Expert Systems with Applications 41 (2014) 3134–3142

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Determination of temporal information granules to improve forecasting in fuzzy time series

Lizhu Wang [a,b,*], Xiaodong Liu [a], Witold Pedrycz [c], Yongyun Shao [b]

[a] Research Center of Information and Control, Dalian University of Technology, Dalian 116024, China
[b] School of Mathematics and System Science, Shenyang Normal University, Shenyang 110034, China
[c] Department of Electrical and Computer Engineering, University of Alberta, Edmonton T6R 2V4, AB, Canada

ARTICLE INFO

Keywords:
Fuzzy time series
Gath–Geva (GG) clustering
Information granules
Enrollment
Segmentation

ABSTRACT

Partitioning the universe of discourse and determining intervals containing useful temporal information and coming with better interpretability are critical for forecasting in fuzzy time series. In the existing literature, researchers seldom consider the effect of time variable when they partition the universe of discourse. As a result, and there is a lack of interpretability of the resulting temporal intervals. In this paper, we take the temporal information into account to partition the universe of discourse into intervals with unequal length. As a result, the performance improves forecasting quality. First, time variable is involved in partitioning the universe through Gath–Geva clustering-based time series segmentation and obtain the prototypes of data, then determine suitable intervals according to the prototypes by means of information granules. An effective method of partitioning and determining intervals is proposed. We show that these intervals carry well-defined semantics. To verify the effectiveness of the approach, we apply the proposed method to forecast enrollment of students of Alabama University and the Taiwan Stock Exchange Capitalization Weighted Stock Index. The experimental results show that the partitioning with temporal information can greatly improve accuracy of forecasting. Furthermore, the proposed method is not sensitive to its parameters.

1. Introduction

For more than one decade, fuzzy time series has successfully been used to deal with various domain problems, such as stock index forecasting (Yu, 2005), hydrometeorology forecasting (Wang, Liu, & Yin, 2012), wastewater treatment (Wen & Lee, 1999), enrollment prediction (Song & Chissom, 1993a), temperature forecasting (Wang & Chen, 2009), etc. The concept of fuzzy time series was firstly introduced by Song and Chissom (1993b). Using the concept of fuzzy time series, they presented the time-invariant fuzzy time series model and the time-variant fuzzy time series model to forecast the enrollments of the University of Alabama. Following Song and Chissom, related studies mainly focus on improving the forecasting accuracy and reducing the computational complexity of the method.

The forecasting process in fuzzy time series consists of the following four steps: (1) partitioning the universe of discourse, (2) defining fuzzy sets and fuzzifying time series with the use of these fuzzy sets(fuzzification), (3) establishing fuzzy logical relationships from the fuzzy time series, and (4) forecasting and defuzzification of the output of fuzzy time series. In recent years, researchers have been realizing many studies to improve and explore all of these four steps. Concerning step (1), Huarng (2001a) observed that the length of intervals in the universe of discourse affects significantly forecasting results in fuzzy time series. In the sequel, they proposed the distribution-based length method and the average-based length method for handling the forecasting problems; Huarng (2006) suggested a different method which is called ratio-based lengths of intervals. Compared with the others of arbitrarily chosen length, Huarng's method has generated more accurate forecasts for enrollment, inventory demand and Taiwan stock price data. Li, Cheng, and Lin (2008) presented a model using fuzzy c-means (FCM) clustering to deal with interval partitioning, which takes the nature of data points into account and produces unequal-sized intervals. Wang and Chen (2009) proposed a method based on clustering techniques to predict the temperature and the Taiwan futures exchange. Kuo et al. (2009) presented a method to forecast the enrollment by involving particle swarm optimization. Yolcu, Egrioglu, Uslu, Basaran, and Aladag (2009), Egrioglu, Aladag, Yolcu, Uslu, and Basaran (2010) and Egrioglu, Aladag, Basaran, Yolcu, and

* Corresponding author. Address: Black building (Room A619), Research Center of Information and Control, Dalian University of Technology, No. 2, Linggong Road, Ganjingzi District, Dalian City, Liaoning Province 116024, China. Tel.: +86 02489795521.

E-mail address: wanglizhu80@126.com (L. Wang).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.10.046

Uslu (2011) proposed a new method based on the use of a single-variable constrained optimization to determine the length of interval. Chi, Fu, and Che (2010) suggested a K-means clustering technique for selecting the length of each interval. Zarandi, Molladavoudi, and Hemmati (2010) proposed Imperialist Competitive Algorithm to determine the length of interval. Bang and Lee (2011) presented a new clustering algorithm of which the structure hierarchically classifies non-linear data; There is not too much work dealing with step (2) besides the use of some fuzzy set theory. Step (3) is also deemed to one of the critical phases to influence forecasting result. Cheng, Chen, Teoh, and Chiang (2008) presented an adaptive expectation model for the Taiwan Stock Exchange Capitalization Weighted Stock Index(TAIEX) forecasting; Chen and Wang (2010) proposed a high-order fuzzy time series forecasting method using fuzzy-trend logical relationships; To reduce computational complexity, Chen (1996) presented an efficient forecasting procedure by grouping fuzzy logical relationships into rules and performing simplified arithmetic operations on these groups. Yu (2005), Cheng, Teoh, and Chen (2007), Teoh, Chen, and Cheng (2007), Lee, Efendi, and Ismail (2009) and Hung and Lin (2013) had also made their efforts to improve it. With regard to phase (4), most of the fuzzy time series models' are the same as that of Song and Chissom.

One of the evident limitations of these models is that they consider ad hoc approaches to process the original numeric data, and researchers seldom take the influence of time variable and the distribution of data itself into account when they partition the universe of discourse. Methods such as particle swarm optimization (Kuo et al., 2009), clustering techniques (Chen & Wang, 2010), support vector machines (Chen & Kao, 2013), entropy-based model (Cheng, Chang, & Yeh, 2006), and refined model (Yu, 2005), which utilize heuristics to segment intervals into subintervals in order to produce high forecasting accuracy, are not supported by underlining semantics.

In this paper, we propose a novel approach to determine an unequal length partitioning in consideration of temporal information(time variable) and distribution of data itself. The proposed method is based on GG clustering-based time series segmentation and the concept of information granules. The role of GG clustering-based time series segmentation and information granules is to determine temporal intervals of unequal length so that the model comes with increased accuracy and enhanced interpretability. The advantages of the proposed method can be summarized as follows:

- This approach becomes more comprehensive because of participation of time variable in partitioning the universe of discourse into intervals with unequal length.
- It has been observed that the forecasting accuracy for the two well-known data sets was significantly improved when the proposed method is employed.
- We determine intervals by GG clustering-based time series segmentation and information granules and these intervals carry well-defined semantics.

The proposed method has been experimentally tested on enrollment and the Taiwan Stock Exchange Capitalization Weighted Stock Index time series forecasting. The experimental results show that forecasting accuracy is evidently improved when comparing the proposed method with the equal length partitioning used in the previous studies.

The remaining content of this paper is organized as follows: In Section 2, we provide a brief review of fuzzy time series, GG clustering-based time series segmentation and information granules. In Section 3, we present the proposed method to partition the universe of discourse and determine unequal length intervals with temporal information. The performance of the proposed method on both enrollment and TAIEX time series forecasting are examined in Section 4. Conclusions are presented in Section 5.

2. Related works

In this section, some related background material including fuzzy time series, GG clustering-based time series segmentation and information granules is briefly reviewed.

2.1. Fuzzy time series

We start with a series of pertinent definitions.

Definition 2.1. Let $U = \{u_1, u_2, \ldots, u_n\}$ be the universe of discourse, a fuzzy set A of the universe of discourse U can be defined as follows:

$$A = \frac{f_A(u_1)}{u_1} + \frac{f_A(u_2)}{u_2} +, \cdots, + \frac{f_A(u_n)}{u_n} \tag{1}$$

where f_A is the membership function of the fuzzy set $A, f_A : U \to [0, 1]$ denotes the membership degree of u_i in the fuzzy set A, and $1 \leqslant i \leqslant n$.

Definition 2.2. Let $Y(t)(t = \ldots, 0, 1, 2, \ldots)$, which is a subset of R be the universe of discourse in which fuzzy sets $f_i(t)(i = 1, 2, \ldots)$ are defined. Let $F(t)$ be a collection of $f_i(t)(i = 1, 2, \ldots)$. Then, $F(t)$ is called a fuzzy time series on $Y(t)(t = \ldots, 0, 1, 2, \ldots)$.

Definition 2.3. Let $F(t-1) = A_i$ and $F(t) = A_j$. The relationship between two consecutive observations, $F(t-1)$ and $F(t)$, referred to as a fuzzy logical relationship, can be denoted by $A_i \to A_j$, where A_i is called the left-hand side and A_j the right-hand side of the fuzzy logical relationship.

Definition 2.4. Fuzzy logical relationships with the same fuzzy set located in the left-hand side of the relationships can be further grouped into a fuzzy logical relationship group (Huarng, 2001a). Suppose there are fuzzy logical relationships such that $A_i \to A_{j1},\ A_i \to A_{j2}, \ldots$, they can be grouped into a fuzzy logical relationship group $A_i \to A_{j1}, A_{j2}, \ldots$.

2.2. Gath and Geva clustering-based time series segmentation

The GG clustering algorithm was first proposed by Gath and Geva (1989). On the basis of GG clustering, Abonyi, Feil, Nemeth, and Arva (2005) developed an algorithm for dividing time series into fuzzy segments, which considers time series segmentation as GG clustering with time coordinate as an additional variable.

Suppose that a time series $\chi = \{\mathbf{x}_k | k = 1, \ldots, n\}$ is a finite set of n samples labeled by time-coordinate $\mathcal{T} = \{t_k | k = 1, \ldots, n\}$, the GG clustering-based time series segmentation algorithm forms a fuzzy clustering of the time series dataset by minimizing the following objective function with respect to membership grades $\mu_{i,k}$ and cluster prototypes $\boldsymbol{\eta}_i$, i.e.,

$$J_{GGTS} = \sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{i,k})^m \|\mathbf{z}_k - \boldsymbol{\eta}_i\|^2 = \sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{i,k})^m D^2(\mathbf{z}_k, \boldsymbol{\eta}_i), \tag{2}$$

where $\mathbf{z}_k = \left[t_k, \mathbf{x}_k^T\right]^T$ is the data point containing time coordinate, $m > 1$ is the weighting exponent, $c \geqslant 2$ is the segment(cluster) number, and $D^2(\mathbf{z}_k, \boldsymbol{\eta}_i)$ is the distance measurement between $\mathbf{z}_k$ and the prototype $\boldsymbol{\eta}_i$. The minimization of the objective function is realized under the following constraints: