ARTICLE IN PRESS

Knowledge-Based Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A fuzzy inference system modeling approach for interval-valued symbolic data forecasting

Leandro Maciel [a,*], Rosangela Ballini [b]

[a] São Paulo School of Politics, Economics and Business, Federal University of São Paulo, R. Angélica 100, Osasco, São Paulo, 06132–380, Brazil
[b] Institute of Economics, University of Campinas, R. Pitágoras 353, Campinas, 13083–857, Brazil

ARTICLE INFO

Article history:
Received 28 February 2018
Received in revised form 19 October 2018
Accepted 21 October 2018
Available online xxxx

Keywords:
Symbolic data analysis
Interval-valued data
Fuzzy inference systems
Rule-based models
Time series forecasting

ABSTRACT

This paper suggests a fuzzy inference system (iFIS) modeling approach for interval-valued time series forecasting. Interval-valued data arise quite naturally in many situations in which such data represent uncertainty/variability or when comprehensive ways to summarize large data sets are required. The method comprises a fuzzy rule-based framework with affine consequents which provides a (non)linear framework that processes interval-valued symbolic data. The iFIS antecedents identification uses a fuzzy c-means clustering algorithm for interval-valued data with adaptive distances, whereas parameters of the linear consequents are estimated with a center-range methodology to fit a linear regression model to symbolic interval data. iFIS forecasting power, measured by accuracy metrics and statistical tests, was evaluated through Monte Carlo experiments using both synthetic interval-valued time series with linear and chaotic dynamics, and real financial interval-valued time series. The results indicate a superior performance of iFIS compared to traditional alternative single-valued and interval-valued forecasting models by reducing 19% on average the predicting errors, indicating that the suggested approach can be considered as a promising tool for interval time series forecasting.

1. Introduction

The increasing development of automated systems, small-scale computing devices, sensor networks, and data capture technologies has contributed to the production of large volumes of data. The significant growth in the use of databases resulted in the need to discover regularities and summarize information stored in large data sets [1–3]. In this scenario, extracting valuable information from a variety of sources plays a crucial role in data management and its related decision-making processes. Besides the highly advanced computing power observed recently, it is sometimes not practical to analyze very large data sets. To alleviate this issue, huge data sets can be aggregated by lists, intervals, histograms, frequency distributions, among others, in which, despite their summarization properties, are also able to take into account variability and/or uncertainty inherent to the data. These kinds of data have been mainly studied in the field of *Symbolic Data Analysis* (SDA) [4,5], a new domain related to multivariate analysis, pattern recognition and artificial intelligence for extending classical exploratory data analysis and statistical methods to symbolic data.

The aim of SDA is to provide suitable methods (such as clustering, factorial techniques and decision trees) to manage aggregated data described by multi-valued variables, thus providing a comprehensive way to summarize data sets by means of symbolic data. This results in a smaller and more manageable data set which preserves the essential information, and is subsequently analyzed by means of generalizing the exploratory data analysis and data mining techniques to symbolic data [6,4,7].

The literature on SDA have discussed the management of large databases by focusing units described by variables that are often categorical to construct categories whose extents are units of the database. The reification of categories in concepts yields symbolic data tables, where the units are the concepts and the variables' values are symbolic. From cognitive sciences, SDA can enhance knowledge discovery and data mining in computer sciences [8]. [9] give an overview of the works on the development in SDA by presenting tools and methods designed to deal with symbolic data. A recent discussion regarding SDA is also provided by [10], explaining how it extends the classical data models to take into account more complete and complex information, and discussing some methods for the (multivariate) analysis of symbolic data.

In the context of massive data sets availability, recommender systems and information filtering also play an important role as the SDA do. The task of recommender systems is to turn data on users and their preferences into predictions of users' possible future likes and interests. The study of recommender systems

* Corresponding author.
E-mail addresses: maciel.leandro@unifesp.br (L. Maciel), ballini@unicamp.br (R. Ballini).

https://doi.org/10.1016/j.knosys.2018.10.033
0950-7051/

is at crossroads of science and socio-economic life and its huge potential was first noticed by web entrepreneurs in the forefront of the information revolution. [11] review recent developments in recommender systems and compare and evaluate available algorithms. Regarding economic big data, [12] recently discussed how big data reveal the status of economic development and suggested applications of different types of big data on quantifying macro-economic structures and micro-social status.

According to SDA, in real-life situations imprecise data occur when a collection associated with each unit under analysis represents the uncertain value of the record, or when data presenting variability emerge e.g. due to the aggregation of single observations.[1] Therefore, symbolic variables are better suited than single-valued variables to describe these complex situations. Interval-valued data are a particular kind of symbolic data which arise in practical situations such as recording meteorological data stations [13], daily low and high asset prices [14–16], blood pressure [5], income level in survey data [17], influence of height on salary expectations [18], image processing [19], financial assets volatility forecasting [20], power load and generation, traffic flows, etc. Interval data are also relevant in the case of confidential data applications in companies and government specific areas in which only ranges of values are permitted to be shown.[2] Therefore, tools for interval-valued symbolic data analysis are demanding.

Different approaches have been introduced in the literature to analyze interval-valued data. For instance, [7] suggested central tendency and dispersion measures that are suitable for interval-valued data. Principal component analysis methods designed for interval-valued data were proposed by [21,22] and [23]. Concerning supervised classification, [24] introduced a symbolic classifier as a region-oriented approach for interval-valued data.

Researchers have also considered regression approaches for interval-valued data. [5] first introduced a regression approach, which fits a linear regression model on the center of the intervals and applies the fitted model to the lower and the upper bounds of the predictor variables to obtain a prediction. Extending this approach to the range of intervals, [25] suggested a regression method that fits two separate linear regression models on the center and the range of the intervals. Based on this idea, [26] introduced a bivariate approach which fits two regression models on both the center and the range of the intervals simultaneously as the predictors. A linear model to analyze interval-valued data was also proposed by [27], based on the bivariate generalized linear model by [28]. In [29], a method that fits a constrained linear regression model on the center and range of the interval values is suggested. [17] proposed a symbolic covariance method based on the symbolic sample covariance introduced by [26]. Related works concerning regression approaches for symbolic interval-valued data include [30–32,6,33].

SDA also provides a number of clustering methods for interval-valued data. These methods differ in the type of the considered symbolic data, in their cluster structures and/or in the considered clustering criteria [34]. The literature has addressed hard clustering methods for interval-values data, broadly divided into hierarchical [35,36] and non-hierarchical approaches [37,38]. A number of authors have also analyzed the problem of fuzzy clustering for symbolic data [39,40]. Fuzzy clustering generates a fuzzy partition based on the idea of partial membership expressed by the degree of membership of each pattern in a given cluster. For instance, [34] suggests adaptive and non-adaptive fuzzy c-means clustering methods for partitioning symbolic interval data. Partitioning fuzzy k-means clustering models for interval-valued data were also introduced by [41]. A possibilistic fuzzy c-means clustering algorithm is addressed in [42]. Symbolic fuzzy classification using a fuzzy radial basis function network and self-organizing maps are proposed in [43] and [44,45], respectively. More recently, [46] suggested a fuzzy clustering algorithm for interval-valued data based on the concept of participatory learning.

When considering a chronological sequence of interval-valued data, interval time series (ITS) arises quite naturally. The tools for ITS data analysis are very compelling and have been considered in the SDA framework. Modeling and forecasting of ITS has the advantage of taking into account the variability and/or uncertainty, and it reduces the amount of random variations relative to that found in classic single-valued time series [47]. The literature on the methodologies for ITS forecasting fall roughly into two categories in relation to the method in which interval data is handled, i.e., splitting single-valued methods or interval-valued methods [47]. In the first category, the lower and upper bounds of interval data are treated as two independent single-valued variables, such as in the works of [48–50]. On the other hand, in the second category the lower and upper bounds of interval data are treated using interval arithmetic as interval-valued data, as in the interval Holts exponential smoothing method (Holt^I) [51], the vector autoregression/vector error correction model [52,53], multilayer perceptron (MLP) [51], interval MLP (iMLP) [54], the complex-valued radial basis function neural network [47], and the hybrid Holt^I multi-output support vector regression model [14]. A comprehensive literature review of the presented methodologies and techniques employed for ITS forecasting can be found in [14,55].

This paper comes within the framework of ITS modeling and forecasting by proposing a fuzzy inference system (FIS) designed to process symbolic interval-valued data, namely iFIS. FISs have been successfully applied in fields such as automatic control, data classification, decision analysis, expert systems, and computer vision [56–60]. In this paper, the iFIS assumes a functional fuzzy rule-based model structure, where the consequent part is expressed as a (non)linear relationship between the input variables and the output variable [61]. The identification of a iFIS concerns the identification of the antecedents and consequents of the fuzzy rules. Rules antecedents are identified using the fuzzy c-means clustering approach for symbolic interval-valued data using adaptive quadratic Euclidian distances, proposed by [34]. The advantage of adaptive distances is that the clustering algorithm is able to recognize clusters of different shapes and sizes. Finally, consequent parameters of the rules are estimated using the least squares algorithm designed for interval-valued data, as suggested in [27]. Computational experiments comprise the prediction of synthetic ITS with linear and chaotic dynamics and real interval stock price time series data forecasting. iFIS prediction performance is compared with traditional single-valued time series methods and interval-valued forecasting approaches in terms of accuracy measures and statistical tests.

In summary, the novelty and contributions of this work can be outlined as follows. First, we build on the literature on SDA by suggesting a fuzzy inference system modeling approach designed to process interval-valued symbolic data naturally, which is a new based for SDA by incorporating a fuzzy clustering and regression mechanisms in an inference framework to identify fuzzy rules antecedents and consequents, respectively. Second, the possibility of forecasting the lower and upper bounds of ITS simultaneously using the proposed approach is examined based on simulated ITS data with different dynamics and on real interval stock price time series data. Third, the suggested approach provides a non-linear mechanism to forecast ITS which, due to its fuzzy nature, is also able to account for imprecise data and vagueness, mainly when financial data is concerned. Finally, the iFIS concerns an ITS forecasting modeling framework that enhances interpretability,

[1] Variability in data occur when each unit represents a specific group/class, or when values express characteristics that float along a period of time.

[2] Interval-valued data are either inherently observed as or processed to be intervals.

providing a linguistic model which could be employed by decision makers in many real-world applications.

Following this introduction, the paper is outlined as follows. Section 2 details the interval-valued fuzzy inference system modeling framework. Experimental results on forecasting simulated ITS data and real interval stock market time series data are discussed in Section 3. Finally, concluding remarks are provided in Section 4.

2. Fuzzy inference system for interval-valued data

The overall formulation of the suggested interval fuzzy inference system (iFIS) for interval-valued data forecasting is detailed in this section. The definition of interval-valued data and interval time series (ITS) are presented first. Further, the iFIS model structure and its identification are described.

2.1. Interval-valued data and ITS

An interval-valued variable $\mathbf{x}$ is defined as a closed bounded set of real numbers in the form:

$$\mathbf{x} = [x^L, x^U]^T \in \Im, \tag{1}$$

where $\Im = \{[x^L, x^U]^T : x^L, x^U \in \Re, x^L \leq x^U\}$ is the set of closed intervals of the real line $\Re$, x^L and x^U are the lower and upper bounds of interval $\mathbf{x}$, respectively, and with the superscript T denoting the transpose of the vector.

The center or mid-point of an interval-valued variable $\mathbf{x}$, denoted as x^c, is calculated as:

$$x^c = \frac{x^L + x^U}{2}. \tag{2}$$

Similarly, the range or half-length of $\mathbf{x}$, x^r, is:

$$x^r = \frac{x^U - x^L}{2}. \tag{3}$$

Notice that, if the center x^c and range x^r of an interval-valued data $\mathbf{x}$ are known, its lower and upper bounds can be recovered as $x^L = x^c - x^r$ and $x^U = x^c + x^r$, respectively.

An interval-valued time series (ITS) is a sequence of interval-valued variables observed in consecutive time steps t ($t = 1, 2, \ldots, n$) expressed as a two-dimensional vector $\mathbf{x}_t = [x_t^L, x_t^U]^T \in \Im$, where n denotes the number of intervals in the time series, i.e. the sample size.

In order to process interval-valued variables by the iFIS modeling framework, the basic concepts of interval arithmetic must be addressed. Interval arithmetic extends traditional arithmetic to operate on intervals. This work uses the arithmetic operations introduced by [62]:

$$\begin{aligned}
\mathbf{x} + \mathbf{y} &= [x^L + y^L, x^U + y^U]^T, \\
\mathbf{x} - \mathbf{y} &= [x^L - y^U, x^U - y^L]^T, \\
\mathbf{xy} &= \left[\min\{x^L y^L, x^L y^U, x^U y^L, x^U y^U\},\right. \\
&\quad \left.\max\{x^L y^L, x^L y^U, x^U y^L, x^U y^U\}\right]^T, \\
\mathbf{x}/\mathbf{y} &= \mathbf{x}\,(1/\mathbf{y})\,, \text{ with } 1/\mathbf{y} = [1/y^U, 1/y^L]^T.
\end{aligned} \tag{4}$$

Interval arithmetic subsumes classic arithmetic. This means that if an operation of interval arithmetic takes real numbers as operands, considering them as intervals of length zero, then we obtain the same result as if the operation were performed using traditional arithmetic.

2.2. iFIS model structure

The interval-valued fuzzy inference system (iFIS) modeling approach is formed by a set of functional fuzzy rules of the following form:

$$\mathcal{R}_i : \text{ IF } X \text{ is } \mu_i \text{ THEN } \mathbf{y}_i = f\,(X, \boldsymbol{\beta})\,, \tag{5}$$

where $\mathcal{R}_i$ is the ith fuzzy rule, $i = 1, 2, \ldots, c$, c is the number of fuzzy rules in the rule base, $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]^T$, $\mathbf{x}_j = [x_j^L, x_j^U]^T \in \Im$, $j = 1, \ldots, p$, is the input comprised by p interval-valued variables, μ_i is the fuzzy set of the antecedent of the ith fuzzy rule whose membership function is $\mu_i(X) : \Im^p \to [0, 1]$, $\mathbf{y}_i = [y_i^L, y_i^U]^T \in \Im$ is the output of the ith rule, $f(\cdot)$ represents an interval-valued affine function, and $\boldsymbol{\beta}$ is the matrix of real-valued parameters of the consequent of the ith rule.

Fuzzy inference using iFIS rules in Eq. (5) is similar to the classic Takagi–Sugeno [61] model counterpart except that the arithmetic operations are the interval operations instead of the usual operations with real numbers. The output of the inference system is computed as:

$$\mathbf{y} = \sum_{i=1}^{c} \left(\frac{\mu_i(X)\mathbf{y}_i}{\sum_{j=1}^{c} \mu_j(X)} \right). \tag{6}$$

The expression in (6) can be rewritten using normalized degrees of activation:

$$\mathbf{y} = \sum_{i=1}^{c} \lambda_i \mathbf{y}_i, \tag{7}$$

where

$$\lambda_i = \frac{\mu_i(X)}{\sum_{j=1}^{c} \mu_j(X)}, \tag{8}$$

is the normalized firing level of the ith rule.

The TS model uses parameterized fuzzy regions and associates each region with a local affine model. The nonlinear nature of the rule-based model emerges from the fuzzy weighted combination of the collection of the multiple local affine models. The contribution of a local model to the model output is proportional to its degree of activation.

Identifying the iFIS comprises two tasks: (i) learning the antecedent part of the model with a fuzzy clustering algorithm for interval-valued data, and (ii) estimation of the parameters of the affine consequents. The ith fuzzy cluster defines μ_i, the antecedent of the ith fuzzy rule. The cluster structure defines the structure of the model itself given that to each cluster there is a correspondent fuzzy rule whose consequent is an interval affine function, i.e. an affine local model. iFIS learning is described as follows.

2.3. iFIS antecedents identification

The identification of iFIS antecedents uses the adaptive fuzzy c-means clustering algorithm for interval-valued data (IFCM), proposed by [34], which concerns an extension of the classical fuzzy c-means clustering algorithm [63] for symbolic interval data. In this approach, the adequacy criterion is based on a suitable adaptive squared Euclidean distance. The main idea is that a different distance is associated to each cluster to compare clusters and their representatives that change at each iteration, i.e., the distance is not definitively determined and differs from one class to another. The advantage of these adaptive distances is that the clustering algorithm is able to find clusters of different shapes and sizes [34].

Let $\Omega = \{1, \ldots, n\}$ be a set of n patterns (each pattern is indexed by t) describing p symbolic interval variables $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ (each variable is indexed by j). Each pattern t is represented as a

Please cite this article in press as: L. Maciel, R. Ballini, A fuzzy inference system modeling approach for interval-valued symbolic data forecasting, Knowledge-Based Systems (2018), https://doi.org/10.1016/j.knosys.2018.10.033.

vector of intervals $X = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$, where $\mathbf{x}_j = [x_j^L, x_j^U]^T \in \Im$. Also, each prototype $\mathbf{v}_i$ of cluster i, $i = 1, \ldots, c$, is represented as a vector of intervals $V_i = [\mathbf{v}_{i1}, \ldots, \mathbf{v}_{ip}]$, where $\mathbf{v}_{ij} = [v_{ij}^L, v_{ij}^U]^T \in \Im$, $j = 1, \ldots, p$.

As in the standard fuzzy c-means algorithm [63], IFCM aims to find a fuzzy partition of a set of patterns in c clusters and a corresponding set of prototypes $\{V_1, \ldots, V_c\}$ that minimizes a W criteria measuring the fit of the clusters to their representatives (prototypes). W is based on an adaptive squared Euclidean distance for each cluster and is defined as:

$$W = \sum_{i=1}^{c}\sum_{t=1}^{n}(\mu_{it})^m \phi_i(X_t, V_i),$$
$$= \sum_{i=1}^{c}\sum_{t=1}^{n}(\mu_{it})^m \sum_{j=1}^{p}\gamma_{ij}\left[(x_j^L - v_{ij}^L)^2 + (x_j^U - v_{ij}^U)^2\right], \tag{9}$$

where $\phi(\cdot)$ is an adaptive Euclidian distance that accesses the dissimilarity between a pair of vectors of intervals defined for each class and parameterized by the vectors of weights $\boldsymbol{\gamma}_i = [\gamma_{i1}, \ldots, \gamma_{ip}]$, $X_t = [\mathbf{x}_{1t}, \ldots, \mathbf{x}_{pt}]$ is a vector of intervals describing the tth pattern, $V_i = [\mathbf{v}_{i1}, \ldots, \mathbf{v}_{ip}]$ is a vector of intervals describing the prototype of cluster i, μ_{it} is the membership degree of pattern t in cluster i, and m is the fuzzification parameter (usually $m = 2$).

The optimal fuzzy partition is obtained by Picard iterations for which the criteria W in (9) is locally minimized. The algorithm starts with an initial membership degree partition and alternates between a representation step and an allocation step until convergence (W reaches a stationary value representing a local minima) [34]. The representation step, which defines the best prototypes and the best distances, has two stages. First, the membership degrees μ_{it} of each pattern t in cluster i and the vector of weights $\boldsymbol{\gamma}_i = [\gamma_{i1}, \ldots, \gamma_{ip}]$ are fixed. Prototypes $V_i = [\mathbf{v}_{i1}, \ldots, \mathbf{v}_{ip}]$, for $i = 1, \ldots, c$ and $j = 1, \ldots, p$, that minimize the clustering criterion W are updated as follows:

$$v_{ij}^L = \frac{\sum_{t=1}^{n}(\mu_{it})^m x_{jt}^L}{\sum_{t=1}^{n}(\mu_{it})^m}, \tag{10}$$

$$v_{ij}^U = \frac{\sum_{t=1}^{n}(\mu_{it})^m x_{jt}^U}{\sum_{t=1}^{n}(\mu_{it})^m}. \tag{11}$$

In the second stage of the representation step, the membership degrees μ_{it} and the prototypes V_i are fixed. The vector of weights $\boldsymbol{\gamma}_i = [\gamma_{i1}, \ldots, \gamma_{ip}]$ minimizing W under $\gamma_{ij} > 0$ and $\prod_{j=1}^{p}\gamma_{ij} = 1$, for $i = 1, \ldots, c$ and $j = 1, \ldots, p$, is updated with the following expression:

$$\gamma_{ij} = \frac{\left\{\prod_{h=1}^{p}\left[\sum_{t=1}^{n}(\mu_{it})^m\left((x_{ht}^L - v_{ih}^L)^2 + (x_{ht}^U - v_{ih}^U)^2\right)\right]\right\}^{\frac{1}{p}}}{\sum_{t=1}^{n}(\mu_{it})^m\left[(x_{jt}^L - v_{ij}^L)^2 + (x_{jt}^U - v_{ij}^U)^2\right]}. \tag{12}$$

Finally, in the allocation step, which defines the best fuzzy partition, the prototypes V_i and the vector of weights $\boldsymbol{\gamma}_i$ are fixed. Thus, the membership degrees μ_{it} which minimize W under $\mu_{it} \geq 0$ and $\sum_{t=1}^{c}\mu_{it} = 1$ are updated as:

$$\mu_{it} = \left[\sum_{h=1}^{c}\left(\frac{\sum_{j=1}^{p}\gamma_{ij}\left[(x_{jt}^L - v_{ij}^L)^2 + (x_{jt}^U - v_{ij}^U)^2\right]}{\sum_{j=1}^{p}\gamma_{hj}\left[(x_{jt}^L - v_{hj}^L)^2 + (x_{jt}^U - v_{hj}^U)^2\right]}\right)^{\frac{1}{m-1}}\right]^{-1}. \tag{13}$$

By fixing the number of clusters c ($2 \leq c < n$), an iteration limit k_{max}, and an error criteria ϵ, the clustering method comprises the iteration between the representation and allocation steps as in the following algorithm. The process produces the vector of cluster prototypes $V_i = [\mathbf{v}_{i1}, \ldots, \mathbf{v}_{ip}]$ and the respective membership degrees μ_{it} of each pattern t with respect to each cluster i, for $t = 1, \ldots, n$ and $i = 1, \ldots, c$, that locally minimize the criterion W. The proof of expressions in (10)–(13) are found in [63] and [34].

2.4. iFIS consequents identification

Once the cluster structure is appropriately defined, i.e. the clusters or fuzzy rules represented by their respective prototypes, the identification of the iFIS requires estimating the consequent parameters of the fuzzy rules. Notice that the iFIS consequents, as defined in (5), concern the fitting of a classical regression model used to predict the values of a dependent quantitative variable in relation to the values of independent quantitative variables for both lower and upper interval bounds. For single-valued variables, the least square method is used to find the optimal parameters by minimizing the sum of the square of residuals. Hence, in order to take into account the variability and/or uncertainty inherent to the data, an approach to fit a linear regression model to interval-valued symbolic data must be considered.

iFIS antecedents identification algorithm

Initialization:
1. Set c, ϵ and k_{max}
2. Initialize μ_{it}, $i = [1, c]$, $j = [1, n]$, $\mu_{it} \geq 0$, $\sum_{t=1}^{c}\mu_{it} = 1$
3. $k = 1$

Representation step:
4. **for** $i = 1, \ldots, c$ **and** $j = 1, \ldots, p$ **do**
5. compute $V_i = [\mathbf{v}_{i1}, \ldots, \mathbf{v}_{ip}]$ using Eqs. (10)–(11)
6. **end for**
7. **for** $i = 1, \ldots, c$ **and** $j = 1, \ldots, p$ **do**
8. update $\boldsymbol{\gamma}_i = [\gamma_{i1}, \ldots, \gamma_{ip}]$ with (12)
9. **end for**

Allocation step:
10. **for** $t = 1, \ldots, n$ **and** $i = 1, \ldots, c$ **do**
11. compute μ_{it} using Eq. (13)
12. **end for**

Stopping criterion:
13. **if** $|W_{k+1} - W_k| \leq \epsilon$ **or** $k > k_{max}$ **then**
14. stop
15. **else**
16. $k = k + 1$ and **return** to step **4**
17. **end if**

Different approaches have been introduced in the literature concerning regression analysis for interval-valued data. [5] suggested a method based on the minimization of the center error, in which the lower and upper bounds of the dependent variable are predicted, respectively, from the lower and upper bounds of the independent variable using the same vector of parameters. On the other hand, [64] proposed a min–max approach, based on the minimization of the errors from two independent linear regression models on the lower and upper bounds of the intervals. This is equivalent to supposing independence between the values of lower and upper bounds of the intervals.

In this paper, iFIS consequents identification adopts the center-range method, proposed by [33], to fit a linear regression model to interval-valued symbolic interval data. This approach includes the information given by both the center and the range of an interval on a linear regression model in order to improve the models prediction performance. Based on experiments using synthetic and real symbolic data, [33] confirms the importance of considering the range information in models to predict symbolic interval data, where the center-range method exhibited significantly better performance than the approaches of [5] and [64].

Let us consider a set of $t = 1, \ldots, n$ examples that are described by $p + 1$ symbolic interval-valued variables $\mathbf{y}_t, \mathbf{x}_{1t}, \ldots, \mathbf{x}_{pt}$. Each fuzzy rule i, $i = 1, \ldots, c$ corresponds to a linear relationship. To facilitate notation, henceforth we omit the index i related to each cluster or fuzzy rule. Thus, the output of iFIS regarding each fuzzy rule, considering the center-range method, can be written as:

$$y_t^c = \beta_0^c + \beta_1^c x_{1t}^c + \cdots + \beta_p^c x_{pt}^c + \epsilon_t^c,$$

Please cite this article in press as: L. Maciel, R. Ballini, A fuzzy inference system modeling approach for interval-valued symbolic data forecasting, Knowledge-Based Systems (2018), https://doi.org/10.1016/j.knosys.2018.10.033.

$$y_t^r = \beta_0^r + \beta_1^r x_{1t}^r + \cdots + \beta_p^r x_{pt}^r + \epsilon_t^r, \tag{14}$$

where y^c and y^r are the center and range of $\mathbf{y}$, respectively, x_j^c and x_j^r are the center and range of $\mathbf{x}_i$, respectively, $j = 1, \ldots, p$, and ϵ^c and ϵ^r are the corresponding residuals for lower and upper interval bounds equations, respectively.

In this approach, the squared sum of deviations represents the sum of the center square error plus the sum of the range square error, considering independent vectors of parameters to predict the center and the range of the intervals [33]:

$$\begin{aligned} S &= \sum_{t=1}^{n} \left((\epsilon_t^c)^2 + (\epsilon_t^r)^2\right) \\ &= \sum_{t=1}^{n} (y_t^c - \beta_0^c - \beta_1^c x_{1t}^c - \cdots - \beta_p^c x_{pt}^c)^2 \\ &\quad + \sum_{t=1}^{n} (y_t^r - \beta_0^r - \beta_1^r x_{1t}^r - \cdots - \beta_p^r x_{pt}^r)^2, \end{aligned} \tag{15}$$

where $\{\beta_0^c, \beta_1^c, \ldots, \beta_p^c\}$ and $\{\beta_0^r, \beta_1^r, \ldots, \beta_p^r\}$ are the parameters associated with the affine center and range equations, respectively.

Differentiating (15) with respect to the parameters and setting the results equal to zero to obtain the normal equations, the least squares estimates of $\{\beta_0^c, \beta_1^c, \ldots, \beta_p^c\}$ and $\{\beta_0^r, \beta_1^r, \ldots, \beta_p^r\}$ which minimize (15) can be written in matrix notation as follows [33]:

$$\hat{\boldsymbol{\beta}} = [\hat{\beta}_0^c, \hat{\beta}_1^c, \ldots, \hat{\beta}_p^c, \hat{\beta}_0^r, \hat{\beta}_1^r, \ldots, \hat{\beta}_p^r]^T = (\mathbf{A})^{-1}\mathbf{b}, \tag{16}$$

where $\mathbf{A}$ is a $2(p+1) \times 2(p+1)$ matrix and $\mathbf{b}$ is a $2(p+1)$ vector denoted as:

$$\mathbf{A} = \begin{pmatrix} n & \sum_t x_{1t}^c & \cdots & \sum_t x_{pt}^c & 0 & \cdots & 0 \\ \sum_t x_{1t}^c & \sum_t (x_{1t}^c)^2 & \cdots & \sum_t x_{pt}^c x_{1t}^c & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_t x_{pt}^c & \sum_t x_{1t}^c x_{pt}^c & \cdots & \sum_t (x_{pt}^c)^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & n & \cdots & \sum_t x_{pt}^r \\ 0 & 0 & \cdots & 0 & \sum_t x_{1t}^r & \cdots & \sum_t x_{pt}^r x_{1t}^r \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \sum_t x_{1t}^r x_{pt}^r & \cdots & \sum_t (x_{pt}^r)^2 \end{pmatrix}, \tag{17}$$

$$\begin{aligned} \mathbf{b} = \Bigg[& \sum_t y_t^c, \sum_t y_t^c x_{1t}^c, \ldots, \sum_t y_t^c x_{pt}^c, \sum_t y_t^r, \\ & \sum_t y_t^r x_{1t}^r, \ldots, \sum_t y_t^r x_{pt}^r \Bigg]^T. \end{aligned} \tag{18}$$

The reconstruction of the interval bounds is based on the center and range estimates, i.e. the predicted $\hat{\mathbf{y}} = [y^L, y^U]^T$ obtained from the estimated values $\hat{y}^c$ and $\hat{y}^r$ as follows:

$$\hat{y}^L = \hat{y}^c - \hat{y}^r, \quad \text{and} \quad \hat{y}^U = \hat{y}^c + \hat{y}^r, \tag{19}$$

where $\hat{y}^c = (\check{\mathbf{x}}^c)^T \hat{\boldsymbol{\beta}}^c$, $\hat{y}^r = (\check{\mathbf{x}}^r)^T \hat{\boldsymbol{\beta}}^r$, $\check{\mathbf{x}}^c = [1, x_1^c, \ldots, x_p^c]^T$, $\check{\mathbf{x}}^r = [1, x_1^r, \ldots, x_p^r]^T$, $\hat{\boldsymbol{\beta}}^c = [\hat{\beta}_0^c, \hat{\beta}_1^c, \ldots, \hat{\beta}_p^c]^T$, and $\hat{\boldsymbol{\beta}}^r = [\hat{\beta}_0^r, \hat{\beta}_1^r, \ldots, \hat{\beta}_p^r]^T$.

The least squares estimates of consequent parameters defined in Eq. (16) are computed for each fuzzy rule in iFIS consequent identification. Therefore, $\hat{\boldsymbol{\beta}}_i$ are the estimate of consequent parameters of the ith fuzzy rule. The iFIS output is then computed as the weighted sum of the local models outputs by their normalized degrees of activation as in Eq. (7).

3. Computational experiments

This section presents the experiments carried out to evaluate the forecasting performance of the suggested iFIS modeling approach. Both synthetic and real financial interval-valued data will be used against traditional statistical time series benchmarks and with linear and nonlinear techniques designed for interval-valued data.

The ARIMA and VECM methods comprise the statistical time series models. The ARIMA is considered a splitting single-valued method, i.e. lower and upper bounds in the intervals are processed individually as independent univariate time series. On the other hand, VECM, the linear Holt's exponential smoothing (Holt$^{\text{I}}$) [51] and the nonlinear interval multilayer perceptron neural network (iMLP) [54] were chosen as the interval-valued methods. Notice that besides VECM as an interval-valued approach, the model does not take into account the interrelations between the lower and upper bounds as an interval structure as Holt$^{\text{I}}$, iMLP and iFIS do.

The selection of the best ARIMA and VECM models was accomplished through the minimization procedure of the Bayesian Information Criterion (BIC) [65]. Their parameters were estimated by maximum likelihood. Holt$^{\text{I}}$ smoothing parameter matrices were estimated by minimizing the interval sum of squared one-step-ahead forecast errors using the limited memory BFGS method for bound constrained optimization as in [51]. The BFGS quasi-Newton method and backpropagation procedure were applied to estimate the iMLP parameters. The number of neurons in the hidden layer was chosen by carrying out simulations to reach the best performance in terms of accuracy measure. This same methodology was used to select the control parameter of iFIS, i.e. the number of clusters or fuzzy rules c. All methods were implemented in MATLAB computing environment.

To access the forecasting performance of the models among the proposed iFISs and selected benchmark approaches, the root mean square error (RMSE) and the mean absolute percentage error (MAPE) accuracy metrics are considered. They are computed as follows:

$$\text{RMSE}^B = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\frac{y_t^B - \hat{y}_t^B}{y_t^B}\right)^2}, \tag{20}$$

$$\text{MAPE}^B = \frac{100}{n}\sum_{t=1}^{n}\frac{|y_t^B - \hat{y}_t^B|}{y_t^B}, \tag{21}$$

where $B = \{L, U\}$ represents the lower and upper interval bounds, $\mathbf{y}_t = [y_t^L, y_t^U]^T$ and $\hat{\mathbf{y}}_t = [\hat{y}_t^L, \hat{y}_t^U]^T$ are the actual and predicted intervals at t, respectively, n is the sample size, and RMSE^L (MAPE^L) and RMSE^U (MAPE^U) are the RMSE (MAPE) for the ITS lower and upper bounds, respectively.

Since the data has an interval structure, it implies that both characteristics (lower and upper bounds) that describe intervals have to be taken into consideration together, using for example dissimilarity measures based on interval distance between the observed interval and its forecast. To quantify the overall accuracy of the fitted and forecasted ITS, the interval average relative variance (ARV$^{\text{I}}$) and the mean distance error (MDE) are considered:

$$\text{ARV}^{\text{I}} = \frac{\sum_{t=1}^{n}\left(y_t^U - \hat{y}_t^U\right)^2 + \sum_{t=1}^{n}\left(y_t^L - \hat{y}_t^L\right)^2}{\sum_{t=1}^{n}\left(y_t^U - \bar{y}^U\right)^2 + \sum_{t=1}^{n}\left(y_t^L - \bar{y}^L\right)^2}, \tag{22}$$

$$\text{MDE} = \frac{1}{n}\sum_{t=1}^{n}\frac{\varpi(\mathbf{y}_t \cup \hat{\mathbf{y}}_t) - \varpi(\mathbf{y}_t \cap \hat{\mathbf{y}}_t)}{\varpi(\mathbf{y}_t \cup \hat{\mathbf{y}}_t)}, \tag{23}$$

where $\bar{\mathbf{y}} = [\bar{y}^L, \bar{y}^U]^T$ is the sample average interval, and $\varpi(\cdot)$ indicates the width of the interval.

Please cite this article in press as: L. Maciel, R. Ballini, A fuzzy inference system modeling approach for interval-valued symbolic data forecasting, Knowledge-Based Systems (2018), https://doi.org/10.1016/j.knosys.2018.10.033.

This paper also considers the computation of descriptive statistics for ITS such as coverage and efficiency rates, as suggested by [66]. The coverage rate is calculated as

$$R^C = \frac{1}{n}\sum_{t=1}^{n}\frac{\varpi(\mathbf{y}_t \cap \hat{\mathbf{y}}_t)}{\varpi(\mathbf{y}_t)}, \tag{24}$$

and the efficiency rate as

$$R^E = \frac{1}{n}\sum_{t=1}^{n}\frac{\varpi(\mathbf{y}_t \cap \hat{\mathbf{y}}_t)}{\varpi(\hat{\mathbf{y}}_t)}. \tag{25}$$

These rates provide additional information on what part of the observed ITS is covered by its forecasts (coverage) and what part of the forecast covers the observed ITS (efficiency). If the observed intervals are fully enclosed in the predicted intervals the coverage rate will be 100%; the efficiency, however, could be less than 100% and reveal that the forecasted ITS is wider than the actual ITS. Hence, these statistics must be considered together. The indication of a good forecast is observed when the coverage and efficiency rates are reasonably high and the difference between them is small [66]. Therefore, we take into account the potential trade-off between the two rates by calculating their average: $\bar{R} = (R^C + R^E)/2$.

Additionally, significant differences among the forecasting methods are evaluated through the analysis of variance (ANOVA) and Turkey's HSD tests [67] in terms of ARV^I considering a 5% level.

3.1. Synthetic interval-valued time series

To calculate the performance of the iFIS modeling approach in interval-valued time series forecasting, four synthetic time series with distinct dynamics were each simulated with 3000 observations, as in the methodology suggested by [68]. The simulation procedure is described as follows:

- First, a generating process with a known structure for the center interval time series is assumed x_t^c, $t = 1, \ldots, n$, obtained from the interval-valued time series $\mathbf{x}_t = [x_t^L, x_t^U]^T$;
- Second, the respective range interval series x_t^r, $t = 1, \ldots, n$, obtained from the interval-valued time series $\mathbf{x}_t = [x_t^L, x_t^U]^T$, are randomly generated from a uniform distribution in the interval $[a, b]$, i.e. $x_t^r \sim U[a, b]$;
- Finally, the interval time series $\mathbf{x}_t = [x_t^L, x_t^U]^T$ is constructed with the relationship related to x_t^c and x_t^r: $x_t^L = x_t^c - x_t^r$ and $x_t^U = x_t^c + x_t^r$;
- Each simulated ITS is divided into an in-sample set with two-thirds of the data (2000 observations) and an out-of-sample set with the remaining one-third of the data (1000 observations). Models training concerns the in-sample set whereas the forecasts are obtained and evaluated considering the out-of-sample data.

Table 1 shows the data-generating process as a combination of the center and range series of the interval-valued time series used to compare the forecasting performance of ARIMA, VECM, $Holt^I$, iMLP and iFIS models. The first two configurations, $Linear_1$ and $Linear_2$, are linear processes, which present a linear correlation between the future and past values of the center series plus a random shock $\epsilon_t \sim N(0, 1)$. Notice that configuration $Linear_1$ is a random walk process with drift, which implies that the history of the process has no relevance to its future dynamic. On the other hand, configurations $Chaotic_1$ and $Chaotic_2$ are nonlinear time series with complex, chaotic behavior. Fig. 1 illustrates each of the simulated interval-valued time series. The vertical line segments correspond to the actual interval-valued data, and the extremes correspond to the lower and upper interval bounds.

Table 1
Simulated interval-valued time series configurations.

Configuration	x^c process	x^r process
$Linear_1$	$x_{t+1}^c = 0.7 + x_t^c + \epsilon_t$	$x_{t+1}^r \sim U[5, 10]$
$Linear_2$	$x_{t+1}^c = 12 + 0.6x_t^c + \epsilon_t$	$x_{t+1}^r \sim U[2, 12]$
$Chaotic_1$	$x_{t+1}^c = 4x_t^c(1 - x_t^c)$	$x_{t+1}^r \sim U[2, 5]$
$Chaotic_2$	$x_{t+1}^c = \sin(12t) + \epsilon_t$	$x_{t+1}^r \sim U[0, 2]$

The results were obtained from Monte Carlo simulation experiments. For each data generating process configuration, 1000 ITS with 3000 observations were simulated.[3] Therefore, at the end of the experiments, the mean of the accuracy measures was calculated based on the 1000 Monte Carlo replications concerning the out-of-sample data. For $Linear_1$, $Linear_2$, $Chaotic_1$ and $Chaotic_2$ configurations, the iFIS best results were achieved based on a structure with 3, 2, 4 and 3 fuzzy rules. Regarding the iMLP, the method considered a network with 12 neurons in the hidden layer for the four synthetic interval-valued series.

Table 2 shows the forecasting performance of the examined methods in terms of accuracy for the different ITS configurations in the 1000 Monte Carlo replications. The best results are highlighted in bold. Lower values of RMSE, MAPE, ARV^I and MDE lead to better forecasts, whereas higher $\bar{R}$ rates are related to better forecasts. The suggested iFIS model achieved a better average performance in nearly all the situations considered, except for the ITS $Chaotic_2$ configuration, in which iMLP provided more accurate results.[4] For the linear interval-valued time series, $Linear_1$ and $Linear_2$, ARIMA, VECM and $Holt^I$ presented similar performances due to their linear structures. Concerning the chaotic interval-valued time series, $Chaotic_1$ and $Chaotic_2$ configurations, significant superior performance of the iFIS and iMLP methods is verified, revealing the advantage of nonlinear techniques for complex time series modeling. Note that, even for the series with a linear correlation structure, $Linear_1$ and $Linear_2$, the fuzzy model achieved better accuracy in the predictions than the ARIMA, VECM and $Holt^I$ models. In terms of the comparison between the methods, iMLP and iFIS provide similar and better results in all interval-valued time-series configurations (Table 2).

When considering the comparison between the three linear methodologies, the $Holt^I$ method only makes more accurate forecasts than ARIMA and VECM for the interval-valued time series with complex dynamics ($Chaotic_1$ and $Chaotic_2$), whereas in the linear configurations ($Linear_1$ and $Linear_2$,), ARIMA and VECM outperform the interval exponential smoothing approach (see Table 2). Further, gains in accuracy by using the nonlinear interval-valued models, iMLP and iFIS, over the linear alternative approaches are more evident when the ITS assume nonlinear chaotic behavior. The ARIMA, VECM and $Holt^I$ are inferior in that they are linear approaches. In addition, in the case of ARIMA and VECM, they ignore the possible mutual dependency between the lower and upper bounds of ITS, evidencing the advantage of interval-valued approaches.

Results were also compared in statistical terms considering the ARV^I values. First, an ANOVA test was performed to verify whether

[3] Experiments were conducted for different number of simulated observations. Models performances were not affected by the size of the data set, except when a small number is used (such as less than 50 simulations), where all methods reduce their accuracy for out-of-sample set as the data do not provide enough information for learning their parameters in the in-sample estimation.

[4] Besides the better accuracy of iMLP, one must note that the results from iFIS and iMLP are very similar. This scenario can be explained by the well-known no free-lunch theorem [69]: if an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems.

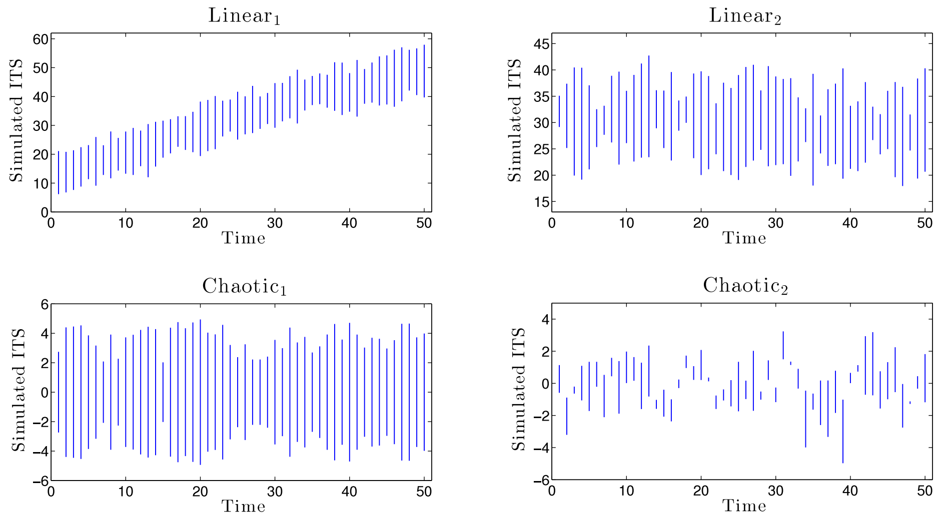

Fig. 1. Examples of simulated interval-valued time series with distinct dynamics.

Table 2
Models performance comparison in terms of accuracy measures for simulated interval-valued time series forecasting.

Method	Metrics						
	$RMSE^L$	$RMSE^U$	$MAPE^L$	$MAPE^U$	ARV^I	MDE	$\bar{R}$
Panel A: Linear$_1$ model configuration							
ARIMA	0.009	0.009	0.129	0.127	0.023	0.258	0.765
VECM	0.007	0.008	0.128	0.125	0.023	0.239	0.769
$Holt^I$	0.018	0.019	0.215	0.211	0.034	0.301	0.791
iMLP	**0.002**	0.002	0.127	0.120	0.022	0.240	0.870
iFIS	**0.002**	**0.001**	**0.118**	**0.115**	**0.018**	**0.237**	**0.875**
Panel B: Linear$_2$ model configuration							
ARIMA	0.171	0.132	14.092	10.282	1.870	0.415	0.762
VECM	0.170	0.130	13.972	9.871	1.852	0.403	0.775
$Holt^I$	0.230	0.198	15.726	12.362	2.019	0.498	0.760
iMLP	0.174	0.129	13.962	9.982	1.801	0.402	0.781
iFIS	**0.169**	**0.125**	**13.900**	**9.847**	**1.755**	**0.368**	**0.813**
Panel C: Chaotic$_1$ model configuration							
ARIMA	0.453	0.376	16.762	16.229	3.356	0.476	0.578
VECM	0.398	0.344	16.201	15.762	3.109	0.419	0.609
$Holt^I$	0.298	0.287	14.720	15.276	2.271	0.366	0.692
iMLP	**0.141**	**0.133**	**9.982**	**11.017**	**1.487**	**0.209**	**0.898**
iFIS	0.146	0.139	10.265	11.094	1.511	0.216	0.892
Panel D: Chaotic$_2$ model configuration							
ARIMA	0.871	0.795	13.722	9.646	4.775	0.747	0.518
VECM	0.789	0.784	13.456	9.342	4.602	0.711	0.546
$Holt^I$	0.653	0.591	11.928	8.911	3.927	0.771	0.672
iMLP	0.541	0.349	7.376	6.329	2.271	0.493	0.736
iFIS	**0.538**	**0.343**	**7.341**	**6.305**	**2.237**	**0.482**	**0.748**

Table 3
ANOVA test results for simulated interval-valued time series.

Configuration	ANOVA test	
	Statistics *F*	p-value
Linear$_1$	34.761	0.000*
Linear$_2$	41.299	0.000*
Chaotic$_1$	36.417	0.000*
Chaotic$_2$	38.530	0.000*

*Indicates significance at 0.05 level.

Table 4
Forecasting models ranking from Turkey's HSD test for simulated interval-valued time series.

Configuration	Rank of methods								
	1		2		3		4		5
Linear$_1$	iFIS	>	iMLP	>**	VECM	>	ARIMA	>**	$Holt^I$
Linear$_2$	iFIS	>	iMLP	>**	VECM	>	ARIMA	>**	$Holt^I$
Chaotic$_1$	iMLP	>	iFIS	>**	$Holt^I$	>**	VECM	>	ARIMA
Chaotic$_2$	iFIS	>	iMLP	>**	$Holt^I$	>*	VECM	>	ARIMA

*Indicates the mean difference between the two competing methods is significant at the 5% level.
**Indicates the mean difference between the two competing methods is significant at the 1% level.

a statistically significant difference exists among the competitive methods. The ANOVA results are presented in Table 3. For the four configurations, the statistics are significant at the 5% level, which indicates that there are significant differences among the examined alternatives. To further compare the significance difference between pairwise methods, the Tukey's HSD test was conducted. The results of the Tukey's HSD test are shown in Table 4 for each ITS configuration in a fashion that the methods are ranked from 1 (the best) to 5 (the worst).

According to the results in Table 4, the proposed iFIS and the iMLP methods perform statistically better than all of the other competitors in all cases, under a 95% confidence level . iMLP was ranked as the best approach for ITS Chaotic$_2$ configuration. However, in all cases the iFIS and iMLP showed equally accurate forecasts in statistical terms, i.e. the difference between the iMLP and the iFIS is not significant at the 0.05 level. Regarding the linear approaches, for Linear$_1$ and Linear$_2$ ITS, VECM and ARIMA outperform the $Holt^I$ method at 95% statistical significance. On the other hand, for Chaotic$_1$ and Chaotic$_2$ ITS configurations, $Holt^I$ achieved forecasts that were statistically superior to ARIMA and VECM. The possible reason for this is that in nonlinear chaotic dynamics, accounting for the interrelations between the ITS lower and upper bounds plays an important role.

3.2. Real financial interval-valued time series

The forecasting performance of the suggested iFIS method was also evaluated on real stock market ITS: S&P 500 and IBOVESPA indexes from the period of January 2004 to December 2015. These are the main indexes of the US and Brazilian stock exchanges. It can be noted that S&P 500 index is a popular choice among researchers to illustrate prediction methods in financial time series. Further, the consideration of these indexes enables us to evaluate the forecasting methods in the real stock market for developed (US) and emerging (Brazil) economies.

The intervals were constructed from a daily range of selected price indexes, i.e., the lowest and highest trading index values for the day were calculated to define the movement on the market for that day. Fig. 2 depicts the daily ITS of S&P 500 and IBOVESPA indexes from January 2, 2005 to April 28, 2005. Again, the first two-thirds of the observations consists of the in-sample set, while the remainder comprises the out-of-sample set.

For the real stock market ITS one-, five-, and ten-step-ahead forecasts were performed to assess the short-, medium-, and long-term forecasting ability of the proposed iFIS method and selected competitors, respectively. To implement multi-step-ahead forecasting (five- and ten-step-ahead), a commonly used iterated strategy is adopted in this study. Further, iMLP and iFIS model representations can be summarized as follows:

$$\hat{\mathbf{y}}_{t+h} \approx f\left(\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-l}\right), \tag{26}$$

where $f(\cdot)$ represents a nonlinear mapping, $\hat{\mathbf{y}}_{t+h} = [y^L_{t+h}, y^U_{t+h}]^T$ is the h-step-ahead forecasted value of the actual ITS $\mathbf{y}_{t+h} = [y^L_{t+h}, y^U_{t+h}]^T$, h is the forecasting horizon ($h = 1, 5$, and 10), and $(l+1)$ is the number of lagged values used as model inputs, chosen by simulations to reach the best performance in terms of accuracy measure.

For S&P 500 and IBOVESPA indexes, simulations indicate five and four ITS lagged values as inputs, respectively, for both the iMPL and the iFIS methods. Additionally, iFIS best results were achieved based on a structure with 3, and 5 fuzzy rules for S&P 500 and IBOVESPA indexes, respectively. The iMLP considered 15 neurons in the hidden layer for both real stock market ITS.

Tables 5 and 6 display the error measures for S&P 500 and IBOVESPA indexes ITS multi-step-ahead forecasting, respectively. The performances of the ARIMA, VECM, Holt^I, iMLP and iFIS models are evaluated by calculating the RMSE, MAPE, ARV^I, MDE and $\bar{R}$ metrics. For both ITS stock indexes and prediction horizons, the ARIMA and VECM models are beaten by the interval-valued approaches (Holt^I, iMLP and iFIS). The iFIS approach produced the best overall results. Overall, the rankings from best to worst are: iFIS, iMLP, Holt^I, VECM and ARIMA for both interval-valued S&P 500 and IBOVESPA (see Tables 5 and 6). The Holt^I method produced forecasts that were worse than the ones obtained by the iMLP and iFIS models. Comparing the results between the iMLP and iFIS methods, the former consistently achieved more accurate predictions.

When comparing the performance of each method across three prediction horizons (i.e., 1, 5, and 10 steps-ahead), the performances of all methods for both S&P 500 and IBOVESPA indexes ITS deteriorates with the increase in prediction horizon (Tables 5 and 6). It is worth noting that for medium- and long-term forecasts, the superiority of iFIS over iMLP is more evident, mainly for the interval-valued IBOVESPA time series where the errors found by the iFIS are significantly lower than those obtained by the iMLP.

Forecasts were further evaluated in statistical terms. Table 7 displays the ANOVA statistics with the corresponding p-values. At a 5% level, the results indicate significant differences among the ARIMA, VECM, Holt^I, iMLP models performance for all prediction horizons, i.e. 1, 5 and 10 steps ahead. Therefore, comparisons between pairwise methods are evaluated using the Tukey's HSD test, as shown in Table 8. Again, the forecasting approaches are ranked from 1 (the best) to 5 (the worst).

According to the results from Table 8, for both S&P 500 and IBOVESPA ITS, the proposed iFIS statistically outperforms all of the alternative approaches at a 95% confidence level for long-term forecasting, i.e. a ten-step-ahead prediction horizon. Concerning the interval-valued S&P 500 index, for $h = 1$ and $h = 5$ forecasting horizons, the iFIS and iMLP perform statistically better than the remaining methods (Table 8). Holt^I, VECM and ARIMA can be considered equally accurate for S&P 500 ITS forecasting for all predicting horizons considered in this study, except for long-term forecasts where VECM provides statistically superior forecasts to ARIMA. When considering the interval-valued IBOVESPA index, iFIS, iMLP and Holt^I outperform VECM and ARIMA methods at 95% statistical significance (Table 8). Finally, VECM and ARIMA provide equally accurate forecasts.

Summing up, from the experimental results provided in this paper, the main findings can be summarized as follows: (i) concerning the synthetic data sets, the interval-valued approaches, iMLP and iFIS, do improve forecasting accuracy compared to the alternative methods, more evidently when the data dynamics is complex, as the improvements are higher for the chaotic interval time series; (ii) concerning the financial interval indexes, results are similar in favor of iMLP and iFIS, and mostly important, when the forecasting horizon increases, the advantages of these interval-valued methodologies are more evident, in contrast with the lost of prediction accuracy of the other approaches for long-term horizons; (iii) generally for both data sets, it is worth noting that when data interval relationship is considered by the model, accuracy is

Table 5
Models performance comparison in terms of accuracy measures for S&P 500 interval-valued time series multi-step-ahead forecasting.

Method	Metrics						
	RMSE^L	RMSE^U	MAPE^L	MAPE^U	ARV^I	MDE	$\bar{R}$
Panel A: one-step-ahead prediction horizon							
ARIMA	0.273	0.287	1.672	1.284	0.863	0.711	0.501
VECM	0.198	0.201	1.459	1.192	0.817	0.699	0.546
Holt^I	0.120	0.124	1.283	1.019	0.792	0.659	0.602
iMLP	0.017	0.019	1.102	0.992	0.767	0.630	0.637
iFIS	**0.014**	**0.012**	**1.073**	**0.889**	**0.730**	**0.601**	**0.674**
Panel B: five-step-ahead prediction horizon							
ARIMA	0.311	0.302	1.876	1.487	1.277	0.788	0.489
VECM	0.210	0.245	1.811	1.404	1.245	0.773	0.512
Holt^I	0.134	0.153	1.562	1.354	1.109	0.710	0.573
iMLP	0.021	0.025	1.452	1.253	0.928	0.698	0.617
iFIS	**0.018**	**0.015**	**1.373**	**1.154**	**0.848**	**0.640**	**0.633**
Panel C: ten-step-ahead prediction horizon							
ARIMA	0.378	0.335	1.876	2.377	1.980	0.780	0.465
VECM	0.264	0.308	1.811	2.233	1.876	0.783	0.488
Holt^I	0.174	0.198	1.562	2.172	1.738	0.738	0.502
iMLP	0.031	0.036	1.452	1.982	1.562	0.701	0.544
iFIS	**0.027**	**0.024**	**1.121**	**1.861**	**1.471**	**0.698**	**0.568**

Table 6
Models performance comparison in terms of accuracy measures for IBOVESPA interval-valued time series multi-step-ahead forecasting.

Method	Metrics						
	RMSE^L	RMSE^U	MAPE^L	MAPE^U	ARV^I	MDE	$\bar{R}$
Panel A: one-step-ahead prediction horizon							
ARIMA	0.155	0.167	2.430	2.513	0.876	0.679	0.508
VECM	0.146	0.154	2.366	2.334	0.845	0.646	0.530
Holt^I	0.098	0.102	2.204	2.177	0.810	0.615	0.576
iMLP	0.029	0.034	1.985	1.911	0.759	0.593	0.587
iFIS	**0.025**	**0.024**	**1.941**	**1.869**	**0.753**	**0.589**	**0.604**
Panel B: five-step-ahead prediction horizon							
ARIMA	0.278	0.222	3.409	3.019	1.354	0.780	0.481
VECM	0.236	0.187	3.341	2.902	1.220	0.773	0.483
Holt^I	0.102	0.093	2.873	2.728	1.023	0.753	0.528
iMLP	0.056	0.053	2.663	2.563	0.981	0.710	0.534
iFIS	**0.044**	**0.042**	**2.445**	**2.346**	**0.883**	**0.682**	**0.573**
Panel C: ten-step-ahead prediction horizon							
ARIMA	0.346	0.315	3.994	3.872	2.388	0.793	0.457
VECM	0.309	0.295	3.827	3.653	2.293	0.782	0.476
Holt^I	0.127	0.128	3.674	3.546	1.873	0.756	0.498
iMLP	0.071	0.079	3.543	3.282	1.726	0.728	0.517
iFIS	**0.054**	**0.051**	**3.276**	**3.112**	**1.542**	**0.690**	**0.521**

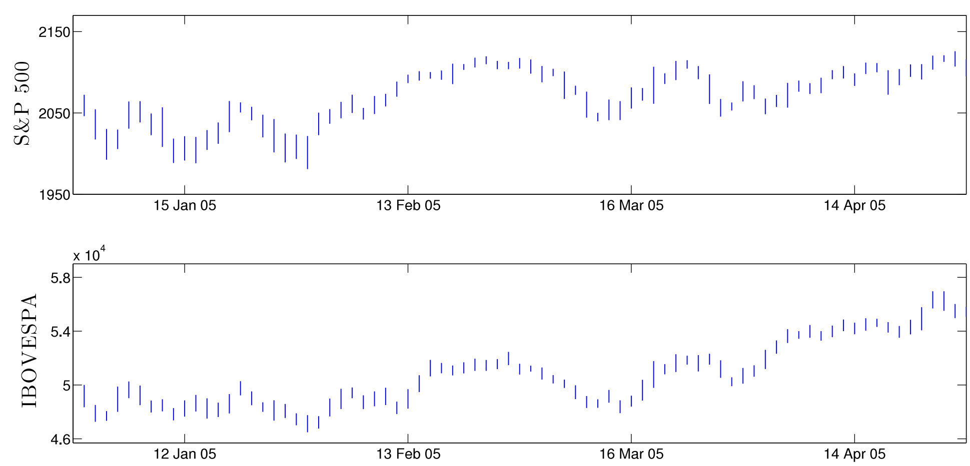

Fig. 2. Part of the interval-valued time series of S&P 500 and IBOVESPA stock indexes.

Table 7
ANOVA test results for real stock market interval-valued time series.

Prediction horizon	ANOVA test	
	Statistics F	p-value
Panel A: Interval-valued S&P 500 index		
$h = 1$	35.142	0.000*
$h = 5$	37.815	0.000*
$h = 10$	39.281	0.000*
Panel B: Interval-valued IBOVESPA index		
$h = 1$	28.331	0.000*
$h = 5$	34.009	0.000*
$h = 10$	33.662	0.000*

*Indicates significance at 0.05 level.

Table 8
Forecasting models ranking from Turkey's HSD test for real stock market interval-valued time series.

Prediction horizon	Rank of methods 1		2		3		4		5
Panel A: Interval-valued S&P 500 index									
$h = 1$	iFIS	>	iMLP	>**	Holt^I	>	VECM	>	ARIMA
$h = 5$	iFIS	>	iMLP	>**	Holt^I	>	VECM	>	ARIMA
$h = 10$	iFIS	>*	iMLP	>**	Holt^I	>	VECM	>**	ARIMA
Panel B: Interval-valued IBOVESPA index									
$h = 1$	iFIS	>	iMLP	>*	Holt^I	>**	VECM	>	ARIMA
$h = 5$	iFIS	>	iMLP	>	Holt^I	>**	VECM	>	ARIMA
$h = 10$	iFIS	>*	iMLP	>	Holt^I	>**	VECM	>	ARIMA

*Indicates the mean difference between the two competing methods is significant at the 5% level.
**Indicates the mean difference between the two competing methods is significant at the 1% level.

consistently improves as evidenced by the results from iMLP and iFIS; (iv) finally, the differences between iMLP and iFIS are more significant for the real stock market ITS, where iFIS provided lowest error values, suggesting the advantage of iFIS over iMLP due to its fuzzy nature in the modeling of uncertain data, as observed in financial prices.

4. Conclusion

The significant development of data collection technologies has contributed to the production of huge volumes of data. Therefore, approaches able to extract valuable information from large databases are demanding. When data are represented by symbolic interval-valued variables, it makes it possible to summarize the data and provide a way to account for the variability and/or uncertainty inherent to the data. In this domain, interval time series (ITS) forecasting plays an increasingly important role in areas such as financial markets, meteorology, and traffic flow management since it is seen as an additional tool for organizations and practitioners in policy and decision-making processes.

This paper presented a symbolic interval-valued fuzzy inference system (iFIS) for ITS forecasting. iFIS concerns a fuzzy rule-based modeling framework which provides a (non)linear framework to process interval-valued data. Additionally, the suggested approach is also able to account for imprecise data and vagueness due to its fuzzy nature. The iFIS antecedents identification uses a fuzzy c-means clustering algorithm for interval-valued data with adaptive distances, whereas parameters of the linear consequents are estimated with a center-range methodology to fit a linear regression model to symbolic interval data. The forecasting performance of the iFIS, in terms of accuracy measures and statistical tests, was compared to single- and interval-valued competitive methods through Monte Carlo experiments using both synthetic interval-valued time series with linear and chaotic dynamics, and real financial interval-valued time series.

The Monte Carlo simulations on both simulated ITS and real interval stock price time series indicate the high forecasting performance of the iFIS model, which is able to statistically outperform the alternative methods. iFIS provides a reduction of approximately 39% and 27%, on average, for the prediction errors concerning the synthetic and real-world data, respectively, when compared to the traditional forecasting methods. When the interval-valued time series exhibits chaotic behavior or multi-step-ahead prediction horizons are concerned, the superiority of the iFIS over the alternative forecasters is more significant. It demonstrated that the suggested iFIS methodology can be considered as a promising tool for interval time series forecasting. Future research includes the automatic selection of the number of clusters in iFIS antecedents identification, its extension for dealing with different symbolic data representations and also the development of an evolving framework in order to model data as streams. The evaluation of results on stock market data using economic criteria is also of great interest to financial managers and analysts.

Acknowledgments

The authors thank the Brazilian Ministry of Education (CAPES) and the São Paulo Research Foundation (FAPESP) for their support.

References

[1] L. Zhou, S. Pan, J. Wang, A.V. Vasilakos, Machine learning on big data: Opportunities and challenges, Neurocomputing 237 (2017) 350–361.
[2] P. Conde-Clemente, G. Trivino, J.M. Alonso, Generating automatic linguistic descriptions with big data, Inform. Sci. 380 (2017) 12–30.
[3] H. Wang, Z. Xu, W. Pedrycz, An overview on the roles of fuzzy set techniques in big data processing: Trends, challenges and opportunities, Knowl.-Based Syst. 118 (2017) 15–30.
[4] M. Chavent, H.-H. Bock, Clustering methods for symbolic objects, in: H.-H. Bock, E. Diday (Eds.), Analysis of Symbolic Data, Springer Berlin Heidelberg, Berlin, Heidelberg, 2000, pp. 294–341.
[5] L. Billard, E. Diday, Regression analysis for interval-valued data, in: H.A.L. Kiers, J.-P. Rasson, P.J.F. Groenen, M. Schader (Eds.), Data Analysis, Classification, and Related Methods, Springer Berlin Heidelberg, Berlin, Heidelberg, 2000, pp. 369–374.
[6] M.A.O. Domingues, R.M.C.R. Souza, F.J.A. Cysneiros, A robust method for linear regression of symbolic interval data, Pattern Recognit. Lett. 31 (2010) 1991–1996.
[7] L. Billard, E. Diday, From the statistics of data to the statistics of knowledge: Symbolic data analysis, J. Amer. Statist. Assoc. 98 (462) (2003) 470–487.
[8] E. Diday, Chapter 38 - categorization in symbolic data analysis, in: H. Cohen, C. Lefebvre (Eds.), Handbook of Categorization in Cognitive Science, Elsevier Science Ltd, Oxford, 2005, pp. 845–867.
[9] E. Diday, F. Esposito, An introduction to symbolic data analysis and the SODAS software, Intell. Data Anal. 7 (6) (2003) 583–601.
[10] M. Noirhomme-Fraiture, P. Brito, Far beyond the classical data models: Symbolic data analysis, Stat. Anal. Data Min. 4 (2011) 157–170.
[11] L. Lu, M. Medo, C.H. Yeung, Y.C. Zhang, Z.K. Zhang, T. Zhou, Recommender systems, Phys. Rep. 519 (2012) 1–49.
[12] J. Gao, T. Zhou, Big data reveal the status of economic development, J. Univ. Electron. Sci. Technol. China 45 (4) (2016) 625–633.
[13] S. Qin, F. Liu, J. Wang, Y. Song, Interval forecasts of a novelty hybrid model for wind speeds, Energy Rep. 1 (2015) 8–16.
[14] T. Xiong, C. Li, Y. Bao, Interval-valued time series forecasting using a novel hybrid Holt^I and MSVR model, Econ. Modell. 60 (2017) 11–23.
[15] W. Lin, G. González-Rivera, Interval-valued time series models: Estimation based on order statistics exploring the Agriculture Marketing Service data, Comput. Statist. Data Anal. 100 (2016) 694–711.
[16] Y.W. Cheung, An empirical model of dialy highs and lows, Int. J. Financ. Econ. 12 (1) (2007) 1–20.
[17] W. Xu, Symbolic Data Analysis: Interval-valued Data Regression (Ph.D. thesis), University of Georgia, 2010.
[18] X. Yang, J. Gao, J.H. Liu, T. Zhou, Height conditions salary expectations: Evidence from large-scale data in China, Physica A 501 (2018) 86–97.
[19] I. Couso, O. Strauss, H. Saulnier, Kendall's rank correlation on quantized data: An interval-valued approach, Fuzzy Sets and Systems 343 (2018) 50–64.
[20] Y. Sun, A. Han, Y. Hong, S. Wang, Threshold autoregressive models for interval-valued time series data, J. Econometrics (2018) http://dx.doi.org/10.1016/j.jeconom.2018.06.009.
[21] N.C. Lauro, F. Palumbo, Principal component analysis of interval data: A symbolic data analysis approach, Comput. Stat. 15 (1) (2007) 73–87.
[22] L. Billard, J. Le-Rademacher, Principal component analysis for interval data, WIREs Comput. Stat. 4 (6) (2012) 535–540.
[23] B. Liu, Y. Shen, W. Zhang, X. Chen, X. Wang, An interval-valued intuitionistic fuzzy principal component analysis model-based method for complex multi-attribute large-group decision-making, European J. Oper. Res. 245 (1) (2015) 209–225.
[24] M. Ichino, H. Yaguchi, E. Diday, A fuzzy symbolic pattern classifier, in: E. Diday, Y. Lechevallier, O. Opitz (Eds.), Ordinal and Symbolic Data Analysis, Springer Berlin Heidelberg, Berlin, Heidelberg, 1996, pp. 92–102.
[25] E. de A. Lima Neto, F.A.T. de Carvalho, C.P. Tenorio, Univariate and multivariate linear regression methods to predict interval-valued features, in: G.I. Webb, X. Yu (Eds.), AI 2004: Advances in Artificial Intelligence, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 526–537.
[26] L. Billard, Dependencies and Variation Components of Symbolic Interval-Valued Data, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 3–12, http://dx.doi.org/10.1007/978-3-540-73560-1_1.
[27] E.A. Lima Neto, G. Cordeiro, F.A.T. de Carvalho, Bivariate symbolic regression models for interval-valued variables, J. Stat. Comput. Simul. 81 (2011) 1727–1744.
[28] M. Iwasaki, H. Tsubaki, A bivariate generalized linear model with an application to meteorological data analysis, Stat. Methodol. 2 (2005) 175–190.
[29] E.A. Lima Neto, F.A.T. de Carvalho, Constrained linear regression models for symbolic interval-valued variables, Comput. Statist. Data Anal. 54 (2010) 333–347.
[30] S. Dias, P. Brito, Off the beaten track: A new linear model for interval data, European J. Oper. Res. 258 (2017) 1118–1130.
[31] C. Lim, Interval-valued data regression using nonparametric additive models, J. Korean Stat. Soc. 45 (2016) 358–370.
[32] A. Blanco-Fernández, A. Colubi, G. González-Rodríguez, Confidence sets in a linear regression model for interval data, J. Statist. Plann. Inference 142 (2012) 1320–1329.
[33] E.A. Lima Neto, F.A.T. de Carvalho, Center and range method for fitting a linear regression model to symbolic interval data, Comput. Statist. Data Anal. 52 (3) (2008) 1500–1515.
[34] F.A.T. de Carvalho, Fuzzy c-means clustering methods for symbolic interval data, Pattern Recognit. Lett. 28 (4) (2007) 423–437.
[35] D.S. Guru, B.B. Kiranagi, Multivalued type dissimilarity measure and concept of mutual dissimilarity value for clustering symbolic patterns, Pattern Recognit. 38 (2005) 151–256.
[36] K.C. Gowda, T.R. Ravi, Agglomerative clustering of symbolic objects using the concepts of both similarity and dissimilarity, Pattern Recognit. Lett. 16 (1995) 647–652.
[37] F.A.T. de Carvalho, R.M.C.R. Souza, M. Chavent, Y. Lechevallier, Adaptive Hausdorff distances and dynamic clustering of symbolic interval data, Pattern Recognit. Lett. 27 (3) (2006) 167–179.
[38] R.M.C.R. Souza, F.A.T. de Carvalho, Clustering of interval data based on city-block distances, Pattern Recognit. Lett. 25 (3) (2004) 353–365.
[39] P. D'Urso, R. Massari, L. De Giovanni, C. Cappelli, Exponential distance-based fuzzy clustering for interval-valued data, Fuzzy Optim. Decis. Mak. 16 (2017) 51–70.
[40] P. D'Urso, J.M. Leski, Fuzzy c-ordered medoids clustering for interval-valued data, Pattern Recognit. 58 (2016) 49–67.
[41] F.A.T. de Carvalho, C.P. Tenório, Fuzzy K-means clustering algorithms for interval-valued data based on adaptive quadratic distances, Fuzzy Sets and Systems 161 (23) (2010) 2978–2999.
[42] Z. Ji, Y. Xia, Q. Sun, G. Cao, Interval-valued possibilistic fuzzy C-means clustering algorithm, Fuzzy Sets and Systems 253 (2014) 138–156.
[43] K. Mali, S. Mitra, Symbolic classification, clustering and fuzzy radial basis function network, Fuzzy Sets and Systems 152 (2005) 553–564.
[44] P. D'Urso, L. De Giovanni, R. Massari, Self-organizing maps for imprecise data, Fuzzy Sets and Systems 237 (2014) 63–89.
[45] M. Yang, W. Hung, D. Chen, Self-organizing map for symbolic data, Fuzzy Sets and Systems 203 (2012) 49–73.
[46] L. Maciel, R. Ballini, F. Gomide, R.R. Yager, Participatory learning fuzzy clustering for interval-valued data, in: Proceedings of the 16th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2016), Eindhoven, The Netherlands, 2016, pp. 1–8.
[47] T. Xiong, Y. Bao, Z. Hu, R. Chiong, Forecasting interval time series using a fully complex-valued RBF neural network with DPSO and PSO algorithms, Inform. Sci. 305 (2015) 77–92.
[48] P.M.M. Rodrigues, N. Salish, Modeling and forecasting interval time series with threshold models, Adv. Data Anal. Classif. 9 (1) (2015) 41–57.
[49] J. Arroyo, A.M. San Roque, C. Maté, A. Sarabia, Exponential smoothing methods for interval time series, in: Proceedings of the 1st European Symposium on Time Series Prediction, 2007, pp. 231–240.
[50] N.M. Fiess, R. MacDonald, Towards the fundamentls of technical analysis: Analysing the information content of high, low and close prices, Econ. Modell. 19 (3) (2002) 353–374.
[51] A.L.S. Maia, F.A.T. de Carvalho, Holt's exponential smoothing and neural network models for forecasting interval-valued time series, Int. J. Forecast. 27 (3) (2011) 740–759.
[52] M. Caporin, A. Ranaldo, P.S. Magistris, On the predictability of stock prices: A case for high and low prices, J. Bank. Finance 37 (2013) 5132–5146.
[53] C. García-Ascanio, C. Maté, Electric power demand forecasting using interval time series: A comparison between VAR and iMLP, Energy Policy 38 (2010) 715–725.
[54] A.M. Roque, C. Maté, J. Arroyo, A. Sarabia, iMLP: Applying multi-layer perceptrons to interval-valued data, Neural Process. Lett. 25 (2) (2007) 157–169.
[55] J. Arroyo, R. Espínola, C. Maté, Different approaches to forecast interval time series: A comparison in finance, Comput. Econ. 27 (2) (2011) 169–191.
[56] F. Bacani, L.C. Barros, Application of prediction models using fuzzy sets: A bayesian inspired approach, Fuzzy Sets and Systems 319 (2017) 104–116.
[57] T. Korol, A fuzzy logic model for forecasting exchange rates, Knowl.-Based Syst. 67 (2014) 49–60.
[58] P. Singh, B. Borah, High-order fuzzy-neuro expert system for time series forecasting, Knowl.-Based Syst. 46 (2013) 12–21.
[59] F.M. Poulzols, A. Lendasse, A.B. Barros, Autoregressive time series prediction by means of fuzzy inference systems using nonparametric residual variance estimation, Fuzzy Sets and Systems 161 (4) (2010) 471–497.
[60] M.O. Efe, O. Kaynak, A novel optimization procedure for training of fuzzy inference systems by combining variable structure systems technique and Levenberg–Marquardt algorithm, Fuzzy Sets and Systems 122 (1) (2001) 153–165.
[61] T. Takagi, M. Sugeno, Fuzzy identification of systems and its applications to modeling and control, IEEE Trans. Syst. Man Cybern. SMC-15 (1) (1985) 116–132.
[62] R.E. Moore, R.B. Kearfott, M.J. Cloud, Introduction to Interval Analysis, SIAM Press, Philadelphia, 2009.

Please cite this article in press as: L. Maciel, R. Ballini, A fuzzy inference system modeling approach for interval-valued symbolic data forecasting, Knowledge-Based Systems (2018), https://doi.org/10.1016/j.knosys.2018.10.033.

[63] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithm, Plenum, New York, 1981.
[64] L. Billard, E. Diday, Symbolic regression analysis, in: Proceedings of the Eighenth Conference of the International Federation of Classification Societies (IFCS'02), Springer, Poland, 2002, pp. 281–288.
[65] G. Schwarz, Estimating the dimension of model, Ann. Statist. 6 (2) (1978) 461–464.
[66] P.M.M. Rodrigues, N. Salish, Modeling and forecasting interval time series with threshold models, Adv. Data Anal. Classif. 9 (1) (2015) 41–57.
[67] F. Ramsay, D. Schaefer, The Statistical Sleuth, Duxbury, Boston, Massachucets, 1996.
[68] A.L.S. Maia, F.A.T. de Carvalho, T.B. Ludermir, Forecasting models for interval-valued time series, Neurocomputing 71 (16–18) (2008) 3344–3352.
[69] D. Wolpert, W. Macready, No free lunch theorems for optimization, IEEE Trans. Evol. Comput. 1 (1) (1997) 67–82.

Please cite this article in press as: L. Maciel, R. Ballini, A fuzzy inference system modeling approach for interval-valued symbolic data forecasting, Knowledge-Based Systems (2018), https://doi.org/10.1016/j.knosys.2018.10.033.