Increasing number of clusters on fuzzy time series (fts) forecasting method

R W Ningrum1‡, B Surarso2 and Farikhin2

1Master Program of Mathematics, Department of Mathematics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia
2Department of Mathematics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia

‡Corresponding author: rianiwidyaningrum@student.undip.ac.id

Abstract. Determining the effective number of clusters can influence forecasting results in FTS method’s applications. Unfortunately, the issue of how many clusters should be used to improve forecasting results has not been touched in previous researches. We observe for some different number of clusters and compare its Root Mean Squared Errors (RMSEs) results. The numerical simulation using Jakarta Composite Index (IHSG) data and the forecasting results show that RMSEs value decrease when the number of clusters is increased. The RMSEs value drops significantly when \( n \approx 50 \) clusters are used.

1. Introduction

There are many research on prediction methods based on time series data. In a conventional time series method, a special dynamic process are represented by crisp value, but in FTS (FTS) method, the historical data is represented as Linguistic Variable (LV) [1]. There are four major issues of FTS forecasting, clustering, building Fuzzy Logical Relationship (FLR) or fuzzification process, and defuzzification process [2-4].

One of the important thing at forecasting method based on FTS is clustering process. Many recent researches develop automatic clustering, such as Fuzzy C-means clustering algorithm [5], Robust clustering [6], and Firefly clustering algorithm [7]. Firefly Algorithm (FA) is often used in optimization, and very useful for partitioning. Ningrum et al. [7] used firefly algorithm at clustering stage to forecast Jakarta Composite Index (IHSG) based on Chen method. Chen method developed by Shyi-Ming Chen is always used 7 intervals at clustering steps. Classical Chen methods never explained why the universe of discourse always divided into 7 clusters with the same or static length of intervals [8,9]. As the effect of additional FA at Chen method, the clusters don’t have the same width anymore. This automatic clustering algorithm (FA) counts the best distance of the upper and lower limits of each intervals. The numerical simulation used 10 cluster at both method (Chen method and Firefly-chen method). The result shows that combined Firefly-chen method better than Chen method, it confirmed by a smaller error value.

Then, in this our recent research we study how the increasing number of clusters influence the error value of forecasting IHSG with FA at clustering process. FA can generates the dynamic length of intervals. This algorithm is able to find the best fitness value (i.e., the effective interval width) so that it produces the best forecasting results. The IHSG data is used for demonstration. The rest of the article is arranged as follows, Section 1 Introduction, Section 2 discussing basic concept of FTS and...
Firefly clustering algorithm, Section 3 Simulation, Section 4 Results and Discussion, Section 5 Conclusion.

2. Methods

2.1. Preliminaries

In statistics and signal processing, time series is a series of data in the form of observation values measured over a certain period of time, based on the same uniform interval. Here is some basic concept of FTS.

2.1.1. Fuzzy Time Series (FTS)

Definition 2.1 [10]

A fuzzy set $A$ of the universe of discourse $U, U = \{u_1, u_2, ..., u_n\}$, is defined as the following formula [4]:

$$A = \{f_A(u_i), \quad i = 1, 2, ..., n\}$$

(1)

where $f_A$ is the member’s func of the fuzzy set $A$, $f_A: U \rightarrow [0,1]$ $f_A(u_i)$ denotes the degree of membership of $u_i$ in the fuzzy set $A$, and $1 \leq i \leq n$.

Definition 2.2 [11]

Let $D(t)$, a subset of $\mathbb{R}$, be universe of discourse on fuzzy sets $f_i(t)(i = 1, 2, ...)$; $F(t)$ is the collection of $f_i(t)$, then $F(t)$ is called a FTS on $D(t)$.

From Definition 2.2, we can see that $F(t)$ can be regarded as a linguistic variable and $f_i(t)(i = 1, 2, ...)$ can be viewed as possible linguistic values of $F(t)$, where $f_i(t)(i = 1, 2, ...)$ are represented by fuzzy sets. We can also see that $F(t)$ is a function of time $t$.

2.1.2. Firefly Algorithm (FA)

Here are the steps in the FA:

Step 1 : Determine random coefficient $\alpha$ attractiveness value of firefly $\beta$, light absorption level around environment $\gamma$, the number of firefly, and the maximum iteration limit.

Step 2 : Spread firefly randomly

Step 3 : Calculate fitness value

Chen method is used to get the error value for each forecasting point.

Step 4 : Determine the brightness firefly i.e., firefly displacement distance which resulting a forecast value with the smallest errors.

Step 5 : The dim firefly move to brightness firefly using firefly position displacement formula:

$$r_{ij} = \left\lVert x_i - x_j \right\rVert = \sqrt{\sum_{k=1}^{d}(x_{ik} - x_{jk})^2}$$

(2)

In the case 2-D, we obtained the following equation [5]:

$$r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

(3)

where,

$r_{ij}$ : difference distance between initial firefly $i$ and $j$

$x_i$ : firefly $i$ position at location $x$

$x_j$ : firefly $j$ position at location $x$

$x_{ik}$ : $k$-th component from spatial coordinate $x_i$ of $i$-th firefly

$x_{jk}$ : $k$-th component from spatial coordinate $x_j$ of $j$-th firefly

Random coefficient $\alpha$ ranged between $[0,1]$, while $\alpha$ value increase then firefly randomization get bigger. Attractiveness value of firefly $\beta$ also ranged between $[0,1]$. If the value of $\beta$ increased, it will effect on the speed of firefly movement. The coefficient value $\gamma$ states the level of light absorption in the environment around firefly ranges from 1 to 10.
In step 3-4 used Chen method stage 3-6 [6] to get the best fitness. The fitness referred in firefly algorithm is the distance displacement of dim firefly to the brightest firefly.

2.2. Simulation

In this paper, fuzzy member’s functions are limited in symmetric triangular member’s functions and as the effect of firefly algorithm addition at the clustering process, all of the triangle may have different shapes at same intervals. Geometrically a triangular shape with different slopes is illustrated with the following Figure 1 and Figure 2.

![Figure 1](image1.png)  ![Figure 2](image2.png)

Figure 1. Representation of triangular member’s function with \( n = 2 \) clusters.

Figure 2. Representation of triangular member’s function with \( n = 4 \) clusters.

Triangles with small interval as figure 2 have a sharp slope, whereas if the width of the interval is wider, the slope of the triangle will be more sloping. The addition of the number of clusters causes the width of the interval to be smaller, so the triangle has a sharper slope. We simulate the running times number of clusters begin 10 clusters then up to a multiple of 10 until 100 clusters by using Matlab program.

3. Results and Discussion

The calculation process use Firefly-Chen method. The first numerical simulation use \( n = 10 \) cluster and we try to increase the number of clusters. As explained in the preceding sections, firefly clustering algorithm caused the cluster have non-static length of intervals [6].

Simulation is also done by increasing the number of clusters. The increased number of clusters caused the length of each cluster interval to be smaller, so the forecasting results closer to the actual value. The forecasting result’s graphic with \( n = 10 \) cluster have already explained Ningrum [4]. Here is the forecasting result’s graphic with \( n = 50 \) clusters and \( n = 100 \) clusters.

Figure 3 and 4 show the forecasting results of both methods overall is smooth approaching the actual data value. The forecasting result of Figure 3 and Figure 4 is almost has no significant difference. The increasing number of clusters causes the length of each cluster interval to be smaller, so the forecasting results closer to the actual value. In this paper, the RMSEs value are used to compare the forecasting accuracy rates. The simulation used the running times of multiple 10 number of clusters begin with 10 until 100 clusters.
Figure 3. Graphic of forecasting results with \( n = 50 \) cluster

Figure 4. Graphic of forecasting results with \( n = 100 \) cluster
Table 1 shows a comparison of the RMSEs value for each method. From the RMSEs value we can see that the Firefly-Chen method has smaller RMSEs than Chen Method and the modification (increased number of clusters) effectively reduce RMSEs value. In the other words, the increased number of cluster produce better forecasting results. RMSEs value drops sharply at point n=50 clusters and then the RMSEs value decreases not significantly until n = 100 clusters.

### 4. Conclusion

In this paper, we have presented the simulation of the increased number of clusters. Based on firefly clustering algorithm additional process, the triangle of each interval may have different shape depend on differentiation of hypotenuse. The increased number of clusters causes the intervals have smaller width, so the peak value of the triangular function closer to the actual value.

From Table 1, the increased number of clusters can effectively reduce RMSEs value. The numerical simulation using IHSG data and the forecasting results show that RMSEs value decrease when the number of clusters is increased. The RMSEs value drops significantly when \( n = 50 \) clusters are used. In the other words, we conclude that the effective number of cluster should used to forecast IHSG at this case is 50 clusters. It means that if the length of interval get smaller or the more tapered the triangular peaks are able to provide better forecasting results.

### References

[1] Singh P and Borah B 2013 *Eng. Appl. Artif. Intell* 26 2443
[2] Suryono S, Surarso B, Saputra R and Bardadi A 2017 *Adv. Sci. Lett* 23 6638
[3] Dincer N G and Akkus O 2018 *Ecol. Inform* 43 157
[4] Stenafakos C 2016 *Ocean Eng* 121 1
[5] Rubio A, Bermudez J D and Vercher E 2016 *Int. J. Approx. Reason* 75 1 – 12
[6] Aripin A, Suryono S and Surarso B 2016 *AIP Conf. Proc* 1746 020046
[7] Ningrum R W, Surarso B, Farikhin and Safarudin Y M 2018 *IOP Conf.Series : J. Phys. Conf. Ser* 983 0120055
[8] Chen S M and Phuong B D H 2017 *Knowl.-Based Syst* 118 204
[9] Chen S M and Tanuwijaya K 2011 *Experts Syst. Appl* 38 10594
[10] Ding Y and Fu X 2016 *Neurocomputing* 188 233
[11] Singh P and Dhiman G 2018 *J. Comput. Sci* 27 370
[12] Bisht S and Kumar S 2016 *Expert Syst. Appl* 64 557

| Number of clusters | Chen Method | Firefly-Chen Method |
|--------------------|-------------|---------------------|
| 10                 | 228.0103    | 184.7329            |
| 20                 | 205.5390    | 167.1405            |
| 30                 | 177.0839    | 151.6063            |
| 40                 | 168.6163    | 135.4238            |
| 50                 | 144.8957    | 124.0389            |
| 60                 | 142.9726    | 127.4675            |
| 70                 | 141.7997    | 119.3710            |
| 80                 | 140.2965    | 114.0155            |
| 90                 | 116.5619    | 104.5882            |
| 100                | 113.5088    | 101.4061            |