems based upon basis functions that are Bézier–Bernstein polynomial functions including the univariate functions:

$$B_j^{(d)}(s) = \frac{d!}{j!(d-j)!} s^j (1-s)^{(d-j)} \tag{72}$$

where j and d are non-negative integer number, $j \leqslant d$ over the region $s \in [0, 1]$, and the bivariate functions:

$$B_{i,j,k}^{(d)}(u, v, p) = \frac{d!}{i!j!k!} u^i v^j p^k \quad (i + j + k = d) \tag{73}$$

where i, j, and k are non-negative integer numbers, and $u+v+p=1$. The network construction is based on an additive decomposition approach via the ANOVA expansion, in which the univariate neurofuzzy submodel with Bernstein polynomials $B_j^{(d)}(s)$ as basis functions is

$$f_k(x_k) = \sum_{j=0}^{d} w_j B_j^{(d)}(s(x_k)) \tag{74}$$

whilst the bivariate nonlinear neurofuzzy submodel with Bernstein polynomials $B_{i,j,k}^{(d)}(u, v, p)$ as basis functions is

$$f_{k_1k_2}(x_{k_1}, x_{k_2}) = \sum_{i+j+k=d} w_{ijk} B_{ijk}^{(d)}(u(x_{k_1}, x_{k_2}), v(x_{k_1}, x_{k_2}), p(x_{k_1}, x_{k_2})) \tag{75}$$

where $u(x_{k_1}, x_{k_2}) + v(x_{k_1}, x_{k_2}) + p(x_{k_1}, x_{k_2}) = 1$, $0 \leqslant u(x_{k_1}, x_{k_2}), v(x_{k_1}, x_{k_2}), p(x_{k_1}, x_{k_2}) \leqslant 1$, w_j and w_{ijk} are the parameters to be identified from data.

In this method, the principle behind the construction algorithms for each of the univariate submodels $f_k(x_k)$ and the bivariate submodels $f_{k_1k_2}(x_{k_1}, x_{k_2})$ is an inverse procedure of the de Casteljau algorithms that are used in Bézier curve and surface geometric design [53]. Very interestingly, this Bézier–Bernstein polynomial-based neurofuzzy model leads to structural parsimony and Delaunay triangulation input space partition, which is completely different from the grid input space partition. Hence this model possesses the high-level interpretability and overcomes the curse of dimensionality well associated with conventional fuzzy and RBFNs.

5.2.4.6. SVM In the following, we briefly review how the SVM technique is applied to induce 0-order TS fuzzy rules such that the outstanding advantage of SVM in producing parsimonious model structure can be employed to improve the high-level fuzzy model interpretability in terms of compactness of rule base. First, let us revisit the 0-order TS fuzzy systems consisting of the following N fuzzy rules:

$$\textit{Rule}_i: \text{ If } x_1 \textit{ is } A_{i,1} \textit{ and } \cdots x_n \textit{ is } A_{i,n} \text{ then } y_i = w_i \quad (i = 1, \ldots, N) \tag{76}$$

In order for the input space to be thoroughly covered by the fuzzy rule "patches'', the following auxiliary rule is added into the rule base:

$$\textit{Rule}_0: \textit{ if } x_1 \textit{ is } A_{0,1} \textit{ and } \cdots \textit{ and } x_n \textit{ is } A_{0,n} \textit{ then } ; y_0 = b_0 \tag{77}$$

where $A_{0,j}$ denotes the domain of x_j, i.e., its MF $\mu_{0,j}(x_j) \equiv 1$, and $b_0 \in \Re$. The overall output of the system with the auxiliary rule is expressed by

$$F(x) = b_0 + \sum_{i=1}^{L} \phi_i(x) w_i \Bigg/ \left(1 + \sum_{i=1}^{L} \phi_i(x)\right) \tag{78}$$

where $x = [x_1, \ldots, x_n]^{\mathrm{T}}$, ϕ_i is the *firing strength* of the ith rule, which is usually calculated in terms of the product operator. Because $1 + \sum_{i=1}^{N} \phi_i(x) > 0$, which does not impact the signs of output (78), the binary fuzzy classifier can be defined as

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N} \phi_i(x) w_i + b_0\right) \tag{79}$$

Hence, in order to apply SVM learning to (79), $\phi_i(x)$ must be a Mercer kernel. Fortunately, as analysed by Chen and Wang [32], if the MFs of fuzzy sets $A_{i,j}$ are generated from a reference function $\rho(\cdot)$ [47] through location shift, i.e., $\mu_{i,j}(x_j) = \rho(x_j - c_j^{(i)})$, and the Fourier transform of the reference function is non-negative, then $\phi_i(x) = \phi(x, c^{(i)}) = \prod_{j=1}^{n} \rho(x_j - c_j^{(i)})$ is proved to be a Mercer kernel, where $c^{(i)} = [c_1^{(i)}, \ldots, c_n^{(i)}]^{\mathrm{T}}$ is called prototype or kernel centre. Then one can obtain an SVM-based fuzzy classifier with optimal decision function, which has the form,

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N} \phi(x, x^{(i)}) \alpha_i y^{(i)} + b_0\right) \tag{80}$$

where $\{(x^{(l)}, y^{(l)})\}_{l=1}^{N}$ are training samples with $y^{(l)} \in \{-1, 1\}$, and the coefficients α_i are obtained by solving the following quadratic programming (QP) problem:

$$\max_{\alpha_i} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y^{(i)} y^{(j)} \phi(x^{(i)}, x^{(j)}) \tag{81}$$

s.t.

$$C \geqslant \alpha_i \geqslant 0, \quad i = 1, \ldots, N$$
$$\sum_{i=1}^{N} \alpha_i y^{(i)} = 0 \tag{82}$$

where the regularization constant $C > 0$ determines the trade-off between the empirical error and the complexity term.

Due to the nature of this QP problem, only a number of coefficients α_i are non-zero, and the input training samples associated to non-zero α_i are called *support vectors* (SVs), thus a sparse solution of f in (80) is obtained. Let N_s be the number of SVs $\{x_s^{(i)}\}_{i=1}^{N_s} \subseteq \{x^{(l)}\}_{l=1}^{N}$, $\{\alpha_i^0\}_{i=1}^{N_s}$ be the non-zero coefficients, and $y_s^{(i)}$, $i = 1, 2, \ldots, N_s$, are the class labels of the corresponding SVs. Then the bias term b_0 can be computed as follows:

$$b_0 = \frac{1}{N_s} \sum_{j=1}^{N_s} \left(y_s^{(j)} - \sum_{i=1}^{N_s} \alpha_i^0 y_s^{(i)} \phi(x_s^{(j)}, x_s^{(i)}) \right) \tag{83}$$

and a parsimonious fuzzy classifier is obtained as

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N_s} \phi(x, x_s^{(i)}) \alpha_i^0 y_s^{(i)} + b_0 \right) \tag{84}$$

whose fuzzy rules are induced as follows [29,199]:

- Definition of the premise parts of fuzzy rules: the MFs of the ith rule are induced as $\mu_{i,j}(x_j) = \rho(x_j - c_j^{(i)})$, where $c^{(i)} = [c_1^{(i)}, \ldots, c_n^{(i)}]^{\mathrm{T}}$ is the ith SV $c^{(i)} \in \{x_s^{(i)}\}_{i=1}^{N_s}$.
- Definition of the consequent parts of fuzzy rules: the consequent value of the ith rule is induced as $w_i = \alpha_i^0 y_s^{(i)}$, $i = 1, \ldots, N_s$, where α_i^0 represent non-zero coefficients, and $y_s^{(i)}$ are the class labels of the SVs $x_s^{(i)}$.

In this way, given an SVM classifier with N_s SVs, we get an equivalent 0-order TS fuzzy system that consists of $N_s + 1$ rules including the auxiliary rule R_0 as described in (77).

It is noted that the SVM is basically a binary classifier [175,176], but different schemes can be applied to the basic SVM algorithm to handle multiple class pattern classification problems [76,185]. These schemes include combination of 1-to-rest classifiers, pairwise coupling and pairwise with majority voting, extending the formulation of SVM to support the multiple class problem.

Generally speaking, SVM is a promising statistical system modelling technique due to its capability of producing parsimonious solution. The invention of SVM was driven by the statistical learning theory that is rooted in VC dimension, which makes its derivation even more profound [175,176]. The VC dimension theory shows that the SVM solution is found by minimizing both the error on the training set (empirical risk) and the complexity of the hypothesis space expressed in terms of VC-dimension. In this sense, the decision function found by SVM is a trade-off between learning error and model complexity. It was shown that if the empirical risk is zero, the expectation value of the probability of committing an error on a test example is bounded by the ratio between the expectation value of the number of SVs and the number of training vectors [39]. In other words, if a relatively small number of SVs construct the optimal hyperplane, the generalization ability will be high-even in an infinite dimensional space. Hence, SVM classifiers usually achieve good generalization performance and avoid overfitting problem.

From the perspective of interpretable fuzzy system modelling, the SVM-based fuzzy classifier produces parsimonious rule base with as few rules as possible to fit the system well, which leads to a good high-level interpretable fuzzy model. In the SVM-based fuzzy classifier, each SV in the SVM corresponds to one rule in the fuzzy model, which means

that the number of fuzzy rules is irrelevant to the dimension of the input space. Hence, the SVM technique provides an efficient way for fuzzy modelling to evade the "curse of dimensionality'' in high-dimensional input space [199]. What's more, after the SVM learning process, not only the SVs, i.e., the prototypes, are generated, its classifier defined by the SVs in a decision surface is also produced. One does not need to specify the number of prototypes in advance, as these multiplier coefficients α_i^0 are naturally obtained from solving a QP problem. That is to say, all the three tasks in designing a prototype-based classifier are completed together and automatically addressed from data in one model structure [199].

However, how to select appropriate parameters for kernel functions so as to improve the generalization performance remains an open problem for most kernel machine models [23,64,102,150–152]. Facing so many parameters in SVM-based fuzzy classifiers, methods based on exhaustive search become intractable. In order to attack this problem, Zhou and Gan proposed to use L2-SVM technique to induce parsimonious fuzzy rule base [199], in which kernel parameters are learned optimally from data according to the radius-margin bound and features ranking is performed simultaneously in an integrated manner so that fuzzy rules can be generated optimally from data.

5.2.5. Tree structures

In [134], Pal and Chakraborty proposed fuzzy interactive dichotomizer (ID3)-type decision trees to extract fuzzy rules from examples. By minimizing the gain in terms of fuzzy entropy, the best feature is selected at each node of the decision tree. Based on a stopping criterion, a decision is made whether a node will be a leaf node. The decision tree was initially proposed by Quinlan in a crisp case [141] and was developed by many researchers [88,184]. Ichihashi et al. suggested an iterative implementation of ID3 principle based on a neuro-fuzzy learning algorithm [81]. In the method for generating fuzzy rules proposed by Jakel et al. using decision tree induction mechanism, the linguistic hedges are used to form linguistic terms derived, which is helpful to make rule antecedents more compact [84]. Mikut proposed a data-driven interpretable fuzzy modelling method in which rule hypotheses are produced by a decision tree and their pruning favour simple premises [121]. In fact, the tree represents a subspace of all the possible rules. Moreover the information entropy measures are used for tree induction, pruning and selection, which allow the user to influence the compromise between accuracy and interpretability. From the perspective of interpretability, the main advantage of fuzzy decision trees is that they can offer a compact rule base by using only the locally most influential variables to generate incomplete rules for a given partition [67]. However, the greedy characteristics of fuzzy decision trees leads to its over-sensitivity to noise, outliers or irrelevant attributes.

5.2.6. Rule merging/aggregating techniques

In order to improve the interpretability of neuro-fuzzy classifier NEFCLASS [125], Klose et al. presented a greedy approach to merging induced fuzzy rules in a modified NEFCLASS system via three phases [104,105]: *input pruning on data set level*, *input pruning on rule level* and *rule merging*. In [62], Gegov suggested a formal simplification technique by removing the inherent redundancy in an inconsistent fuzzy rule base. This type of redundancy is expressed by the presence of inconsistent rules and removed by aggregating such rules to obtain consistent rule base. The overall process of aggregating inconsistent rules is illustrated as follows:

Step 1: Put all inconsistent rules in groups whereby the rules in each group have the same permutation of linguistic values of inputs and different permutations of linguistic values of outputs.

Step 2: For each group of rules, find a single equivalent rule whose effect on the defuzzified output is the same as the effect of all rules.

Step 3: If it is not possible to apply *Step* 2, then find a subset of equivalent rules whose effect on the defuzzified output is the same as the effect of the whole set of rules for this group.

Step 4: For each group of rules, keep either the single equivalent rule or the subset of equivalent rules and remove all the other rules.

Moreover, the implementation of the above steps can be done using Boolean matrices or binary relations as indicated for instance in Table 1 or (8) [62]. And a similar mechanism is used to remove the redundancy expressed by the presence of non-monotonic fuzzy rules. Interestingly, Gegov suggested a formal approach to manipulation of multiple rule bases by applying the merging/splitting techniques, such as vertical merging/splitting manipulation of rule bases, horizontal merging/splitting manipulation of rule bases, output merging/splitting manipulation of rule bases [62]. However, the interpretability improvement techniques reviewed in this paper focus on operations on fuzzy sets and rules in a single rule base.

5.2.7. Rule ranking and selection techniques

In fuzzy modelling, one commonly used strategy to generate compact rule base starts with an oversized number of partitions in input space, then refines the rule base by selecting the influential fuzzy rules and removing the redundant ones. An important advantage of this selection is to reduce the possible redundancy existing in the rule base, thus improve the model interpretability and generalization capability. The key point is to rank fuzzy rules in terms of an appropriate index.

5.2.7.1. Rule ranking methods by considering rule base structure The firing strength matrix of a fuzzy system is defined as

$$G = \begin{bmatrix} p_1(x^{(1)}) & \cdots & p_L(x^{(1)}) \\ \vdots & \vdots & \vdots \\ p_1(x^{(N)}) & \cdots & p_L(x^{(N)}) \end{bmatrix}_{N\times L} \tag{85}$$

where $p_i(x^{(k)})$ is defined in (12). In G, each column corresponds to one fuzzy rule, important fuzzy rules correspond to the columns of the matrix that are linearly independent of each other [122]. The SVD of G plays an important role in rule ranking and rule subset selection. As indicated in [190], the redundant fuzzy partitions (corresponding to the linear dependent or zero-valued columns) are associated with near zero singular values of G. The smaller are the singular values, the less influential are the associated fuzzy rules. Mouzouris and Mendel first applied the SVD-QR with column pivoting algorithm to G for identifying the most important fuzzy rules from a given rule base [122,123]. Furthermore, this algorithm was popularized by Yen et al. in fuzzy modelling [190,189]. In [189], statistical information criteria together with SVD based rule selection technique are used to seek the fitness-complexity trade-off. The advantage is that the number of rules necessary for a compact rule base can be determined automatically. The *SVD-QR* with *column pivoting* algorithm was originally proposed by Golub et al. [66] to solve the problem of subset selection in regression analysis and used by Kanjilal and Banerjee [98] to select hidden nodes in a feedforward neural network. The estimation of an effective rank is needed in the *SVD-QR* with *column pivoting* algorithm.

Another way of fuzzy rule ranking based on matrix decomposition is to apply the pivoted QR decomposition directly to the matrix G [155,158], in which R-values defined as the absolute values of diagonal elements of matrix R in QR decomposition tend to track the singular values of the matrix G and can be used for rule ranking to identify the influential rules. The pivoted QR decomposition algorithm for ranking fuzzy rules is addressed as follows:

Step 1: Calculate the QR decomposition of G and get the permutation matrix Π via $G\Pi = QR$, where Q is an unitary matrix, R is an upper triangular matrix. The absolute values of the diagonal elements of R, denoted as $|R_{ii}|$, decrease as i increases and are named as R-values.

Step 2: Rank fuzzy rules in terms of the R-values and the permutation matrix Π. In Π each column has one element taking value 1 and all the other elements taking value 0. Each column of Π corresponds to a fuzzy rule. The numbering of the jth most important rule in the original rule base is the same as the numbering of the row where the "1'' element of the jth column is located. For example, if the "1'' of the 1st column is in the 4th row, then the 4th rule is the most important one and its importance is measured as $|R_{11}|$. The rule corresponding to the first column is the most important, and in descending order the rule corresponding to the last column is the least important.

By applying the pivoted QR decomposition algorithm to the induced fuzzy classifier, each rule can be assigned an R-value. The importance of a fuzzy rule is measured by its associated R-value. However, the pivoted QR decomposition algorithm and the SVD-QR with column pivoting algorithm when applied to fuzzy rule ranking ignore the effect of the rule consequents.

5.2.7.2. Rule ranking methods by considering output contributions of fuzzy rules Another scheme of fuzzy rule ranking is to consider the output contribution of the fuzzy rules. One appropriate tool is the OLS technique [30]. Wang and Mendel first applied the OLS method to fuzzy rule selection [183], in which the OLS was used to select the most important fuzzy basis functions, each of which was associated with a fuzzy rule. The OLS-based methods select the most important fuzzy rules based on their contributions of variance to the variance of the output. For the 0-order TS fuzzy system defined in (3) and (4) with $a_{ki} = 0$ ($i = 1, \ldots, L$; $k = 1, \ldots, n$), the firing strength matrix reduces to the firing matrix G in (85). As stated above, each column of G corresponds to one fuzzy rule. The OLS method transforms the columns of G into a set of orthogonal basis vectors in the interest of identifying the individual contribution of each

rule. The ratio defined by (20) offers a simple criterion of ordering the rules and selecting a subset of important rules in a forward regression manner [181,183]. The first rules with the largest error reduction ratios will be selected. The selection stops when the cumulated explained variance is satisfactory, i.e., the output is reconstructed well enough, which occurs at step r when the criterion defined in (21) is satisfied.

However, when applied to fuzzy rule selection and ranking, because the OLS-based method is guided by the approximation capabilities of the rules (fitting error) without paying attention to the premise structures, it may take certain risk since low variance does not necessarily suggest the corresponding rule is unimportant in the model. So Setnes and Hellendoorn additionally considered the dependency between the current rule to be selected and the set of rules previously selected using OLS [158]. Mastorocostas et al. proposed a constrained OLS as an improvement to the basic OLS approach for producing compact fuzzy rule base [117]. In the algorithm, both the premise part redundancy and input variable selection for consequent part are considered in the process of model construction. In an effort of building an interpretable 0-order TS fuzzy rule base from data [46], Destercke et al. run the OLS algorithm twice: the first pass of OLS algorithm is to select most important fuzzy rules and the second pass to optimize the rule consequents. No rule selection was made, but the k-means algorithm was used to reduce the number of distinct rule consequents.

Different from the rule selection based on the OLS approach, Zhou and Gan suggested another fuzzy rule ranking index considering the effect of rule consequents [199], which is called *α-values of fuzzy rules*. It can be seen from Section 5.2.4.6 that in the SVM-based fuzzy modelling, for each induced fuzzy rule, its associated Lagrangian multiplier α_i^0 actually determines the depth of the effect of the rule consequent, hence, they can be used to rank the rule output contributions and act as an fuzzy rule index to select the most influential fuzzy rules [199].

5.2.7.3. Rule ranking methods by considering both rule base structure and output contributions of fuzzy rules As indicated in [155], it is highly desired for a rule ranking method to take into account both the rule base structure and the output contribution of the fuzzy rules in order to generate a compact rule base with good generalization performance. However, currently this kind of more reasonable rule ranking scheme is very rare. Zhou and Gan proposed the so-called ω-values of fuzzy rules to rank the fuzzy rules [199]:

$$\omega_i = \frac{\tilde{\alpha}_0^{(i)} \cdot |R_{ii}|}{\max_i \tilde{\alpha}_0^{(i)} \cdot \max_i |R_{ii}|} \tag{86}$$

where $|R_{ii}|$ are the R-values of fuzzy rules as addressed above via pivoted QR decomposition for evaluating the fuzzy rules by considering the rule base structure only, whilst $\tilde{\alpha}_0^{(i)}$ are the α-values of fuzzy rules measuring the rule output contributions of fuzzy rules. In such a way, the defined ω-values of fuzzy rules rank the fuzzy rules by considering both the rule base structure and the output contribution of the fuzzy rules.

5.2.8. Multi-objective optimization

It is noteworthy that all the above constructing techniques for high-level interpretability of fuzzy models focus on the criterion of rule base parsimony and simplicity without consideration of other criteria of high-level interpretability, such as consistency and completeness. No doubt, a desirable fuzzy model with high-level interpretability should take into account all the criteria as indicated in Section 4.2.

5.2.8.1. Multi-objective optimization with evolutionary strategies and GAs for interpretable fuzzy modelling Jin et al. in [92] proposed a systematic design paradigm for generating complete, consistent and compact fuzzy systems based on (μ, λ)-evolution strategies (ES), in which the structure of the fuzzy rules determining compactness of the fuzzy system is evolved along with the parameters of the fuzzy system, the completeness is guaranteed by checking the completeness of the fuzzy partitioning of input variables, and the completeness of the rule structure.

A (μ, λ)-(ES) algorithm [12,154] is to deal with mixed optimization, which can be described by the following notation:

$$(\mu, \lambda)\text{-}ES := (I, \mu, \lambda; m, s, \sigma; f, g) \tag{87}$$

where I is a string of real or integer numbers representing an individual in the population, μ and λ are the numbers of the parents and offsprings, respectively, σ is the parameter to control the step size, m represents the mutation operator,

which is the main operator in the mechanism of ES, s stands for the selection method and in this case the parents will be selected only from the λ descendants, f is the objective function to be minimized and g is the constraining function to which the variables are subject. The variables to be optimized and the step-size control parameter are mutated in the following way:

$$\sigma_i' = \sigma_i \cdot \exp(\tau_1 \cdot \mathrm{N}(0,1) + \tau_2 \cdot \mathrm{N}_i(0,1)), \quad i = 1, \ldots, Q \tag{88}$$

$$I_i' = I_i + \sigma_i' \cdot \mathrm{N}_i(0,1), \quad i = 1, \ldots, Q_1 \tag{89}$$

$$I_i' = I_i + \lfloor \sigma_i' \cdot \mathrm{N}_i(0,1) \rfloor, \quad i = Q_1 + 1, \ldots, Q \tag{90}$$

where $\mathrm{N}(0,1)$ and $\mathrm{N}_i(0,1)$ are normally distributed random numbers with mean of zero and variance of 1, Q is the total number of variables to be optimized, Q_1 is the number of real variables and naturally $Q - Q_1$ denotes the number of the integer variables and $\lfloor \cdot \rfloor$ is the maximal integer smaller than the real number inside, τ_1 and τ_2 are two global step control parameters.

In this method, the parameters representing the fuzzy operators and MFs are encoded with real variables, while the rule structure parameters are encoded with integer numbers. In order to prevent the step-size from converging to zero, it was re-initialized with a value of 1.0, when σ becomes very small. In the evolution process, the completeness of the fuzzy partitioning of each input variable is examined using the fuzzy similarity measure (26), and similar fuzzy sets are merged for partition distinguishability. As for the rule consistency, the consistency index defined in (39) cannot be directly applied in the (μ, λ)-(ES) algorithm, so a degree of inconsistency of a rule base RB_1 generated from data was suggested as follows:

$$f_{\mathrm{Incons}} = \sum_{i=1}^{N_1} Incons(i) \tag{91}$$

where $Incons(i)$ represents the degree of inconsistency of the ith rule in the rule base RB_1 with N_1 rules, which is defined as

$$Incons(i) = \sum_{\substack{1 \leqslant k \leqslant N_1 \\ k \neq i}} [1 - Cons(RB_1(i), RB_1(k))] + \sum_{1 \leqslant k \leqslant N_2} [1 - Cons(RB_1(i), RB_2(k))] \tag{92}$$

where RB_2 represents the rule base extracted from prior knowledge, N_2 denotes the number of rules in base RB_2. The degree of inconsistency of a rule base f_{Incons} is incorporated in the objective function of the (μ, λ)-(ES) algorithm to prevent from generating fuzzy rules that seriously contradict with each other or even with the heuristic knowledge.

In order to find the right trade-off between global model performance and fuzzy model interpretability, Cococcioni et al. [37] proposed a Pareto-based multi-objective evolutionary approach to generate a set of Mamdani fuzzy systems from numerical data, in which a variant of the $(2+2)$ Pareto archived evolutionary strategy ($(2+2)$PAES) was adopted. The $(2+2)$PAES determines an approximation of the optimal Pareto front by concurrently minimizing the root MSE and the complexity. Complexity was measured as sum of the conditions which compose the antecedents of the rules. Thus, the parsimonious Mamdani fuzzy systems with a small number of rules and a small number of input variables were obtained.

To seek good trade-off between number of fuzzy rules and global model accuracy, Alcala et al. [4] employed the SPEA2 algorithm to get the Pareto solutions with the least number of possible rules but still presenting high accuracy, in which appropriate genetic operators and some modifications were applied. Ishibuchi and Nojima [83] used a multi-objective fuzzy genetics-based machine learning algorithm to examine the interpretability-accuracy trade-off in fuzzy rule-based classifiers. Each fuzzy rule was represented by its antecedent fuzzy sets as an integer string of fixed length, whilst the rule base was represented as a concatenated integer string of variable length, then the multi-objective GA simultaneously maximizes the accuracy of rule sets and minimizes their complexity. In [82], Ishibuchi et al. showed two GA-based approaches to extraction of linguistically interpretable fuzzy rules from high-dimensional data. One is rule selection where a small number of fuzzy rules were selected from a large number of pre-specified candidate rules. The other is fuzzy genetics-based machine learning where rule sets were evolved by genetic operations. These two approaches searched for non-dominated rule sets with respect to the following three objectives: to maximize the

classification performance of correctly training patterns, to minimize the number of fuzzy rules, and to minimize the total number of antecedent conditions.

5.2.8.2. Global learning and local learning for interpretable TS fuzzy local linear models We argue that TS fuzzy model interpretability due to its rule structure hails from the interaction between global model and its local linear models [93,190]. In [190], Yen et al. proposed a novel scheme that combines global learning and local learning by minimizing a combined objective function to improve the TS local model interpretability in the sense of the definition in Section 4.2.5). This combined learning algorithm is described as follows:

$$\min_f(\alpha J_{\mathrm{G}} + \beta J_{\mathrm{L}}) \tag{93}$$

s.t.

$$\alpha + \beta = 1 \tag{94}$$

where J_{G} is the global model objective function defined in (44), while J_{L} is the local model objective function defined in (42), α and β are weighting coefficients to impact the balance between global model performance and local model interpretability.

Johansen and Babuska applied a slight extension of (93) to solve the optimization problem [93],

$$\min_f \left(J_{\mathrm{G}} + \sum_{i=1}^{L} \beta_i J_{\mathrm{L}_i} \right) \tag{95}$$

s.t.

$$H_i(\theta_i) = 0, \quad F_i(\theta_i) \leqslant 0 \quad (i = 1, \ldots, L) \tag{96}$$

$$\beta_i \geqslant 0 \quad (i = 1, \ldots, L) \tag{97}$$

where $J_{\mathrm{L}_i} = \sum_K^N w_i(k)[y^{(k)} - y_i(k)]^2$, H_i and F_i are affine functions of consequent parameters θ_i (corresponding to translating *a priori* knowledge about the system being modelled into these constraints on consequent parameters), L denotes the number of TS rules, $\beta_i \geqslant 0$ are the parameters for the set of Pareto-optimal solutions of the underlying multi-objective optimization problem, which essentially determine the trade-off between the possibly conflicting objectives of global model accuracy and local model interpretation.

5.2.8.3. Multi-objective learning incorporating all criteria It is pinpointed that multi-objective optimization is a suitable and promising technique in interpretable fuzzy system modelling, because so many criteria should be considered to achieve a fuzzy model with good interpretability and global performance. However, currently, the efforts of using multi-objective optimization techniques to improve high-level interpretability of fuzzy models only consider part of the criteria, mainly the parsimony of rule base. Given the appropriate formalized constraint expressions for the high-level interpretability criteria, the multi-objective optimization technique has the great potential of being applied to construct high-level interpretable fuzzy models with the consideration of all the criteria including the global model objective, in which Pareto-optimal solution is to determine the trade-off between global model performance and high-level interpretability of fuzzy models.

The characteristics of the existing construction techniques for high-level interpretability of fuzzy models are summarized in Table 3, in which some drawbacks of these techniques are also indicated, whilst the advantages lie in their abilities to achieve high-level interpretable fuzzy models by considering corresponding criteria in the modelling process.

5.3. Constructive techniques for both low-level interpretability and high-level interpretability of fuzzy models

In the previous two subsections, we reviewed the constructive techniques for achieving *low-level interpretability* and *high-level interpretability* separately. Undoubtedly, a more attractive and efficient constructive technique for improving fuzzy model interpretability should automatically take into account both the semantic criteria and syntactic criteria to

Table 3
Techniques for high-level interpretability of fuzzy models

Schemes	High-level interpretability criteria					Comments
	Rule base parsimony	Readability of single rule	Consistency Consistency	Completeness Completeness	Transparency of rule structure	
ANOVA decomposition	Yes	No	No	No	Yes (Mamdani model) No (TS model)	Losing transparent solution if the object being modelled contains high-dimensional interactions
Input variable selection	Yes	No	No	No	Yes (Mamdani model) No (TS model)	High computing overhead
Grid partitioning	No	Yes	No	No	Yes (Mamdani model) No (TS model)	Curse of dimensionality
Scatter partitioning	Yes	No	No	No	Yes (Mamdani model) No (TS model)	Poor rule readability
Fuzzy tree	Yes	No	No	No	Yes (Mamdani model)	Over-sensitivity to noise, outliers or irrelevant attributes
Rules merging	Yes	No	No	No	Yes (Mamdani model)	High computing overhead
Rule ranking	Yes	No	No	No	Yes (Mamdani model)	High computing overhead
Multi-objective optimization	Yes	Yes	Yes	Yes	Yes	High computing overhead

obtain both low-level interpretable and high-level interpretable fuzzy models in one model structure. However, very few efforts have been made in this aspect. Indeed, some low-level interpretability improvement techniques on fuzzy set level such as merging redundant fuzzy sets can reduce the redundancy in rule base, thus improve the high-level interpretability, but if one does not consider the criteria for high-level interpretability at the same time, inconsistent rules may be produced and rule base may still remain redundancy.

It seems that some prototype-based fuzzy modelling techniques, for example the family of FCM [14,50], can be used to automatically construct both low-level interpretable and high-level interpretable fuzzy models in one model structure. Because the training samples are clustered into important homogeneous regions (i.e., prototypes) which are characterized by multi-dimensional fuzzy sets, a rule is associated to each region, i.e., the premise part of each rule is a multi-dimensional fuzzy set, and the MFs on individual variables can be obtained by projecting the multi-dimensional fuzzy set onto the corresponding antecedent individual variables. Indeed, initialization of input space partitioning by prototypes plays an important role in improving fuzzy model interpretability [80,148]. If an appropriate clustering validation index for efficiently determining the number of prototypes is applied, fuzzy clustering can generate a parsimonious rule base. However, two problems remain in fuzzy clustering for constructing both low-level interpretable and high-level interpretable fuzzy model. One is that fuzzy rules with multi-dimensional fuzzy sets in premise parts may possess poor interpretability, because it is hard to render the multi-dimensional MFs meaningful linguistic terms. Another problem is that even though the MFs on individual variables are used to form input space partitioning, the distinguishability of these MFs obtained by projecting multi-dimensional MFs are not always guaranteed due to too much overlapping [10,67].

Currently, how to automatically construct both low-level interpretable and high-level interpretable fuzzy model in one model structure by considering all the criteria still remains an open way of research.

6. Open problems and future research topics on fuzzy model interpretability

Data-driven fuzzy system modelling is widely and increasingly used for modelling tasks with tremendously successful applications in many areas, such as forecasting, classification, system modelling and control, fault diagnosis, data mining and knowledge discovery, particularly in modelling complex systems and processes whose first-principle models are

unknown. However, interpretability of the fuzzy models resulted from data-driven fuzzy modelling with adaptive learning is most probably lost due to the accuracy-oriented nature. Hence, interpretable data-driven fuzzy system modelling methods are in great demand. They not only possess the capabilities to predict the system behaviours, but also can be interpreted with sensible fuzzy MFs. Whilst some progresses have been made in describing the theoretical and practical aspects of interpretable fuzzy modelling, as reviewed in this paper, interesting issues have cropped up.

Interpretable fuzzy modelling has recently received increasing attention in the communities of fuzzy logic, machine learning and intelligent data analysis. The methods developed for interpretable fuzzy modelling have great potentials of being extended to other areas concerning knowledge extraction and discovery, such as data mining, knowledge-based systems, etc.

In order to construct interpretable fuzzy models with good global model performance, formalized expressions of criteria for low-level interpretability and high-level interpretability are very important. It should be noted that the formalized definition efficiently used for Mamdani fuzzy models could not be suitable for TS fuzzy models, such as the consistency measure of fuzzy rules [92]. It is highly desirable to develop some formalized definitions of low-level interpretability criteria and high-level interpretability criteria, which fit within both Mamdani models and TS models and can be easily integrated with global model performance measure aiming at good trade-off between the two conflicting modelling objectives.

Although some researchers have made efforts to measure the interpretability of fuzzy models, such as via similarity measure [156], possibility measure [77,118,119] and entropy measures [58,59,198], a general way of measuring fuzzy model interpretability, which can be widely accepted like the accuracy measurement, is still an unsolved problem. The framework of low-level interpretability and high-level interpretability of fuzzy models proposed in this paper gives a new line of thought for this effort, that is, a definition of fuzzy model interpretability measurement should take into account both low-level interpretability and high-level interpretability, whereas the existing efforts were mainly made on the measurements of low-level interpretability only.

It can be seen from Tables 2 and 3 that most existing techniques for designing interpretable fuzzy models only consider individual criteria in the modelling process. A promising scheme is to take into account all the criteria in achieving standardized fuzzy partition or high-level interpretable fuzzy models.

Particularly, the majority of the current efforts on improving fuzzy model interpretability focus on constructive techniques for achieving low-level interpretability and high-level interpretability separately. Some researchers have suggested to build interpretable fuzzy models in terms of some criteria for both low-level interpretability and high-level interpretability in modelling process [83,180]; however, separated methods were used to improve individual low-level interpretability and high-level interpretability. Undoubtedly, a more attractive and efficient constructive technique for improving fuzzy model interpretability should aim at achieving interpretable fuzzy model automatically with both low-level interpretability and high-level interpretability within one model structure by considering all the criteria.

Fuzzy rule ranking and selection is an important way of constructing high-level interpretable fuzzy models. More efficient fuzzy rule ranking indexes considering both the rule base structure and contributions of outputs will be of great help in identifying the most influential rules and removing the redundant ones. No doubt, it is highly desirable for a fuzzy rule ranking index to fit both Mamdani fuzzy models and general TS fuzzy models.

The scheme of multi-objective learning combines global learning and local learning, which is suitable for improving fuzzy model interpretability. An open problem here is how to automatically identify optimal weighting coefficients in a combined learning objective function.

The goal of interpretable fuzzy modelling is to construct a fuzzy model with good trade-off between global model accuracy and model interpretability. However, the costs and complexity analyses of the developed methods in achieving model interpretability should be considered as well. Currently only a few researchers have been aware of the issue of interpretability costs [46].

It is pinpointed that human's prior knowledge plays pivotal role in interpreting fuzzy models. How to interpret a built fuzzy model is strongly affected by subjectivity [38,60,61]. Hence an appeal is made to fuzzy logic and machine learning communities to select some benchmark examples for demonstrating the performances of fuzzy modelling techniques in improving model interpretability, though this is a hard task [38,149].

Currently, as a new research domain, interpretable fuzzy modelling focuses on the interpretability improvement of fuzzy models with type-1 fuzzy sets as reviewed in this paper; however, when type-2 fuzzy sets [95,120,194]. are constructed from data by adaptive-learning algorithms, preserving interpretability of type-2 fuzzy models during adaptation is also an important issue.

7. Conclusions

This paper suggests that the interpretability of fuzzy models can be considered in the framework of *low-level interpretability* at fuzzy set level and *high-level interpretability* at fuzzy rule level. The significance of this taxonomy lies in that it provides a useful guideline for interpretable fuzzy system modelling, and that confusions between fuzzy model interpretability achieved by adjusting MFs and the one obtained by applying transparent modelling techniques in traditional statistical system modelling can be avoided. Moreover, this paper reviews various data-driven interpretable fuzzy modelling techniques in literature from the perspective of low-level interpretability and high-level interpretability. And some open problems and potential future research directions have also been identified in this paper.

Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive suggestions that have been incorporated in this paper, and thank Professor Bob John and Dr. Alexander Gegov for providing some materials.

References

[1] J. Abonyi, R. Babuska, H.B. Verbruggen, F. Szeifert, Incorporating prior knowledge in fuzzy model identification, Internat. J. Systems Sci. 31 (5) (2000) 657–667.
[2] H. Akaike, Fitting autoregressive models for prediction, Ann. Inst. of Statist. Math. 21 (1969) 243–247.
[3] R. Alcala, J. Alcala-Fdez, J. Casillas, O. Cordon, F. Herrera, Hybrid learning models to get the interpretability-accuracy trade-off in fuzzy modelling, Soft Comput. 10 (2006) 717–734.
[4] R. Alcala, J. Alcala-Fdez, M.J. Gacto, F. Herrera, A multi-objective genetic algorithm for tuning and rule selection to Obtain accurate and compact linguistic fuzzy rule-based systems, Internat. J. Uncertainty Fuzziness Knowledge-Based Systems 15 (5) (2007) 539–557.
[5] R. Alcalá, J. Alcala-Fdez, F. Herrera, J. Otero, Genetic learning of accurate and compact fuzzy rule based systems based on the 2-tuples linguistic representation, Internat. J. Approx. Reason. 44 (1) (2007) 45–64.
[6] R. Alcalá, J. Casillas, O. Cordón, F. Herrera, Building fuzzy graphs: features and taxonomy of learning non-grid-oriented fuzzy rule-based systems, Internat. J. Intelligent Fuzzy Systems 11 (3–4) (2001) 99–119.
[7] T. G. Amaral, V. F. Pires, and M. M. Crisostomo, An approach to improve the interpretability of neuro-fuzzy systems, in: Proc. IEEE International Conf. on Fuzzy Systems (FUZZ-IEEE), 16–21 July, Vancouver, BC, Canada, 2006, pp. 8502–8509.
[8] H. C. Anderson, The controller output error method, Ph.D. Dissertation, University of Queensland, Australia, 1998.
[9] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (1950) 337–404.
[10] R. Babuska, Fuzzy Modelling for Control, Kluwer Academic Publishers, Boston, USA, 1998.
[11] R. Babuska, Construction of fuzzy systems-interplay between precision and transparency, in: Proc. of ESIT, Aachen, Germany, 2000, pp. 445–452.
[12] T. Back, Parallel optimization of evolutionary algorithms, in: Parallel Problem Solving from Nature, Springer, Berlin, Germany, 1994, pp. 418–427.
[13] M.R. Berthold, K.-P. Huber, Constructing fuzzy graphs from examples, Intelligent Data Anal. 3 (1) (1999) 37–53.
[14] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, New York, 1981.
[15] J.C. Bezdek, I. Anderson, An application of the *c*-varieties clustering algorithm to polygonal curve fitting, IEEE Trans. Systems Man Cybernet. 15 (5) (1985) 637–641.
[16] J.C. Bezdek, N.R. Pal, Two soft relative of learning vector quantization, Neural Networks 8 (5) (1995) 729–743.
[17] R.J. Bhansali, D.Y. Downham, Some properties of the order of an autoregressive model selected by a generalization of Akaike's EPF criterion, Biometrika 64 (1977) 547–551.
[18] M. Bikdash, A highly interpretable form of Sugeno inference systems, IEEE Trans. Fuzzy Systems 7 (6) (1999) 686–696.
[19] U. Bodenhofer, P. Bauer, Toward an axiomatic treatment of interpretability, in: Proc. Sixth Internat. Conf. on Soft Computing (IIZUKA2000), 2000, pp. 334–339.
[20] U. Bodenhofer, P. Bauer, A formal model of interpretability of linguistic variables, in: J. Casillas, O. Cordón, F. Herrera, L. Magdalena (Eds.), Interpretability Issues in Fuzzy Modelling, Studies in Fuzziness and Soft Computing, Vol. 128, Springer, Heidelberg, 2003.
[21] K. M. Bossley, Neurofuzzy modelling approaches in system identification, Ph.D. Dissertation, University of Southampton, UK, 1997.
[22] M. Brown, C. Harris, Neurofuzzy Adaptive Modelling and Control, Prentice-Hall, NY, 1994.
[23] F. Camastra, A. Verri, A novel kernel method for clustering, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 801–805.
[24] in: J. Casillas, O. Cordón, F. Herrera, L. Magdalena (Eds.), Accuracy Improvements in Linguistic Fuzzy Modelling, Springer, Berlin, 2003.
[25] in: in:J. Casillas, O. Cordón, F. Herrera, L. Magdalena (Eds.), Interpretability Issues in Fuzzy Modelling, Springer, New York, 2003.
[26] J. Casillas, O. Cordón, M.J. del Jesus, F. Herrera, Genetic tuning of fuzzy rule deep structures preserving interpretability and its interaction with fuzzy rule set reduction, IEEE Trans. Fuzzy Systems 13 (1) (2005) 13–29.
[27] G. Castellano, A.M. Fanelli, C. Mencar, Generation of interpretable fuzzy granules by a double-clustering technique, Arch. Control Sci. Special Issue Granular Comput. 12 (4) (2002) 397–410.
[28] G. Castellano, A.M. Fanelli, C. Mencar, A neuro-fuzzy network to generate human-understandable knowledge from data, Cognitive Systems Res. 3 (2) (2002) 125–144.

[29] S. Chen, Basis pursuit, Ph.D. thesis, Department of Statistics, Stanford University, November, 1995.
[30] S. Chen, C.F.N. Cowan, P.M. Grant, Orthogonal least squares learning algorithm for radial basis function networks, IEEE Trans. Neural Networks 2 (2) (1991) 302–309.
[31] S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM Rev. 43 (1) (2001) 129–159.
[32] Y. Chen, J.Z. Wang, Support vector learning for fuzzy rule-based classification systems, IEEE Trans. Fuzzy Systems 11 (6) (2003) 716–728.
[33] V. Cherkassky, F. Mulier, Learning from Data: Concepts, Theory, and Methods, Wiley, New York, 1998.
[35] S.L. Chiu, Fuzzy model identification based on cluster estimation, J. Intelligent Fuzzy Systems 2 (3) (1994) 267–278.
[36] S. Chiu, J.J. Cheng, Automatic rule generation of fuzzy rule base for robot arm posture selection, in: Proc. of NAFIPS Conf., San Antonio, TX, December 1994, pp. 436–440.
[37] M. Cococcioni, P. Ducange, B. Lazzerini, F. Marcelloni, A Pareto-based multi-objective evolutionary approach to the identification of Mamdani fuzzy systems, Soft Comput. 11 (11) (2007) 1013–1031.
[38] O. Cordón, F. Herrera, Author's reply (to comments on the benchmarks in 'a proposal for improving the accuracy of linguistic modelling' and related articles), IEEE Trans. Fuzzy Systems 11 (6) (2003) 866–869.
[39] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learning 20 (3) (1995) 197–273.
[40] R.N. Dave, R. Krishnapuram, Robust clustering method: a unified view, IEEE Trans. Fuzzy Systems 5 (2) (1997) 270–293.
[41] J.L. Deng, Control problems of grey systems, Systems Control Lett. 1 (5) (1982) 288–294.
[42] A. De Luca, S. Termini, A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory, Inform. and Control 20 (1972) 301–312.
[43] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussions), J. Roy. Statist. Soc. B 39 (1977) 1–39.
[44] J.V. de Oliveira, A design methodology for fuzzy system interfaces, IEEE Trans. Fuzzy Systems 3 (4) (1995) 404–414.
[45] J.V. de Oliveira, Semantic constraints for membership function optimization, IEEE Trans. Systems Man Cybernet. Part A 29 (1) (1999) 128–138.
[46] S. Destercke, S. Guillaume, B. Charnomordic, Building an interpretable fuzzy rule base from data using orthogonal least squares-application to a depollution problem, Fuzzy Sets and Systems 158 (18) (2007) 2078–2094.
[47] D. Dubois, H. Prade, Operations on fuzzy numbers, Internat. J. System Sci. 9 (6) (1978) 613–626.
[48] D. Dubois, H. Prade, Fuzzy Sets and Systems: Theory and Applications, Academic, New York, 1980.
[49] D. Dubois, H. Prade, L. Ughetto, Checking the coherence and redundancy of fuzzy knowledge bases, IEEE Trans. Fuzzy Systems 5 (3) (1997) 398–417.
[50] J.C. Dunn, A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters, J. Cybernet. 3 (3) (1973) 32–57.
[51] H.W. Engl, M. Hanke, A. Neubauer, Regularization of Inverse Problems, Kluwer Academic Publishers, Dordrecht, 1999.
[52] J. Espinosa, J. Vandewalle, Constructing fuzzy models with linguistic integrity from numerical data-AFRELI algorithm, IEEE Trans. Fuzzy Systems 8 (5) (2000) 591–600.
[53] G. Farin, Curves and Surfaces for Computer-Aided Geometric Design: A Practical Guide, Academic, Boston, MA, 1994.
[54] A. Fiordaliso, A constrained Takagi–Sugeno fuzzy system that allows for better interpretation and analysis, Fuzzy Sets and Systems 118 (2) (2001) 307–318.
[55] J.H. Fridedman, Multivariate adaptive regression splines (with discussion), Ann. Statist. 19 (1) (1991) 79–141.
[56] B. Fritzke, Growing cell structures—a self-organizing network for unsupervised and supervised learning, Neural Networks 7 (9) (1994) 1441–1460.
[57] B. Fritzke, Incremental neuro-fuzzy systems, in: Proc. of Application of Soft Computing, SPIE Internat. Symp. on Optical Science, Engineering and Instrumentation, San Diego, USA, 1997, pp. 86–97.
[58] T. Furuhashi, On interpretability of fuzzy models, in: N.R. Pal, M. Sugeno (Eds.), Advances in Soft Computing—2002 AFSS International Conference on Fuzzy Systems, Calcutta, India, Lecture Notes in Computer Science, Vol. 2275, February 3–6, 2002, Springer, Berlin, 2002, pp. 12–19.
[59] T. Furuhashi and T. Suzuki, On interpretability of fuzzy models based on conciseness measure, in: Proc. Tenth IEEE Internat. Conf. on Fuzzy Sets (FUZZ-IEEE'01), Melbourne Australia, December, 2001, pp. 284–287.
[60] Q. Gan, C.J. Harris, Fuzzy local linearization and local basis function expansion in nonlinear system modelling, IEEE Trans. Systems Man Cybernet.—Part B 29 (4) (1999) 559–565.
[61] Q. Gan, C.J. Harris, A hybrid learning scheme combining EM and MASMOD algorithms for fuzzy local linearization modelling, IEEE Trans. Neural Networks 12 (1) (2001) 43–53.
[62] A. Gegov, Complexity Management in Fuzzy Systems—A Rule Base Compression Approach, Springer,Berlin, 2007.
[63] J. Geweke, R. Meese, Estimating regression models of finite but unknown order, Internat. Econom. Rev. 22 (1981) 55–71.
[64] M. Girolami, Mercer kernel-based clustering in feature space, IEEE Trans. Neural Networks 13 (3) (2002) 780–784.
[65] S. Gottwald and U. Petri, An algorithmic approach toward consistency checking systems of fuzzy control rules, in: Proc. Third European Congress on Intelligent Techniques and Soft Computing, Aachen, Germany, 1995, pp. 682–687.
[66] G. H. Golub, V. Klema, and G. W. Stewart, Rank degeneracy and least squares problems, Technical Report TR-456, Department of Computing Science, University of Maryland, College Park, MD, USA, 1976.
[67] S. Guillaume, Designing fuzzy inference systems from data: an interpretability-oriented review, IEEE Trans. Fuzzy Systems 9 (3) (2001) 426–443.
[68] S. Guillaume, B. Charnomordic, Generating an interpretable family of fuzzy partitions from data, IEEE Trans. Fuzzy Systems 12 (3) (2004) 324–335.
[69] S.R. Gunn, J.S. Kandola, Structural modelling with sparse kernels, Mach Learning 48 (1–3) (2002) 137–163.
[70] D.E. Gustafson, W. Kessel, Fuzzy clustering with a fuzzy covariance matrix, in: K.S. Fu (Ed.), Proc. IEEE-CDC, Vol. 1.2, IEEE Press, Piscataway, NJ, 1979, pp. 761–766.

[71] E.J. Hannan, The estimation of the order of an ARMA process, Ann. Statist. 8 (1980) 1071–1081.
[72] E.J. Hannan, B. Quinn, The determination of the order of an autoregression, J. Roy. Statist. Soc. Ser. B 41 (1979) 190–191.
[73] W. Hardel, Applied Nonparametric Regression, Cambridge University Press, Cambridge, MA, 1990.
[74] C.J. Harris, X. Hong, Q. Gan, Adaptive Modelling, Estimation and Fusion From Data—A Neurofuzzy Approach, Springer, Berlin, 2002.
[75] T. Hastie, R. Tibshirani, Generalized Additive Models, Chapman & Hall, CRC, London, Boca Raton, FL, 1990.
[76] T. Hastie, R. Tibshirani, Classification by pairwise coupling, Technical Report, Stanford University and University of Toronto, 1996.
[77] H.A. Hefny, Comments on 'Distinguishability quantification of fuzzy sets', Inform. Sci. 177 (21) (2007) 4832–4839.
[78] F. Hoeppner, Fuzzy shell clustering algorithms in image processing: fuzzy c-rectangular and 2-rectangular shells, IEEE Trans. Fuzzy Systems 5 (4) (1997) 599–613.
[79] X. Hong, C.J. Harris, Generalized neurofuzzy network modelling algorithms using Bézier–Bernstein polynomial functions and additive decomposition, IEEE Trans. Neural Networks 11 (4) (2000) 889–902.
[80] F. Hoppner, F. Klawonn, Obtaining interpretable fuzzy models from fuzzy clustering and fuzzy regression, in: Proc. Fourth Internat. Conf. on Knowledge-based Intelligent Engineering Systems and Allied Tech (KES), Brighton, UK, 2000, pp. 162–165.
[81] H. Ichihashi, T. Shirai, K. Nagasaka, T. Miyoshi, Neuro-fuzzy ID3: a method of inducing fuzzy decision trees with linear programming for maximizing entropy and an algebraic method for incremental learning, Fuzzy Sets and Systems 81 (1) (1996) 157–167.
[82] H. Ishibuchi, T. Nakashima, T. Murata, Three-objective genetics-based machine learning for linguistic rule extraction, Inform. Sci. 136 (1–4) (2001) 109–133.
[83] H. Ishibuchi, Y. Nojima, Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning, Internat. J. Approx. Reason. 44 (1) (2007) 4–31.
[84] J. Jakel, L. Groll and R. Mikut, Tree-oriented hypothesis generation for interpretable fuzzy rules, in: Proc. Seventh European Congress on Intelligent Techniques and Soft Computing, Aachen, Germany, 1999, pp. 279–280.
[85] J.-S. Jang, C.-T. Sun, Functional equivalence between radial basis function networks and fuzzy inference systems, IEEE Trans. Neural Networks 4 (1) (1993) 156–159.
[86] J.-S. Jang, C.-T. Sun, E. Mizutani, Neuro Fuzzy and Soft Computing, Prentice-Hall, Englewood Cliffs, NJ, 1997.
[87] J.-S.R. Jang, ANFIS: adaptive-network-based fuzzy inference system, IEEE Trans. Systems Man Cybernet. 23 (3) (1993) 665–685.
[88] C.Z. Janikow, Fuzzy decision trees: issues and methods, IEEE Trans. Systems Man Cybernet. Part B 28 (1) (1998) 1–14.
[89] C. Jerome, B. Noel, H. Michel, A new fuzzy clustering technique based on PDF estimation, in: Proc. Information Processing and Managing of Uncertainty (IPMU'2002), 2002, pp. 225–232.
[90] F. Jimenez, G. Sanchez, A.F. Gomez-Skarmeta, H. Roubos, R. Babuska, Fuzzy modelling with multi-objective neuro-evolutionary algorithms, in: Proc. of IEEE Internat. Conf. on Systems, Man and Cybernetics (IEEE SMC'02), Yasmine Hammamet, Tunisia, 2002, pp. 253–258.
[91] Y. Jin, Fuzzy modelling of high-dimensional systems: complexity reduction and interpretability improvement, IEEE Trans. Fuzzy Systems 8 (2) (2000) 212–220.
[92] Y. Jin, W. Seelen, B. Sendhoff, On generating FC^3 fuzzy rule systems from data using evolution strategies, IEEE Trans. Systems Man Cybernet. Part B 29 (6) (1999) 829–845.
[93] T.A. Johansen, R. Babuska, Multi-objective identification of Takagi–Sugeno fuzzy models, IEEE Trans. Fuzzy Systems 11 (6) (2003) 847–860.
[94] T.A. Johansen, R. Shorten, R. Murray-Smith, On the interpretation and identification of dynamic Takagi–Sugeno fuzzy models, IEEE Trans. Fuzzy Systems 8 (3) (2000) 297–313.
[95] R.I. John, P. Innocent, Modelling uncertainty in clinical diagnosis using fuzzy logic, IEEE Trans. System Man Cybernet. Part B: Cybernet. 35 (5) (2005) 1340–1350.
[96] S.B. Jorgensen, K.M. Hangos, Grey box modelling for control: qualitative models as a unifying framework, Internat. J. Adaptive Control Signal Process. 9 (6) (1995) 547–562.
[97] J.S. Kandola, Interpretable modelling with sparse kernels, Ph.D. Thesis, Department of Electronics and Computer Science, University of Southampton, 2001.
[98] P.P. Kanjilal, D.N. Banerjee, On the application of orthogonal transformation for the design and analysis of feedforward networks, IEEE Trans. Neural Networks 6 (5) (1995) 1061–1070.
[99] N. Karayiannis, G.W. Mi, Growing radial basis neural networks: merging supervised and unsupervised learning with network growth techniques, IEEE Trans. Neural Networks 8 (6) (1997) 1492–1506.
[100] T. Kavli, ASMOD—an algorithm for adaptive spline modelling of observation data, Internat. J. Control 58 (4) (1993) 947–967.
[101] E. Kim, M. Park, S. Kim, M. Park, A transformed input-domain approach to fuzzy modelling, IEEE Trans. Fuzzy Systems (6) (1998) 596–604.
[102] S.W. Kim, B.J. Oommen, On utilizing search methods to select subspace dimensions for kernel-based nonlinear subspace classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 27 (1) (2005) 136–141.
[103] A. Klose, A. Nürnberger, Applying Boolean transformations to fuzzy rule bases, in: Proc. Seventh European Congress on Intelligent Techniques and Soft Computing (EUFIT'99), Verlag Mainz, Aachen, Germany, 1999, p. 6.
[104] A. Klose, A. Nürnberger, D. Nauck, Some approaches to improve the interpretability of neuro-fuzzy classifiers, in: Proc. Sixth European Congress on Intelligent Techniques and Soft Computing (EUFIT'98), Verlag und Druck Mainz GmbH, Aachen, 1998, pp. 629–633.
[105] A. Klose, A. Nürnberger, D. Nauck, Improved NEFCLASS pruning techniques applied to a real world domain, in: G. Krell, B. Michaelis, D. Nauck, R. Kruse (Eds.), Proceedings of Neural Networks in Applications (NN'99), Logisch GmbH, Magdeburg, Germany, 1999, pp. 47–52.
[106] T. Kohonen, Self-organization and Associative Memory, third ed., Springer, Berlin, 1989.
[107] S. Kullback, Information Theory and Statistics, Wiley, New York, 1959.
[108] M. Kumar, R. Stoll, N. Stoll, A robust design criterion for interpretable fuzzy models with uncertain data, IEEE Trans. Fuzzy Systems 14 (2) (2006) 314–328.

[109] A. Laha, N.R. Pal, Some novel classifiers designed using prototypes extracted by a new scheme based on self-organizing feature map, IEEE Trans. Systems Man Cybernet. and Cybernet. B 31 (6) (2001) 881–890.
[110] M.S. Lewicki, T.J. Sejnowski, Learning overcomplete representations, Neural Comput. 12 (2) (2000) 337–365.
[112] P. Lindskog, Fuzzy identification from a grey box modelling point of view, in: H. Hellendoorn, D. Driankov (Eds.), Fuzzy Model Identification, Springer, Berlin, 1997, pp. 3–50.
[113] A. Lotfi, H.C. Andersen, A.C. Tsoi, Interpretation preservation of adaptive fuzzy inference systems, Internat. J. Approx. Reason. 15 (4) (1996) 379–394.
[114] E.H. Mamdani, Applications of fuzzy algorithm for control a simple dynamic plant, Proc. IEEE 121 (12) (1974) 1585–1588.
[115] E.H. Mamdani, Advances in the linguistic synthesis of fuzzy controllers, Internat. J. Man Mach. Stud. 8 (6) (1976) 669–678.
[116] E.H. Mamdani, Application of fuzzy logic to approximate reasoning using linguistic systems, IEEE Trans. Comput. 26 (12) (1977) 1182–1191.
[117] P.A. Mastorocostas, J.B. Theocharis, V.S. Petridis, A constrained orthogonal least-squares method for generating TSK fuzzy models: application to short-term load forecasting, Fuzzy Sets and Systems 118 (2) (2001) 215–233.
[118] C. Mencar, G. Castellano, A.M. Fanelli, Distinguishability quantification of fuzzy sets, Inform. Sci. 177 (2007) 130–149.
[119] C. Mencar, G. Castellano, A.M. Fanelli, A. Bargiela, Similarity vs. possibility in measuring fuzzy sets distinguishability, in: Proc. Fifth Internat. Conf. on Recent Advances in Soft Computing (RASC), Nottingham, UK, 2004, pp. 354–359.
[120] J.M. Mendel, R.I. John, F. Liu, Interval type-2 fuzzy logic systems made simple, IEEE Trans. Fuzzy Systems 14 (6) (2006) 808–821.
[121] R. Mikut, J. Jäkel, L. Gröll, Interpretability issues in data-based learning of fuzzy systems, Fuzzy Sets and Systems 150 (2) (2005) 179–197.
[122] G.C. Mouzouris, J.M. Mendel, Designing fuzzy logic systems for uncertain environments using a singular-value-QR decomposition method, in: Proc. Fifth IEEE Internat. Conf. on Fuzzy Systems, New Orleans, LA, 1996, pp. 295–301.
[123] G.C. Mouzouris, J.M. Mendel, A singular-value-QR decomposition based method for training fuzzy logic systems in uncertain environments, J. Intelligent and Fuzzy Systems 5 (1997) 367–374.
[124] D. D. Nauck, Measuring interpretability in rule-based classification systems, Proc. IEEE Internat. Conf. on Fuzzy Systems (FUZZ-IEEE'03), St. Louis, Missouri, USA, 2003, pp. 196–201.
[125] D. Nauck, R. Kruse, A neuro-fuzzy method to learn fuzzy classification rules from data, Fuzzy Sets and Systems 89 (1997) 277–288.
[126] D. Nauck, R. Kruse, A neuro-fuzzy approach to obtain interpretable fuzzy systems for function approximation, in: Proc. IEEE Internat. Conf. on Fuzzy Systems (FUZZ-IEEE), Anchorage, AK, May 4–9, 1998, pp. 1106–1111.
[127] D. Nauck, R. Kruse, Obtaining interpretable fuzzy classification rules from medical data, Artificial Intelligence in Medicine 16 (1999) 149–169.
[128] O. Nelles, Nonlinear system identification with local linear neuro-fuzzy models, Ph.D. Thesis, Darmstadt University, Darmstadt, Germany, 1999.
[129] O. Nelles, A. Fink, R. Babuska, M. Setnes, Comparison of two construction algorithms for Takagi-Sugeo fuzzy models, Internat. J. Appl. Math. Comput. Sci. 10 (4) (2000) 835–855.
[130] O. Nelles, R. Isermann, Basis function networks for interpolation of local linear models, in: Proc. IEEE Conf. on Decision and Control (CDC), Kobe, Japan, 1996, pp. 470–475.
[131] S. M. Omohundro, The Delaunay triangulation and function learning, Technical Report TR-90-001, International Computer Science Institute, Berkeley, 1990.
[132] R.P. Paiva, A. Dourado, Interpretability and learning in neuro-fuzzy systems, Fuzzy Sets and Systems 147 (1) (2004) 17–38.
[133] N.R. Pal, Soft computing for feature analysis, Fuzzy Sets and Systems 103 (2) (1999) 201–221.
[134] N.R. Pal, S. Chakraborty, Fuzzy rule extraction from ID3-type decision trees for real data, IEEE Trans. Systems Man Cybernet. Part B 31 (5) (2001) 745–754.
[135] W. Pedrycz, J.C. Bezdek, R.J. Hathaway, G.W. Rogers, Two nonparametric models for fusing heterogeneous fuzzy data, IEEE Trans. Fuzzy Systems 6 (3) (1998) 411–425.
[136] W. Pedrycz, J.V. de Oliveira, Optimization of fuzzy models, IEEE Trans. Systems Man Cybernet. Part B 26 (4) (1996) 627–636.
[137] C.A. Pena-Reyes, M. Sipper, A fuzzy-genetic approach to breast cancer diagnosis, Artificial Intelligence in Medicine 17 (2) (1999) 131–155.
[138] C.A. Pena-Reyes, M. Sipper, Fuzzy CoCo: a cooperative-coevolutionary approach to fuzzy modelling, IEEE Trans. Fuzzy Systems 9 (5) (2001) 727–737.
[139] C.A. Pena-Reyes, M. Sipper, Fuzzy CoCo: balancing accuracy and interpretability of fuzzy models by means of coevolution, in: J. Casillas, O. Cordón, F. Herrera, L. Magdalena (Eds.), Accuracy Improvements in Linguistic Fuzzy Modelling Studies in Fuzziness and Soft Computing, Vol. 129, Springer, Berlin, 2003.
[140] T.A. Plate, Accuracy versus interpretability in flexible modelling: implementing a tradeoff using Gaussian process models, Behaviourmetrika Special Issue on Interpreting Neural Network Models 26 (1999) 29–50.
[141] R. Quinlan, Induction of decision trees, Mach. Learning 1 (1986) 81–106.
[142] A. Riid, Transparent fuzzy systems: modelling and control, Ph.D. Dissertation, Department of Computer Control, Tallinn Technical University, Estonia, 2002.
[143] A. Riid, R. Isotamm, E. Rustern, Transparent analysis of first-order Takagi–Sugeno systems, in: Proc. 10th Internat. Symp. on System-Modelling-Control, Zakopane, Poland, 2001, pp. 165–170.
[144] A. Riid, E. Rustern, Transparent fuzzy systems and modelling with transparency protection, in: Proc. IFAC Symp. on Artificial Intelligence in Real Time Control, Budapest, 2000, pp. 229–234.
[145] J. Rissanen, Modeling by shortest data description, Automatica 14 (1978) 465–471.
[146] I. Rojas, H. Pomares, J. Ortega, A. Prieto, Self-organized fuzzy system generation from training examples, IEEE Trans. Fuzzy Systems 8 (1) (2000) 23–26.

[147] H. Roubos, M. Setnes, Compact and transparent fuzzy models and classifiers through iterative complexity reduction, IEEE Trans. Fuzzy Systems 9 (4) (2001) 516–524.
[148] J.A. Roubos, R. Babuska, Comments on the benchmarks in 'a proposal for improving the accuracy of linguistic modelling' and related articles, IEEE Trans. Fuzzy Systems 11 (6) (2003) 861–865.
[149] J.A. Roubos, M. Setnes, J. Abonyi, Learning fuzzy classification rules from labeled data, Inform. Sci. 150 (1–2) (2003) 77–93.
[150] A. Ruiz, P.E. López-de-Teruel, Nonlinear kernel-based statistical pattern analysis, IEEE Trans. Neural Networks 12 (1) (2001) 16–32.
[151] B. Schölkopf, S. Mika, A. Smola, G. Ratsch, K.-R. Müller, Kernel PCA pattern reconstruction via approximate pre-images, in: L. Niklasson, M. Boden, T. Ziemke (Eds.), Proc. Eighth Internat. Conf. on Artificial Neural Networks, Perspectives in Neural Computing, Springer, Berlin, 1998, pp. 147–152.
[152] B. Schölkopf, A.J. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation 10 (5) (1998) 1299–1319.
[153] W. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[154] H.P. Schwefel, Evolution and Optimum Seeking, Wiley, New York, 1995.
[155] M. Setnes, R. Babuska, Rule base reduction: some comments on the use of orthogonal transforms, IEEE Trans. Systems Man Cybernet. Part C 31 (2) (2001) 199–206.
[156] M. Setnes, R. Babuska, U. Kaymak, H.R. van Nauta Lemke, Similarity measures in fuzzy rule base simplification, IEEE Trans. Systems Man Cybernet. Part B 28 (3) (1998) 376–386.
[157] M. Setnes, R. Babuska, H.B. Verbruggen, Rule-based modelling: precision and transparency, IEEE Trans. Systems Man Cybernet. Part C 28 (1) (1998) 165–169.
[158] M. Setnes, H. Hellendoorn, Orthogonal transforms for ordering and reduction of fuzzy rules, Proc. Ninth IEEE Internat. Conf. on Fuzzy Systems, San Antonio, USA, 2000, pp. 700–705.
[159] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, M. Anthony, Structural risk minimization over data-dependent hierarchies, IEEE Trans. Inform. Theory 44 (5) (1998) 1926–1940.
[160] Q. Shen, J.G. Marin-Blazquez, Microtuning of membership functions: accuracy vs. interpretability, in: Proc. IEEE Internat. Conf. on Fuzzy Systems (FUZZ-IEEE'02), Honolulu, Hawai, USA, 2002, pp. 168–173.
[161] R. Silipo, M.R. Berthold, Input features' impact on fuzzy decision processes, IEEE Trans. Systems Man Cybernet. Part B 30 (6) (2000) 821–834.
[163] M. Stitson, A. Gammerman, V. Vapnik, V. Vovk, C. Watkins, J. Weston, Support vector regression with ANOVA decomposition kernels, in: B. Scholkopf, C.J.C. Burges, A.J. Smola (Eds.), Advances in Kernel Methods—Support Vector Learning, MIT Press, Cambridge, MA, 1999, pp. 285–292.
[164] C.J. Stone, M. Hansen, C. Kooperberg, Y.K. Truong, Polynomial splines and their tensor products in extended linear modelling, Ann. Statist. 25 (4) (1997) 1371–1470.
[165] M. Sugeno, G.T. Kang, Structure identification of fuzzy model, Fuzzy Sets and Systems 28 (1) (1988) 15–33.
[166] M. Sugeno, T. Yasukawa, A fuzzy-logic-based approach to qualitative modelling, IEEE Trans. Fuzzy Systems 1 (1) (1993) 7–31.
[167] T. Suzuki, T. Furuhashi, Conciseness of fuzzy models, in: J. Casillas, O. Cordón, F. Herrera, L. Magdalena (Eds.), Interpretability Issues in Fuzzy Modelling, Studies in Fuzziness and Soft Computing, Vol. 128, Springer, Heidelberg, 2003.
[168] T. Takagi, M. Sugeno, Fuzzy identification of systems and its applications to modelling and control, IEEE Trans. Systems Man Cybernet. 15 (1) (1985) 116–132.
[169] M. Taniguchi, On the selection of the order of the spectral density of a stationary process, Ann. Instit. Statist. Math. 32 (1980) 401–409.
[170] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Statist. Soc. B 58 (1) (1996) 267–288.
[171] A.N. Tikhonov, Solution of incorrectly formulated problems and the regularization method, Soviet Math. Dokl. 4 (1963) 1035–1038.
[172] A.N. Tikhonov, V.Y. Arsenin, Solutions of Ill-Posed Problems, Wiley, New York, 1977.
[173] J.T. Tou, R.C. Gonzales, Pattern Recognition Principles, Addison-Wesley, Reading, MA, 1974.
[174] E. Van Broekhoven, V. Adriaenssens, B. De Baets, Interpretability-preserving genetic optimization of linguistic terms in fuzzy models for fuzzy ordered classification: an ecological case study, Internat. J. Approx. Reason. 44 (1) (2007) 65–90.
[175] V. Vapnik, The Nature of Statistical Learning, Springer, New York, 1995.
[176] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[177] G. Wahba, Spline Models for Observational Data, SIAM, Philadelphia, 1990.
[178] G. Wahba, Y. Wang, C. Gu, R. Klein, B. Klein, Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy, Ann. Statist. 23 (16) (1995) 1865–1895.
[179] H. Wang, S. Kwong, Y. Jin, W. Wei, K. Man, Agent-based evolutionary approach to interpretable rule-based knowledge extraction, IEEE Trans. Systems Man Cybernet. Part C 29 (2) (2005) 143–155.
[180] H. Wang, S. Kwong, Y. Jin, W. Wei, K.F. Man, Multi-objective hierarchical genetic algorithm for interpretable fuzzy rule-based knowledge extraction, Fuzzy Sets and Systems 149 (1) (2005) 149–186.
[181] L. Wang, R. Langari, Building Sugeno-type models using fuzzy discretization and orthogonal parameter estimation techniques, IEEE Trans. Fuzzy Systems 3 (4) (1995) 454–458.
[182] L. Wang, G. Libert, Combining pattern recognition techniques with Akaike's information criteria for identifying ARMA models, IEEE Trans. Signal Process. 42 (1994) 1388–1396.
[183] L.X. Wang, J.M. Mendel, Fuzzy basis functions, universal approximation, and orthogonal least squares learning, IEEE Trans. Neural Networks 3 (5) (1992) 807–814.
[184] X.-Z. Wang, D.S. Yeung, E.C.C. Tsang, A comparative study on heuristic algorithms for generating fuzzy decision trees, IEEE Trans. Systems Man Cybernet. Part B 31 (2) (2001) 215–226.

[185] J. Weston, C. Watkins, Multi-class support vector machines. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK, 1998.
[186] J. Wyatt, Nervous about artificial neural networks? (commentary), Lancet 346 (8984) (1995) 1175–1177.
[187] R.R. Yager, D.P. Filev, Unified structure and parameter identification of fuzzy models, IEEE Trans. Systems Man Cybernet. 23 (4) (1993) 1198–1205.
[188] R.R. Yager, D.P. Filev, Approximate clustering via the mountain method, IEEE Trans. Systems Man Cybernet. 24 (1994) 1279–1284.
[189] J. Yen, L. Wang, Application of statistical information criteria for optimal fuzzy model construction, IEEE Trans. Fuzzy Systems 6 (3) (1998) 362–372.
[190] J. Yen, L. Wang, C.W. Gillespie, Improving the interpretability of TSK fuzzy models by combining global learning and local learning, IEEE Trans. Fuzzy Systems 6 (4) (1998) 530–537.
[191] L.A. Zadeh, Fuzzy sets, Inform. Control 8 (1965) 338–353.
[192] L.A. Zadeh, Toward a theory of fuzzy systems, in: Aspects of Network and System Theory, Rinehart and Winston, New York, NY, USA, 1971, pp. 469–490.
[193] L.A. Zadeh, Outline of a new approach to the analysis of complex systems and decision processes, IEEE Trans. Systems Man Cybernet. 3 (1) (1973) 28–44.
[194] L.A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning—1, Inform. Sci. 8 (1975) 199–249.
[195] L.A. Zadeh, Fuzzy logic and the calculi of fuzzy rules and fuzzy graphs: a précis, Internat. J. Multiple-Valued Logic 1 (1996) 1–38.
[196] H. Zhang, G. Wahba, Y. Lin, M. Voelker, M. Ferris, R. Klein, B. Klein, Variable selection and model building via likelihood basis pursuit, J. Amer. Statist. Assoc. 99 (467) (2004) 659–672.
[197] S.-M. Zhou, J.Q. Gan, An unsupervised kernel based fuzzy c-means clustering algorithm with kernel normalization, Internat. J. Comput. Intell. Appl. 4 (4) (2004) 355–373.
[198] S.-M. Zhou, J.Q. Gan, Constructing accurate and parsimonious fuzzy models with distinguishable fuzzy sets based on an entropy measure, Fuzzy Sets and Systems 157 (8) (2006) 1057–1074.
[199] S.-M. Zhou, J.Q. Gan, Constructing parsimonious fuzzy classifiers based on L2-SVM in high-dimensional space with automatic model selection and fuzzy rule ranking, IEEE Trans. on Fuzzy Systems 15 (3) (2007) 398–409.
[200] R. Zwick, E. Carlstein, D.V. Budescu, Measures of similarity among fuzzy concepts: a comparative analysis, Internat. J. Approx. Reason. 1 (2) (1987) 221–242.