INTER-QUARTILE RANGE APPROACH TO LENGTH–INTERVAL ADJUSTMENT OF ENROLLMENT DATA IN FUZZY TIME SERIES FORECASTING

ZUHAIMY ISMAIL,*‡, RISWAN EFENDI,*§ and MUSTAFA MAT DERIS†¶

*Mathematics Department
Universiti Teknologi Malaysia
Skudai Johor Bahru 81310, Malaysia

‡Faculty of Computer Science
Universiti Tun Hussein Onn Malaysia
Batu Pahat Johor 86400, Malaysia

§zuhaimy@utm.my

¶wanchaniago@gmail.com

†mustafa@uthm.edu.my

Received 7 January 2013
Revised 18 April 2013
Published 22 August 2013

Various methods have been presented to investigate the length of data interval and partition number of universe of discourse in fuzzy time series to achieve high level forecasting accuracy. However, the interval length is still an issue and thus, influencing the forecasting accuracy. This paper proposes a new approach using the average inter-quartile range to improve the interval length and subsequently the partition number of universe of discourse. Moreover, in forecasting method, the first-differencing data is also considered to obtain better forecast. The enrollment data of Alabama University is used as an example and the efficiency of the proposed method is compared with the previous methods. The result shows that the proposed method improves the accuracy and efficiency of the yearly enrollment forecasting opportunities.

Keywords: Fuzzy time series; enrollment; interval length; inter-quartile range; first-differencing.

1. Introduction

In the present prediction system, many soft computing methods are proposed for forecasting and predicting of nonlinear time series data by researchers. For example, genetics algorithm, neural network, fuzzy logic, fuzzy neural network, evolutionary algorithm, etc. These methods aimed to cover the limitations of classical time series methods especially the linearity assumptions in Ref. 1. This is also one of the reasons why the fuzzy time series (FTS) concept was introduced in Ref. 2. In this concept, the requirement of the statistical assumptions is not needed and the number of

§Corresponding author.
observations is also not limited in the forecasting. Thus, FTS approaches have been widely used in real life such as enrollment, stocks index prices, temperature and financial prediction.

In FTS forecasting, there are three determinative factors that have been considered to achieve the high level forecasting accuracy. First, the lengths of intervals and partition number of universe of discourse. Second, the weight of fuzzy logical relationships (FLRs). Third, the effective steps and the proper rules to obtain the forecast values. These factors are still investigated until present by researchers. Through this paper, the first and third factors are explored in detail.

Many contributions have been done by researchers in determining of interval length and partition number. For example, the universe of discourse is divided into seven equal length intervals in Ref. 5. Furthermore, in Ref. 6, the development of Chen’s work is discussed by re-dividing intervals based on the highest frequency into sub-intervals. Meanwhile, in Ref. 14, the highlighting of the interval length influences the forecasting performance and proposed two different methods are presented by using the average and the distribution, for defining the length. Additionally, the ratio-based lengths is explored in Ref. 23, where this way is more appropriate than the equal lengths of intervals. The effective way is introduced using average-based lengths of interval in Ref. 24 and applied to Chen’s model (Ref. 5). An optimization of interval length is also proposed in Ref. 25 by using the polynomial interpolation. On the other hand, a new bandwidth interval is introduced using forecasting in Ref. 26. The latest work in 2012, in Ref. 1, the intervals are re-divided into sub-intervals using highest frequency as mentioned in Chen’s and Hsu’s works (Ref. 6), but the total number of intervals should not exceed integer \((N/2 + 1)\) where \(N\) is total number of terms in the series. However, the optimal and effective lengths of data intervals and partitions number of universe of discourse are not yet resolved mathematically and no standard rule can be followed.

In this paper, we described the inter-quartile range approach to determine the effective length of intervals and partitions number of universe of discourse in FTS. Through this approach, the length of intervals is assigned by using the sum of data ranges from \(Q_1\) (quartile-1) and data minimum, \(Q_2\) (quartile-2) and \(Q_1\), \(Q_3\) (quartile-3) and \(Q_2\), data maximum and \(Q_3\) will be divided with the highest repetitive of frequency data among quartiles. By using this approach, the determining of length intervals and dividing of partitions number are more simplified and easier to use when compared with some related methods. Moreover, the data differencing are carried out in a forecasting model in order to achieve at the highest level of forecasting accuracy. To verify the efficiency of the proposed method, yearly enrollment data of Alabama University are used as an empirical analysis and the forecasting results are compared with some existing proposed methods.

The rest of paper is organized as follows: In Sec. 2, the fundamental theories in FTS are described. The proposed method is presented in Sec. 3. In Sec. 4, the algorithm for enrollment forecasting is explained. The discussion and a brief conclusion are explored in final section.
2. The Fundamental Theories in FTS

In Refs. 2–5 and 16, the fundamental concepts of FTS have been established originally as:

**Definition 1.** Fuzzy time series (Refs. 2–4)

Let \( Y(t) \) (\( t = 0, 1, 2, \ldots \)), a subset of real numbers, be the universe of discourse on which the fuzzy sets \( f_i(t) \), \( i = 1, 2, \ldots \) are defined in the universe of discourse \( Y(t) \) and \( F(t) \) is a collection of \( f_i(t) \), \( i = 1, 2, \ldots \). Then \( F(t) \) is an FTS defined on \( Y(t) \) (\( t = 0, 1, 2, \ldots \)). Therefore, \( F(t) \) is the linguistic time series variable, where \( f_i(t) \), \( i = 1, 2, \ldots \), are the possible linguistic values of \( F(t) \).

**Definition 2.** First order fuzzy relation (Refs. 2–4)

Suppose \( F(t) \) is caused by \( F(t - 1) \) denoted by \( F(t - 1) \rightarrow F(t) \), then this relationship can be represented by:

\[
F(t) = F(t - 1) \circ R(t, t - 1)
\]

where “ \( \circ \) ” represent an operator, \( R(t, t - 1) \) is a fuzzy relationship between \( F(t) \) and \( F(t - 1) \). \( F(t - 1) \) is the first-order model of \( F(t) \).

**Definition 3.** Fuzzy logical relationships (FLRs) (Ref. 5)

Let \( F(t - 1) = A_i \) and \( F(t) = A_j \). The relationship between two consecutive data (called a fuzzy logical relationship, FLR), i.e., \( F(t) \) and \( F(t - 1) \), can be denoted by \( A_i \rightarrow A_j, i, j = 1, 2, \ldots, p \) (where \( p \) is the interval or the subinterval number) is called the left-hand side (LHS) of the FLR, and \( A_j \) is the right-hand side (RHS) of the FLR.

**Definition 4.** Fuzzy logical groups (FLG) (Ref. 16)

Let \( A_i \rightarrow A_j, A_i \rightarrow A_k, \ldots, A_i \rightarrow A_p \) be the FLRs with the same LHH which are grouped into an ordered FLG (called a fuzzy logical group) by putting all their RHS together as on the RHS of the FLG. It can be written as below:

\[
A_i \rightarrow A_j, A_k, \ldots, A_p \quad i, j, k, \ldots, p = 1, 2, \ldots, n (n \in \mathbb{N}).
\]

There are five major steps in developing the FTS as proposed in Ref. 2, which are:

- Define the universe of discourse (\( U \)) and divide it into several equal length intervals.
- Fuzzify each interval into linguistic time series values (\( A_i, i = 1, 2, \ldots, p, p \) is the partition number).
- Establish fuzzy logical relationships among linguistic time series values (\( A_i \rightarrow A_j, i, j = 1, 2, \ldots, p \))
- Establish forecasting rule.
- Determine the forecast value.
3. The Proposed Approach Based on Inter-Quartile Range

In FTS forecasting, the forecast values are very dependent with the length of intervals and partitions number of universe of discourse. Basically in grouped data, length of interval is defined as (Ref. 30):

\[ l = \frac{R}{k}, \]  

(3)

where \( l \) is a length of interval, \( R \) is a range of data, and \( k \) is a number of classes (partitions number). This reason encourages us to investigate the appropriate and effective lengths of interval based on the inter-quartile range approach. Moreover, range data and the quartiles range are defined as (Ref. 31):

(i) Range \((R)\) is the difference between the largest \((X_{\text{max}})\) and smallest \((X_{\text{min}})\) observations in a set of data that is,

\[ R = X_{\text{max}} - X_{\text{min}}. \]  

(4)

(ii) The first quartile \((Q_1)\) is the value such that 25.0 percent of the observations are smaller and 75.0 percent larger,

\[ Q_1 = \frac{(n + 1)}{4}. \]  

(5)

(iii) The second quartile \((Q_2)\), is the value such that 50.0 percent of the observations are smaller and 50.0 percent are larger,

\[ Q_2 = \frac{(n + 1)}{2}. \]  

(6)

(iv) The third quartile \((Q_3)\), is the value such that 75.0 percent of the observations as smaller and 25.0 percent larger,

\[ Q_3 = \frac{3(n + 1)}{4}. \]  

(7)

Let \( X_t \) \((t = 1, 2, \ldots, n)\) be a time series data and \((X_{\text{min}}, X_{\text{max}}) \in X_t\). Define the universe of discourse \( U \) as \([X_{\text{min}} - P_1, X_{\text{max}} + P_2]\) where \( P_1 \) and \( P_2 \) are proper positive numbers. Moreover, \( U \) also can be presented as \([a, b]\). Completely, the effective interval length and partition number can be determined by using the proposed approach as follows:

**Step 1:** Determine \(Q_1\), \(Q_2\) and \(Q_3\) by using Eqs. (8)-(10).

**Step 2:** Divide \( U \) into four lengths of intervals by using quartiles as in Table 1.

<table>
<thead>
<tr>
<th>Interval</th>
<th>Quartile Ranges</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>([a, Q_1])</td>
<td>(R_1)</td>
<td>(f_1)</td>
</tr>
<tr>
<td>([Q_1, Q_2])</td>
<td>(R_2)</td>
<td>(f_2)</td>
</tr>
<tr>
<td>([Q_2, Q_3])</td>
<td>(R_3)</td>
<td>(f_3)</td>
</tr>
<tr>
<td>([Q_3, b])</td>
<td>(R_4)</td>
<td>(f_4)</td>
</tr>
</tbody>
</table>
Step 3: Divide each range of intervals as mentioned in Table 1 by using the highest repetitive frequency \((f_h)\).

\[
l_1 = R_1/f_h, \quad l_2 = R_2/f_h, \quad l_3 = R_3/f_h, \quad l_4 = R_4/f_h.
\]

Thus, the effective length of interval \((l)\) and partition number \((p)\) can be written as:

\[
l = \text{Average}(l_1, \ldots, l_4) \quad (8)
\]

\[
p = 4 \times f_h \quad (9)
\]

Step 4: Re-divide each interval in Table 1 into equally length sub-intervals by using effective length of interval as calculated in Eq. (8). On the other hand, Eq. (9) can be used to determine partition number of interval.

4. Enrollment Forecasting

This section introduces the proposed algorithm for enrollment forecasting, by applying the same yearly enrollment data of Alabama University from 1971 until 1992 as in Refs. 2, 4, 5, 14, 25–28. This algorithm can be derived by following steps as follow:

Step 1: Compute \(Q_1, Q_2\) and \(Q_3\) and divide universe of discourse \(U\) by using proposed method as mentioned in Sec. 3. The results can be presented in Table 2. Table 2 shows \(f_1 = f_3 = f_4 = 5\), then lengths of interval \(l_1 = 2104/5, l_2 = 573/5, l_3 = 1495/5\), and \(l_4 = 2110/5\). Thus, the effective length of interval \(= \text{average of } (l_1, \ldots, l_4) = 341.10\). The total of partitions number \(= 5 + 5 + 5 + 5 = 20\) intervals:

\[
u_1 = [13054.10, 13368.26], u_2 = [13368.26, 13682.46], \ldots, u_{20} = [19481.10, 19882.10].
\]

Furthermore, there are also 20 of midpoints intervals:

\[
m_1 = (13054.10 + 13368.20)/2, \ldots, m_{20}
\]

\[
= (19481.10 + 19882.10)/2 \text{ or } 13211.5, \ldots, 19651.10.
\]

Step 2: Define fuzzy sets for observations. Each linguistics observation \(A_i\) can be defined by intervals \(u_1, \ldots, u_{20}\), respectively. Each \(A_i\) can be represented as in the

<table>
<thead>
<tr>
<th>Quartile</th>
<th>Interval</th>
<th>Frequency</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X_{\text{min}} = 13055)</td>
<td>[13054.10–15159.10]</td>
<td>5</td>
<td>2104</td>
</tr>
<tr>
<td>(Q_1 = 15159)</td>
<td>[15159.10–15732.10]</td>
<td>7</td>
<td>573</td>
</tr>
<tr>
<td>(Q_2 = 15732)</td>
<td>[15732.10–17227.10]</td>
<td>5</td>
<td>1495</td>
</tr>
<tr>
<td>(Q_3 = 17227)</td>
<td>[17227.10–19337.10]</td>
<td>5</td>
<td>2110</td>
</tr>
<tr>
<td>(X_{\text{max}} = 19337)</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
following Eq. (10), and the value, $k_j$, is determined by:

- If $j = i - 1$, then $k_j = 0.5$;
- If $j = i$, then $k_j = 1$;
- If $j = i + 1$ then $k_j = 0.5$; elsewhere $k_j = 0$.

$$A_i = \sum_{j=1}^{n} \frac{k_j}{u_j}$$  \hspace{1cm} (10)

**Step 3:** Transform the yearly enrollment data into linguistic time series values and establish FLRs as in Table 3.

**Step 4:** Calculate the forecast values by using proposed method as follows: In time series analysis, the first differencing is written as (Ref. 32) Hanke and Wichern, 2009):

$$\Delta^1 d = X_t - X_{(t-1)}.$$  \hspace{1cm} (11)

By using Eq. (7), the forecast values of $(X_t)$ can be presented as:

$$X_t = X_{(t-1)} + \Delta^1 d.$$  \hspace{1cm} (12)

In this proposed method, the forecast value can be modeled by using midpoints of intervals from first-order fuzzy logical relationship as below:

Let $A_i \rightarrow A_j$ is an FLRs, where $A_i, A_j$ are present and past linguistic time series values at time $(t)$ and $(t - 1)$. Then, the midpoints of $A_i$ and $A_j$ are $m_i$ and $m_j$. By

<table>
<thead>
<tr>
<th>Year</th>
<th>Actual Data</th>
<th>Linguistic Values</th>
<th>FLRs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1971</td>
<td>13055</td>
<td>$A_1$</td>
<td>*</td>
</tr>
<tr>
<td>1972</td>
<td>13563</td>
<td>$A_2$</td>
<td>$A_1 \rightarrow A_2$</td>
</tr>
<tr>
<td>1973</td>
<td>13867</td>
<td>$A_3$</td>
<td>$A_2 \rightarrow A_3$</td>
</tr>
<tr>
<td>1974</td>
<td>14696</td>
<td>$A_6$</td>
<td>$A_3 \rightarrow A_6$</td>
</tr>
<tr>
<td>1975</td>
<td>15460</td>
<td>$A_8$</td>
<td>$A_6 \rightarrow A_8$</td>
</tr>
<tr>
<td>1976</td>
<td>15311</td>
<td>$A_7$</td>
<td>$A_8 \rightarrow A_7$</td>
</tr>
<tr>
<td>1977</td>
<td>15603</td>
<td>$A_8$</td>
<td>$A_7 \rightarrow A_8$</td>
</tr>
<tr>
<td>1978</td>
<td>15861</td>
<td>$A_9$</td>
<td>$A_8 \rightarrow A_9$</td>
</tr>
<tr>
<td>1979</td>
<td>16807</td>
<td>$A_{12}$</td>
<td>$A_9 \rightarrow A_{12}$</td>
</tr>
<tr>
<td>1980</td>
<td>16919</td>
<td>$A_{12}$</td>
<td>$A_{12} \rightarrow A_{12}$</td>
</tr>
<tr>
<td>1981</td>
<td>16388</td>
<td>$A_{10}$</td>
<td>$A_{12} \rightarrow A_{10}$</td>
</tr>
<tr>
<td>1982</td>
<td>15433</td>
<td>$A_8$</td>
<td>$A_{10} \rightarrow A_8$</td>
</tr>
<tr>
<td>1983</td>
<td>15497</td>
<td>$A_8$</td>
<td>$A_8 \rightarrow A_8$</td>
</tr>
<tr>
<td>1984</td>
<td>15145</td>
<td>$A_7$</td>
<td>$A_8 \rightarrow A_7$</td>
</tr>
<tr>
<td>1985</td>
<td>15163</td>
<td>$A_7$</td>
<td>$A_7 \rightarrow A_7$</td>
</tr>
<tr>
<td>1986</td>
<td>15984</td>
<td>$A_9$</td>
<td>$A_7 \rightarrow A_9$</td>
</tr>
<tr>
<td>1987</td>
<td>16859</td>
<td>$A_{12}$</td>
<td>$A_{12} \rightarrow A_{12}$</td>
</tr>
<tr>
<td>1988</td>
<td>18150</td>
<td>$A_{16}$</td>
<td>$A_{12} \rightarrow A_{16}$</td>
</tr>
<tr>
<td>1989</td>
<td>18970</td>
<td>$A_{18}$</td>
<td>$A_{16} \rightarrow A_{18}$</td>
</tr>
<tr>
<td>1990</td>
<td>19328</td>
<td>$A_{19}$</td>
<td>$A_{18} \rightarrow A_{19}$</td>
</tr>
<tr>
<td>1991</td>
<td>19337</td>
<td>$A_{19}$</td>
<td>$A_{19} \rightarrow A_{19}$</td>
</tr>
<tr>
<td>1992</td>
<td>18876</td>
<td>$A_{18}$</td>
<td>$A_{19} \rightarrow A_{18}$</td>
</tr>
</tbody>
</table>
using Eq. (12), the forecast value at time $t$ can be written as:

$$X_t = X_{t-1} + \Delta^1 d$$

$$F(A_i) = F(A_j) + \Delta^1 d$$

$$F(A_i) = m_j + (X_t - X_{t-1})$$  \hspace{1cm} (13)
Equation (13) shows the value of $F(A_j)$ is represented by midpoint $m_j$. Moreover, the forecast values of enrollment are presented in Table 4. The actual enrollment and forecast values are also presented in Fig. 1.

**Step 5:** Compute the error of forecasting by using mean square error (MSE). Furthermore, MSE of proposed method is compared with MSE of some existing methods as in Table 5.

5. Results and Discussion

With the proposed method, it is easier to group data based on quartiles range, and then re-dividing them into sub-intervals by using frequency of spread data from quartiles. Thus, each data can be distributed in the optimal interval, with no empty interval obtained. Moreover, the first differencing approach is used to obtain better forecast. In this study, the application of the first differencing does not remove the trends of data, but it multiplied with midpoint of first-order of FLRs in forecasting model. From this, the proposed method is able to decrease the value of MSE significantly.

Table 5 shows MSE of existing methods are varied from 418,184 until 60,714. These values indicate that to achieve the high level of forecasting accuracy is not easy task, although many contributions have been explored in determining of effective length interval and partition number of universe of discourse. From Table 5, it indicates that MSE of the proposed method is the smallest as compared to other methods. It reduces approximately seven times smaller than the method proposed in Ref. 25. Moreover, this value explains that the proposed method is able to improve the forecasting accuracy significantly.

<table>
<thead>
<tr>
<th>No.</th>
<th>Methods</th>
<th>MSE</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Song and Chisson (1993a)</td>
<td>412,499</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>Song and Chisson (1994)</td>
<td>775,687</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>Sullivan &amp; Woodall (1994)</td>
<td>386,055</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>Chen (1996)</td>
<td>407,507</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>Huang (2001)</td>
<td>78,792</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>Chen (2002)</td>
<td>86,694</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>Aladag et al. (2009)</td>
<td>78,073</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>Egrioglu et al. (2009)</td>
<td>66,661</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>Engrioglu et al. (2010)</td>
<td>60,714</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>Pathak and Singh (2011)</td>
<td>418,184</td>
<td>11</td>
</tr>
<tr>
<td>11</td>
<td>Shah (2012)</td>
<td>61,032</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>Our proposed method</td>
<td>8,293</td>
<td>1</td>
</tr>
</tbody>
</table>

6. Conclusion

Various methods have been presented to investigate the length of data interval and partition number of universe of discourse in FTS to achieve high level forecasting
accuracy. These will resolve the fluctuations of FTS as mentioned in Refs. 14 and 25. However, it still has difficulties in choosing the optimal length of interval. In this paper, we propose a new method to adjust the length of interval and partition number based on inter-quartile range. Furthermore, in forecasting model, the first differencing is also considered to obtain better forecast values. Our method gives very significant effect to reduce the enrollment forecasting error, such as seven times lower as compared to a method proposed in Ref. 25, whilst more than seven times lower as compared to other methods as shown in Table 5. Furthermore, the study of effective lengths interval and forecasting model should be enlarge for future research.

References