ICIC Express Letters
Volume **10**, Number **12**, December **2016**

ICIC International ©2016 ISSN 1881-803X
pp. 2899–2904

PARTITIONING THE UNIVERSE OF DISCOURSE BASED ON AFFINITY PROPAGATION CLUSTERING TO IMPROVE FORECASTING IN FUZZY TIME SERIES

Yanli Shi[1] and Xianchang Wang[2,*]

[1]School of Science
Jilin Institute of Chemical Technology
No. 45, Chengde Street, Longtan District, Jilin 132022, P. R. China

[2]School of Sciences
Dalian Ocean University
No. 52, Heishijiao Street, Shahekou District, Dalian 116023, P. R. China
*Corresponding author: wxcixll@sohu.com

Received April 2016; accepted July 2016

Abstract. *Forecasting based on fuzzy time series is becoming more and more attractive. In fuzzy time series analysis, the intervals' length and the number of intervals are important issues. This paper proposes a new method for partitioning the universe of discourse to improve forecasting in fuzzy time series. The method can generate unequal length intervals for interval partitioning problem by affinity propagation clustering. Meanwhile, the intervals determined by typical cluster members make the fuzzy intervals more meaningful. For evaluating the performance of the proposed approach, an experimental study is reported using the well-known enrollment data of the University of Alabama. The comparison is carried out with regard to other fuzzy time series forecasting methods. The results have shown that the proposed method can improve the accuracy of fuzzy time series forecasting.*

Keywords: Fuzzy time series forecasting, Interval partitioning, Affinity propagation, Enrollment data

1. **Introduction.** Recently, the fuzzy time series forecasting models are becoming more and more popular, due to the fact that traditional time series analyses are not suitable for the linguistic data. Fuzzy time series approach is viewed as a linguistic model which is easily understood by human. Meanwhile, fuzzy time series forecasting models can forecast a linguistic value not a simple meaningless number. These characteristics make fuzzy time series model particularly suitable for forecasting problem. In last few decades, a number of excellent methods have been proposed for forecasting various practical problems, such as university enrollment forecasting [1], temperature forecasting [2], stock index forecasting [3], crop production forecasting [4], export forecasting [5], and rainfall forecasting [6].

The fuzzy time series forecasting models were first introduced by Song and Chissom [7, 8, 9]; following Song and Chissom, related works mainly focus on either improving the forecasting accuracy or reducing the computational complexity. Forecasting results can be affected by the length of intervals for partitioning the universe of discourse [10]. However, the length of intervals has been chosen arbitrarily in many fuzzy time series forecasting models [7, 9, 10, 11, 12]. To deal with this problem, Yolcu et al. [13] introduced a new method to partition the universe of discourse which is based on artificial bee colony (ABC) algorithm. Egrioglu et al. [14] proposed a new method through using a single variable constrained optimization to choose the length of interval.

Recently, clustering methods are employed for partitioning the universe of discourse. Cheng et al. partitioned the length of intervals by fuzzy c-means clustering method

[15, 16, 17]. Egrioglu et al. [18] proposed to use Gustafson-Kessel fuzzy clustering algorithm in the stage of fuzzification. However, the cutpoints of intervals obtained by these models are meaningless. Meanwhile, the number of fuzzy intervals has been also chosen arbitrarily in most existing literature, and the forecasting accuracy rates of the existing methods are not good enough. In this paper, an index of partitioning the universe of discourse is presented to choose the appropriate number of fuzzy intervals. And a new fuzzy time series approach is proposed by using affinity propagation (AP) clustering algorithm [19] in the stage of fuzzification to make the intervals more meaningful. The advantage of AP clustering is that its clustering results come up with typical cluster members (exemplars) not prototypes (e.g., fuzzy c-means clustering, and Gustafson-Kessel clustering), furthermore, AP clustering is a quite fast and effective clustering method. All these characteristics make AP clustering particularly suitable for partitioning the universe of discourse. For evaluating the performance of our approach, the proposed method is compared to previously proposed methods available in the literature by testing them on the well-known enrollment data for the University of Alabama.

The paper is organized as follows. Section 2 gives a brief review of fuzzy time series and AP clustering. Section 3 presents a new fuzzy time series forecasting method and an index for partitioning the universe of discourse. The performance of the proposed method on the well known enrollment data is examined in Section 4. Some conclusions are presented in Section 5.

2. **Fuzzy Time Series.**

The basic concepts of fuzzy time series are first presented by Song and Chissom [7, 8, 9]. The related definitions of fuzzy time series are given as follows.

Definition 2.1. *Let $U = \{u_1, u_2, \ldots, u_n\}$ be the universe of discourse, fuzzy sets $F_1, F_2, \ldots, F_c$ of U can be defined as follows: $F_i = \frac{f_{F_i}(u_1)}{u_1} + \frac{f_{F_i}(u_2)}{u_2} + \cdots + \frac{f_{F_i}(u_n)}{u_n}$ where f_{F_i} $(i = 1, 2, \ldots, c) : U \to [0, 1]$ is the membership function of the fuzzy set F_i, and c is the number of fuzzy sets one defined on U. $f_{F_i}(u_j)$ denotes the membership degree of u_t belonging to the fuzzy set F_i, and $1 \leq t \leq n$.*

Definition 2.2. *If $F(t)$ $(t = 0, 1, 2, \ldots)$ is a collection of F_i $(i = 1, 2, \ldots, c)$, then, $F(t)$ is called a fuzzy time series on $\{u_1, u_2, \ldots, u_n\}$.*

Definition 2.3. *Let $F(t-1) = F_i$ and $F(t) = F_j$, the first order fuzzy time series forecasting model can be defined as $F_i \to F_j$, where $F_i \to F_j$ is a fuzzy logical relationship, F_i is called the left-hand side and F_j is called the right-hand side of the fuzzy logical relationship.*

Huarng [10] grouped the fuzzy logical relationships by combining the same fuzzy set located in the left-hand side. Suppose there are three fuzzy logical relationships such that $F_i \to F_{j1}$, $F_i \to F_{j2}$, $F_i \to F_{j3}$, they can be grouped into a fuzzy logical relationship group $F_i \to F_{j1}, F_{j2}, F_{j3}$.

3. **The Proposed Fuzzy Time Series Forecasting Method.**

Given a time series $U = \{u_1, u_2, \ldots, u_n\}$, let $U_{\min} = \min\{x_i | x_i \in U\}$, $U_{\max} = \max\{x_i | x_i \in U\}$, $U = [U_{\min}, U_{\max}]$ be the universe of discourse. The number of fuzzy sets and the length of intervals have been chosen arbitrarily in the literature. For this reason, in this paper, an index of partitioning the universe of discourse is presented as follows.

3.1. **The proposed index of partitioning the universe of discourse.**

It is assumed that the universe is divided into c $(c \geq 2)$ unequal length subintervals. Calculating exemplar $e_1, e_2, \ldots, e_c$ and clusters $X_1, X_2, \ldots, X_c$ by AP clustering, where e_i is corresponding to X_i, e_i is an exemplar of the cluster X_i. Evaluate the clustering result usually by the following cost function.

$$J_c = \sum_{i=1}^{c} \sum_{x \in X_i} ||x - e_i||^2 \tag{1}$$

In general, as c increases the cost J_c will decrease, and J_2 is the biggest, so, J_c/J_2 is less than 1. To choose appropriate number of fuzzy sets c, penalty term of the number of clusters should be considered, and thus, a partition index is proposed as follows:

$$I_c = \frac{J_c}{J_2} + \frac{\sqrt{c}}{n} \tag{2}$$

The minimum of I_c means the most appropriate number of fuzzy sets should be chosen. The maximum of I_c is up to $1 + \sqrt{c}/n$.

3.2. **The proposed fuzzy time series forecasting method.**

Step 1: Determine the number of clusters c and partition the universe of discourse by exemplars.

According to Formula (2), the appropriate number of fuzzy sets c is chosen by the following formula:

$$c = \arg\min_{i=2,\ldots,n} \{I_i\} \tag{3}$$

The exemplars $e_1, e_2, \ldots, e_c$ are chosen by AP clustering and sorted ascending. The unequal length intervals can be calculated as: $u_1 = [U_{\min}, (e_1 + e_2)/2]$, …, $u_i = [(e_{i-1} + e_i)/2, (e_i + e_{i+1})/2]$, …, $u_c = [(e_{c-1} + e_c)/2, U_{\max}]$.

Step 2: Define fuzzy sets and obtain fuzzy time series.

First, fuzzy sets $F_1, \ldots, F_c$ defined on the universe of discourse U are chosen with different linguistic values such as "very small", "small", "medium small", "medium", "medium large", "large", and "very large". Each fuzzy set F_i is expressed in terms of the intervals $u_1, \ldots, u_c$ as follows: $F_1 = 1/u_1 + 0.5/u_2 + 0/u_3 + \cdots + 0/u_c$, …, $F_i = 0/u_1 + \cdots + 0.5/u_{i-1} + 1/u_i + 0.5/u_{i+1} + \cdots + 0/u_c$, …, $F_c = 0/u_1 + \cdots + 0/u_{c-2} + 0.5/u_{c-1} + 1/u_c$.

The way to fuzzify the historical data is to find which interval it belongs to according to its maximum degree of membership and associate with it the corresponding linguistic value by Definition 2.2.

Step 3: Establish fuzzy logical relationships groups.

Establish fuzzy logical relationships by Definition 2.3 and obtain the fuzzy logical relationships groups by the method proposed by [10].

Step 4: Forecast and defuzzify by exemplars.

Forecasting and defuzzification are following Lee et al.'s rules [20]. Suppose $F(t-1) = F_i$, the exemplars $e_1, e_2, \ldots, e_c$ are chosen by AP clustering:

Case 1: If $F_i \rightarrow F_j$, then $F(t) = F_j$, and a forecast value is equal to e_j;

Case 2: If $F_i \rightarrow F_{j1}, F_{j2}, \ldots, F_{jm}$, then $F(t) = F_{j1}, F_{j2}, \ldots, F_{jm}$, and a forecast value is equal to $(e_{j1} + e_{j2} + \cdots + e_{jm})/m$;

Case 3: if $F_i \rightarrow \emptyset$, then $F(t) = F_i$, and the forecast value is equal to e_i.

4. **Experimental Studies.**

To evaluate the model and analyze its merits, the method is tested on enrollment data for the University of Alabama, see Table 1; years and the corresponding enrollment numbers are listed in this table.

Step 1: Determine the number of clusters c and partition the universe of discourse by exemplars.

For $c = 2, 3, \ldots, n$, according to Formula (3) and AP culstering, given $c = 7$, and the obtained exemplars are $e_1 = 13563$, $e_2 = 14696$, $e_3 = 15433$, $e_4 = 15984$, $e_5 = 16807$, $e_6 = 18150$, $e_7 = 18970$. These exemplars are marked in Figure 1. The unequal length intervals

TABLE 1. The University of Alabama enrollment data

Year	Enrollment	Year	Enrollment
1971	13055	1982	15433
1972	13563	1983	15497
1973	13867	1984	15145
1974	14696	1985	15163
1975	15460	1986	15984
1976	15311	1987	16859
1977	15603	1988	18150
1978	15861	1989	18970
1979	16807	1990	19328
1980	16919	1991	19337
1981	16388	1992	18876

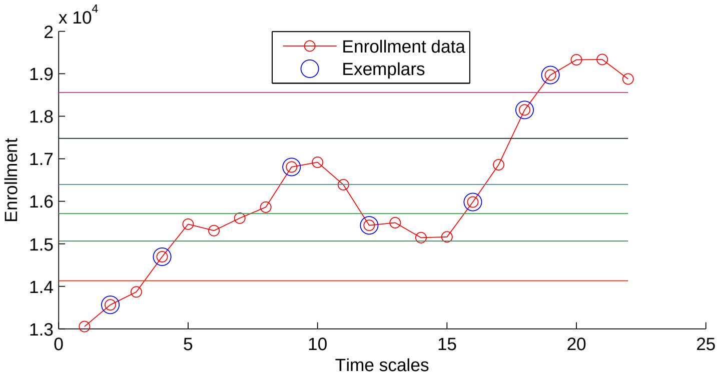

FIGURE 1. The obtained intervals

can be calculated as: $u_1 = [13055, 14129]$, $u_2 = [14129, 15065]$, $u_3 = [15065, 15709]$, $u_4 = [15709, 16396]$, $u_5 = [16396, 17478]$, $u_6 = [17478, 18560]$, $u_7 = [18560, 19337]$. Figure 1 shows these intervals; in Figure 1, these parallel lines partition Y axis into seven intervals, that is $u_1, u_2, \dots, u_7$.

Step 2: Define fuzzy sets and obtain fuzzy time series.

First, fuzzy sets $F_1, F_2, \dots, F_7$ defined on the universe of discourse U are chosen with different linguistic values "very small (very poor enrollment)" is denoted by F_1, "small (poor enrollment)" is denoted by F_2, "medium small (below average enrollment)" is denoted by F_3, "medium (average enrollment)" is denoted by F_4, "medium large (above average enrollment)" is denoted by F_5, "large (good enrollment)" is denoted by F_6, and "very large (excellent enrollment)" is denoted by F_7. Each fuzzy set F_i is expressed in terms of the intervals $u_1, u_2, \dots, u_7$ as follows:

$F_1 = 1/u_1 + 0.5/u_2 + 0/u_3 + 0/u_4 + 0/u_5 + 0/u_6 + 0/u_7$,
$F_2 = 0.5/u_1 + 1/u_2 + 0.5/u_3 + 0/u_4 + 0/u_5 + 0/u_6 + 0/u_7$,
$F_3 = 0/u_1 + 0.5/u_2 + 1/u_3 + 0.5/u_4 + 0/u_5 + 0/u_6 + 0/u_7$,
$F_4 = 0/u_1 + 0/u_2 + 0.5/u_3 + 1/u_4 + 0.5/u_5 + 0/u_6 + 0/u_7$,
$F_5 = 0/u_1 + 0/u_2 + 0/u_3 + 0.5/u_4 + 1/u_5 + 0.5/u_6 + 0/u_7$,
$F_6 = 0/u_1 + 0/u_2 + 0/u_3 + 0/u_4 + 0.5/u_5 + 1/u_6 + 0.5/u_7$,
$F_7 = 0/u_1 + 0/u_2 + 0/u_3 + 0/u_4 + 0/u_5 + 0.5/u_6 + 1/u_7$.

Figure 2 shows the obtained fuzzy time series.

Step 3: Establish fuzzy logical relationships groups.

Establish fuzzy logical relationships by Definition 2.3 and obtain the fuzzy logical relationships groups: $F_1 \to F_1, F_2$; $F_2 \to F_3$; $F_3 \to F_3, F_4$; $F_4 \to F_3, F_5$; $F_5 \to F_4, F_5, F_6$; $F_6 \to F_7$; $F_7 \to F_7$.

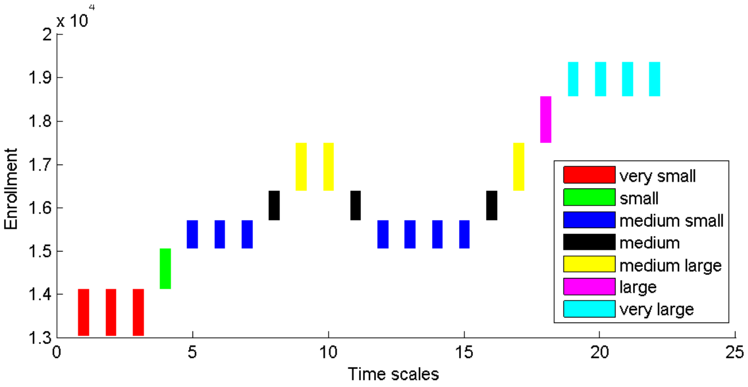

Figure 2. The obtained fuzzy time series

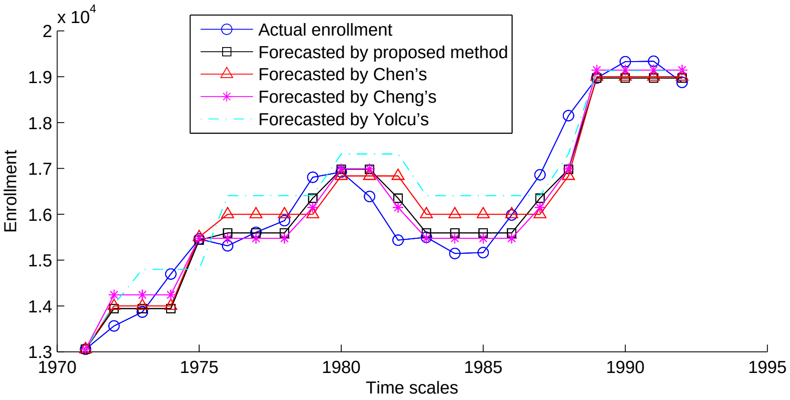

Figure 3. Results of comparative analysis

Step 4: Forecast and defuzzify by exemplars.

After defuzzification, the obtained crisp forecasts are shown in Figure 3. Figure 3 also depicts the forecasting results of Chen's [12], Cheng et al.'s [15] and Yolcu et al.'s [21]. Concerning mean square error (MSE) [21], the enrollment forecasting results of the proposed $MSE = 224690$ greatly outperformed Chen's [12] ($MSE = 407507$), Cheng et al.'s [15] ($MSE = 228918$) and Yolcu et al.'s [21] ($MSE = 648302$). The method can generate unequal length intervals for interval partitioning problem by affinity propagation clustering. Meanwhile, the intervals determined by typical cluster members make the fuzzy intervals more meaningful.

5. **Conclusions.** Recently, the fuzzy time series forecasting models are becoming quite popular. Traditional time series analyses cannot be applied to the linguistic data. Fuzzy time series approaches can be viewed as a linguistic model that makes these kinds of models can be easily understood by human. Meanwhile, fuzzy time series forecasting models can forecast a linguistic value not a number. Although this makes fuzzy approaches very attractive, there are still problems that are needed to be solved. One of these problems is to determine the lengths of intervals. The decision on what the lengths will be is very important for forecasting accuracy. Another problem is how to choose the appropriate number of intervals.

In order to solve this problem, in the proposed method, the AP clustering algorithm is employed to find the cutpoints of intervals by the obtained exemplars, and the forecasting result is computed based on these exemplars. Also, the number of the intervals can be identified by our proposed index of partitioning the universe of discourse to choose the appropriate number of fuzzy set. In order to show the efficiency of the proposed method, the well-known data set, which is the enrollment data at the University of Alabama is examined. As a result, it is obviously observed that the proposed method produces lower MSE values.

There are two main parameters of affinity propagation: preference and damping factor. The following research focuses on the relations between the accuracy of fuzzy time series forecasting models and two main parameters of affinity propagation.

REFERENCES

[1] S. M. Chen and N. Y. Wang, Fuzzy forecasting based on fuzzy-trend logical relationship groups, *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol.40, pp.1343-1358, 2010.

[2] S. M. Chen and J. R. Hwang, Temperature prediction using fuzzy time series, *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol.30, pp.263-275, 2000.

[3] S. M. Chen, G. M. T. Manalu, J. S. Pan and H. C. Liu, Fuzzy forecasting based on two-factors second-order fuzzy-trend logical relationship groups and particle swarm optimization techniques, *IEEE Trans. Cybernetics*, vol.30, pp.1102-1117, 2013.

[4] S. R. Singh, A robust method of forecasting based on fuzzy time series, *Applied Mathematics and Computation*, vol.188, pp.472-484, 2007.

[5] H. L. Wong, Y. H. Tu and C. C. Wang, Application of fuzzy time series models for forecasting the amount of Taiwan export, *Expert Systems with Applications*, vol.37, pp.1465-1470, 2010.

[6] S. T. Li and Y. C. Cheng, Deterministic fuzzy time series model for forecasting enrollments, *Computers and Mathematics with Applications*, vol.53, pp.1904-1920, 2007.

[7] Q. Song and B. S. Chissom, Forecasting enrollments with fuzzy time series – Part I, *Fuzzy Sets and Systems*, vol.54, pp.1-9, 1993.

[8] Q. Song and B. S. Chissom, Fuzzy time series and its models, *Fuzzy Sets and Systems*, vol.54, pp.269-277, 1993.

[9] Q. Song and B. S. Chissom, Forecasting enrollments with fuzzy time series – Part II, *Fuzzy Sets and Systems*, vol.62, pp.1-8, 1994.

[10] K. Huarng, Effective lengths of intervals to improve forecasting in fuzzy time series, *Fuzzy Sets and Systems*, vol.123, pp.387-394, 2011.

[11] J. R. Hwang, S. M. Chen and C. H. Lee, Handling forecasting problems using fuzzy time series, *Fuzzy Sets and Systems*, vol.100, pp.217-228, 1998.

[12] S. M. Chen, Forecasting enrollments based on fuzzy time series, *Fuzzy Sets and Systems*, vol.81, pp.311-319, 1996.

[13] U. Yolcu, O. Cagcag, C. H. Aladag and E. Egrioglu, An enhanced fuzzy time series forecasting method based on artificial bee colony, *Journal of Intelligent and Fuzzy Systems*, pp.1-9, 2013.

[14] E. Egrioglu, C. H. Aladag, M. A. Basaran, U. Yolcu and V. R. Uslu, A new approach based on the optimization of the length of intervals in fuzzy time series, *Journal of Intelligent and Fuzzy Systems*, vol.22, pp.15-19, 2011.

[15] C. H. Cheng, G. W. Cheng and J. W. Wang, Multi-attribute fuzzy time series method based on fuzzy clustering, *Expert Systems with Applications*, vol.34, pp.1235-1242, 2008.

[16] S. T. Li, Y. C. Cheng and S. Y. Lin, A FCM-based deterministic forecasting model for fuzzy time series, *Computers and Mathematics with Applications*, vol.56, pp.3052-3063, 2008.

[17] E. Egrioglu, C. H. Aladag and U. Yolcu, Fuzzy time series forecasting with a novel hybrid approach combining fuzzy c-means and neural networks, *Expert Systems with Applications*, vol.40, pp.854-857, 2013.

[18] E. Egrioglu, C. H. Aladag, U. Yolcu, V. R. Uslu and N. A. Erilli, Fuzzy time series forecasting method based on Gustafson-Kessel fuzzy clustering, *Expert Systems with Applications*, vol.38, pp.10355-10357, 2011.

[19] B. J. Frey and D. Dueck, Clustering by passing messages between data points, *Science*, vol.315, pp.972-976, 2007.

[20] M. H. Lee, Z. Ismail and R. Efendi, Modified weighted for enrollment forecasting based on fuzzy time series, *Matematika*, vol.25, pp.67-78, 2009.

[21] U. Yolcu, E. Egrioglu, V. R. Uslu, M. A. Basaran and C. H. Aladag, A new approach for determining the length of intervals for fuzzy time series, *Applied Soft Computing*, vol.9, pp.647-651, 2009.