2013 12th International Conference on Machine Learning and Applications

Preprocessing in Fuzzy Time Series to Improve the Forecasting Accuracy

Fábio José Justo dos Santos
Computer Department
[1]Federal University of São Carlos; [2]Federal Institute of Education, Science and Technology of São Paulo
[1]São Carlos, [2]Araraquara, Brazil
fabio_santos@dc.ufscar.br

Heloisa de Arruda Camargo
Computer Department
Federal University of São Carlos
São Carlos, Brazil
heloisa@dc.ufscar.br

***Abstract*—The preprocessing in fuzzy time series has an important role to improve the forecast accuracy. The definitions of domain, number of linguistic terms and of the membership function to each fuzzy set, has direct influence in the forecast results. Thus, this paper has the focus on definition of these parameters, before of performing the prediction. The experimental results in enrollments time series show that, when the forecast is performed after proposed preprocessing, the accuracy rate is improved.**

Keywords: preprocessing, fuzzy time series, forecasting

I. INTRODUCTION

Introduced in [1], [2] and [3], Fuzzy Time Series (FTS) arise to overcome two weak points of statistical methods: to deal with data represented by linguistic terms and to do forecasting with little data set.

In [4a], the authors use the membership function in FTS to generate a weighted average forecast. In [5a] is presented a forecast model to multi-attribute FTS of second order. A forecasting model for FTS using k-means was proposed in [6a]. The model perform a weighted forecast defined by fuzzy logical relationship. The applications of these methods are in several areas as enrollments forecasting [2], [3], [7a], temperature prediction [8a], stock index forecasting [9a], electricity demand [10a], [11a], and others.

In the literature, several studies make prediction values, without the preprocessing of the data set [12a], [13a], [14a], [15a]. However, the preprocessing in FTS has an important role to improve the forecasting accuracy rate. Without the preprocessing, the forecast in FTS can be prejudiced by outliers and by fuzzy sets that can do not represent faithfully the data sets in the time series.

According to [16], forecasting accuracy can be enhanced with the use of techniques to define the amount and the support of fuzzy sets to each variable in the FTS. In this paper, we show a preprocessing model which aims to improve the forecasting accuracy performing, identification and exclusion of outliers in time series, defining the amount of linguistic terms that will represent the FTS and defining the support of these terms. To attain these objectives, is proposed the combination of different methods. To demonstrate the efficacy of the proposed method, are performed experiments that confirm the improvement expected in the accuracy after the implementation of the proposed preprocessing model.

The rest of this paper is organized as follows. In section II are introduced the basic concepts of Fuzzy Time Series. The forecasting model used in experiments is shown in section III. The proposed method is explained in section IV. The experiments and conclusions are presented in sections V and VI, respectively.

II. FUZZY TIME SERIES

Based on Zadeh's works [17], [18], Fuzzy Time Series was introduced to deal time series where datum are represented by fuzzy sets, instead of crisp values. The basic definitions are:

Definition 1. Let $Y(t)$ $(t = 0,1,2,\dots)$ be a subset of real numbers. Assume that fuzzy set $f_i(t)$, where $i = 0,1,2,$ ..., is defined in the universe of discourse $Y(t)$. If $F(t)$ is the collection of $f_i(t)$, then $F(t)$ is called a fuzzy time series on $Y(t)$.

Definition 2. If $F(t)$ is caused by $F(t-1)$, i.e., $F(t-1) \rightarrow F(t)$, this fuzzy relationship can be denoted by $F(t) = F(t-1) \circ R(t,t-1)$, where $R(t,t-1)$ is a fuzzy relationship between $F(t)$ and $F(t-1)$, R is the set of relationships joined by the max–min operator denoted by "∘".

Definition 3. If $F(t)$ is caused by n fuzzy sets, i.e., $F(t-n), F(t-n+1), \dots, F(t-1)$, the fuzzy relationship is represented by $A_{i_1}, A_{i_2}, \dots, A_{i_n} \rightarrow A_j$, where, $F(t-n) = A_{i_1}, F(t-n+1) = A_{i_2}$ and $F(t-1) = A_{i_n}$. This relationship is called nth order fuzzy time series model.

Definition 4. All fuzzy logical relationships with the same left-hand sides, can be grouped together into different fuzzy logical relationship groups. If there are two fuzzy logical relationships $A_i \rightarrow A_{j_1}$ and $A_i \rightarrow A_{j_2}$, these two fuzzy logical relationships can be grouped as $A_i \rightarrow A_{j_1}, A_{j_2}$.

The original model proposed in [1], [2] and [3] to predict a number, include six steps: (1) define the universe of discourse; (2) define the linguistic terms, i.e., fuzzy sets, to represent the time series values; (3) fuzzifying the time series; (4) deriving fuzzy logical relationships; (5) forecast, and (6) defuzzifying the forecasting outputs.

III. FORECASTING MODEL

Base on the model proposed in [1], several studies arose to deal with forecasting. In this section, we introduce the fuzzy time series model proposed in [19], which will be used to compare the accuracy of forecasting methods without and

978-0-7695-5144-9/13 $31.00 © 2013 IEEE
DOI 10.1109/ICMLA.2013.185

with the execution of preprocessing model proposed in this paper. In [19], the prediction is performed as follows:

***Step 1*:** Cluster time series Y into c clusters to identify patterns.

Step 2: Rank each cluster and fuzzify the time series $Y(t)$ as fuzzy time series $F(t)$.

Step 3: Establish fuzzy relationships and the fuzzy relationship groups.

Step 4: Forecast and defuzzify the possible outcomes based on fuzzy following rules:

- If the fuzzy relationship group of A_i is empty, such as $A_i \rightarrow \emptyset$, then the forecasting of $F(t+1)$ is defined by cluster center of linguistic term A_i.
- If the fuzzy relationship is one-to-one, such as $A_i \rightarrow A_j$, them the forecasting of $F(t+1)$ is defined by cluster center of linguistic term A_j.
- If the fuzzy relationship group is one-to-many, such as $A_i \rightarrow A_{j_1}, A_{j_2}, A_{j_3}, \ldots A_{j_n}$, then the forecasting of $F(t+1)$ assume that each linguistic term has the same possibility. Therefore, the arithmetic mean of the n cluster centers is defined as the forecasted value.

IV. PREPROCESSING MODEL PROPOSED

In this section, we propose a preprocessing model to Fuzzy Time Series in order to: (1) identify and remove outliers; (2) define the time series domain; (3) define the number of linguistic terms, i.e., fuzzy sets to each variable; (4) define the support and of crisp value which will represent the linguistic terms. This model is based on different methods available in literature and can be represented as showed in the Figure 1.

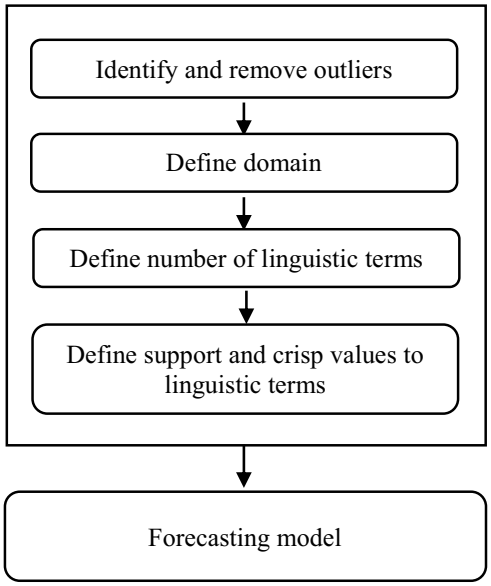

Figure 1. Steps of proposed preprocessing with a forecasting model

The preprocessing showed in Figure 1, is essential to enhance the forecasting accuracy in a FTS. The existence of one or more outliers in sample data, can exert a negative influence in the definition of the number of linguistic terms and in the definition of the support and crisp values that will represent each these linguistic terms.

For outlier detection [20], consider the existence of n historical data in $Y(t)$, where $t = 0,1,2, \ldots, n$. The first step to detect an outlier is define the square matrix H as follows:

$$H = x(x^T x)^{-1} x^T, \tag{1}$$

where

$$x = \begin{bmatrix} 1 & d_1 \\ 1 & d_2 \\ 1 & d_3 \\ \vdots & \vdots \\ 1 & d_n \end{bmatrix}. \tag{2}$$

The next step is to compute the *Residual Student* for each sample in the time series, as showed in the sequence.

$$RStudent_i = \frac{e_i}{\hat{\sigma}^{(i)}\sqrt{1-h_i}} \tag{3}$$

In (3), $\hat{\sigma}^{(i)}$ is the standard deviation without the i th element of time series, h_i is the ith diagonal element in the matrix H and e_i is defined by (4), where d_i is ith element of time series. In this paper, for the ith value be consider an outlier, the $RStudent_i$ should be equal or larger than 2.5.

$$e_i = d_i - \frac{\sum_{k=1}^{n} d_k}{n} \tag{4}$$

After identifying and removing the outliers, the universe of discourse is defined by (5), where D_{min} is the min value in the sample data, D_{max} is the max and σ is the standard deviation of the data.

$$D = [D_{min} - \sigma, D_{max} + \sigma] \tag{5}$$

Based in [21], after defined the time series domain, the algorithm below is applied to calculate the number of linguistic terms that will represent the data in the time series.

Step 1. Sort the n numerical data in an ascending sequence as $d_1, d_2, \ldots, d_i, \ldots, d_n$, where d_1 is the smallest datum among the n numerical data, d_n is the largest datum among the n numerical data, and $1 \leq i \leq n$.

Step 2. Calculate the threshold of condition stop as in (6).

$$\tau = \frac{\sum_{i=1}^{n-1}(d_{i+1} - d_i)}{n-1} \tag{6}$$

Step 3. Put each numerical datum into a cluster as $\{d_1\}, \{d_2\}, \ldots, \{d_i\}, \ldots, \{d_n\}$, where the symbol "{ }" denotes a cluster.

Step 4. Assume that there are p clusters. Calculate the cluster center to each as in (7), where m is the number of

elements in $cluster_k$, d_j are the data in the $cluster_k$, r is the number of data in $cluster_k$, and $1 \leq k \leq p$.

$$cluster\_center_k = \frac{\sum_{j=1}^{m} d_j}{r}, \quad (7)$$

Step 5. Calculate the distance between cluster m and $m + 1$, where $m = 1, 2, \dots, p - 1$.

$$distance_{m,m+1} = |cluster\_center_m - cluster\_center_{m+1}| \quad (8)$$

Step 6. If the shortest distance between the clusters is less than τ, then combine the clusters having the smallest distance between them into a cluster and go back to Step 4.

Fuzzy sets with different sizes can reflect the real internal structure of the data and to produce better results than equally spaced sets. Thus, in this paper the support and the crisp values that represent the linguistic terms, are defined by k-means clustering. The parameter p resulting from the previous algorithm, indicate the number of clusters to the k-means. From the p cluster centers found by k-means $(c_1, c_2, c_3, \dots, c_p)$, the fuzzy sets are defined as follows:

- For linguistic terms L_j, where $1 < j < p$, the center of triangular fuzzy membership function, i.e., the point where the pertinence degree is equal 1, is represented by center of cluster j.
- The left-shoulder and the right-shoulder of linguistic terms L_j, where $1 < j < p$, are defined by the cluster center $j - 1$ and $j + 1$, respectively
- For L_1, the trapezoidal fuzzy set are used. The right-shoulder is the center of L_2. D_{min}, i.e., the smallest value of the domain, is used to the left-shoulder and to the left parameter where the pertinence is equal 1. To the right parameter where the pertinence is equal 1, c_1 is used.
- For L_p, the trapezoidal fuzzy set is used too. The left-shoulder is the center of L_{p-1} and D_{max}, i.e., the highest value of the domain, defines the right-shoulder and the right parameter where the pertinence is equal 1. The left-shoulder where the pertinence is equal 1 is defined by c_p.

After the preprocessing, are performed the fuzzyfication of the time series, the deriving fuzzy logical relationships and the forecasting. In the next section, the experiments show an improvement in the forecasts accuracy.

V. Experiments with Proposed Preprocessing Model

The experiments are performed with the yearly data on enrollments at the University of Alabama between 1971 and 1992. To demonstrate the efficacy of the proposed preprocessing model, the forecasting is performed and compared in a scenario with and without outliers in the dataset. The values are shown in Table I.

In Table II are compared the crisp results that represent each linguistic term after clustering data with and without outliers. According to proposed method, 10 and 11 linguistic terms are used to deriving fuzzy logical relationships in enrollments time series with and without outliers, respectively. In [19], the authors use 7 linguistic terms in original dataset, which are maintained here to the both scenarios.

TABLE I. Enrollments

Year	Enrollments	Enrollments With Outliers
1971	13055	13055
1972	13563	13563
1973	13867	13867
1974	14696	14696
1975	15460	15460
1976	15311	15311
1977	15603	15603
1978	15861	15861
1979	16807	16807
1980	16919	16919
1981	16388	16388
1982	15433	15433
1983	15497	15497
1984	15145	25145*
1985	15163	15163
1986	15984	15984
1987	16859	16859
1988	18150	18150
1989	18970	18970
1990	19328	19328
1991	19337	19337
1992	18876	18876

* outlier in dataset

The actual enrollments and the forecasting values are shown in the Table III. The columns "Forecast Without Preprocessing" and “Forecast After Proposed Preprocessing” shown the forecast results according to [19], nevertheless, the results in the second column were calculated after applying the proposed method.

TABLE II. Cluster Centers

Features	Cluster Centers According to [19]		Cluster Centers With The Proposed Preprocessing Model	
	With Outliers	*Original Dataset*	*With Outliers*	*Original Dataset*
Domain	—	—	[11280;21112]	[11888;22555]
L_1	13619	13459	13309	13055
L_2	15135	14702	13867	13715
L_3	15496	15385	14696	14696
L_4	16061	15950	15373	15206
L_5	16888	16835	15816	15498
L_6	18971	18161	16388	15993
L_7	24990	19144	16862	16388
L_8	—	—	18150	16861
L_9	—	—	18923	18150
L_{10}	—	—	19332	18923
L_{11}	—	—	—	19333

The average error percentage (MAER) and mean square error (MSE) are used to demonstrate the improvement in

forecasting. The statistical numbers showed in Table IV, confirm the better forecasting accuracy in time series after the performing the proposed preprocessing, especially in scenarios with outliers.

TABLE III. FORECASTING ENROLLMENTS OF THE UNIVERSITY OF ALABAMA

Year	Enrollments	Forecast Without Preprocessing		Forecast After Proposed Preprocessing	
		With Outlier	Original Datase	With Outlier	Original Datase
1971	13055				
1972	13563	14377	14242	13588	13715
1973	13867	14377	14242	13588	14206
1974	14696	14377	14242	14696	14206
1975	15460	15778	15474	15373	15498
1976	15311	17920	15474	15595	15566
1977	15603	15778	15474	15595	15566
1978	15861	17920	15474	16339	15566
1979	16807	16192	16146	16339	16862
1980	16919	17307	16988	17133	17133
1981	16388	17307	16988	17133	17133
1982	15433	16192	16146	15373	15498
1983	15497	17920	15474	15595	15566
1984	15145	17920	15474	*	15566
1985	15163	15135	15474	15595	15566
1986	15984	15778	15474	15595	15566
1987	16859	16192	16146	16339	16862
1988	18150	17307	16988	17133	17133
1989	18970	18971	19144	18923	18923
1990	19328	18971	19144	19332	19333
1991	19337	18971	19144	19128	19128
1992	18876	18971	19144	19128	19128

* not performed to outliers

TABLE IV. RESULTS AND COMPARATION FOR ENROLLMENTS FORECASTING

Measure	Forecast Without Preprocessing [19]		Forecast After Proposed Preprocessing	
	With Outlier	Original Datase	With Outlier	Original Datase
MAER	5,23%	2,40%	1,70%	1,57%
MSE	1390438	228918	148992	133513

VI. CONCLUSION

This preprocessing method to fuzzy time series was proposed to enhance the forecasting accuracy. An important feature to forecasting in fuzzy time series, is the definition of numbers of linguistic terms and their respective crisp values. A big number of linguistic terms, can induce an overfitting. A little number of these terms, can produce a poor forecasting. Moreover, the presence of outliers in the time series, can results in a negative influence to the forecast. The experimental results demonstrate that the proposed method improves the forecasting accuracy rate both in time series with outliers, as in time series without outliers. In the future, will be proposed a new method to forecasting added to preprocessing model proposed in this paper.

REFERENCES

[1] Q. Song and B. S. Chissom, “Fuzzy time series and its models,” Fuzzy Sets and Systems, vol. 54, pp. 269–277, 1993.

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

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

[4] S. Bai-qing, X. Shan and W. Berlin, “ Fuzzy Time Series Models for LNSZZS Forecasting,” Journal of Convergence Information Technology (JCIT), vol. 7, No. 19, Oct. 2012, pp. 470-478.

[5] S. Chatterjee, S. Nigam, J.B. Singh and L.N. Upadhyaya, "Application of fuzzy time series in prediction of time between failures & faults in software reliability Assessment," Fuzzy Information and Engineering, vol. 3, 2011, pp. 293-309, doi: 10.1007/s12543-011-0084-7.

[6] C. Kai, F. Fang-Ping and C. Wen-Gang, "A novel forecasting model of fuzzy time series based on K-means clustering," Proc. Second International Workshop on Education Technology and Computer Science, 2010, pp. 223-225, doi: 10.1109/ETCS.2010.249.

[7] S. S. Gangwar and S. Kumar, “Partitions based computational method for high-order fuzzy time series forecasting,” Expert Systems with Applications, vol. 39, 2012, pp. 12158–12164.

[8] N. Y. Wang and S. M. Chen, “Temperature prediction and TAIFEX forecasting based on automatic clustering techniques and two-factors high-order fuzzy time series,” Expert Systems with Applications, vol. 36, no. 2, Mar. 2009, pp. 2143-2154.

[9] S. Chen and C. Chen, “TAIEX forecasting based on fuzzy time series and fuzzy variation groups,” IEEE Transactions on Fuzzy Systems, vol. 19, Febr. 2011, No. 1, pp. 1-12.

[10] W. Shen, V. Babushkin, Z. Aung and W. L. Woon, “An ensemble model for day-ahead electricity demand time series forecasting,” Proc. of the fourth international conference on Future energy systems, 2013, pp. 51-62, doi: 10.1145/2487166.2487173.

[11] F. M. Álvarez, A. Troncoso, J. C. Riquelme and J. S. A. Ruiz, “Energy time series forecasting based on pattern sequence similarity,” IEEE Transactions on Knowledge and Data Engineering, vol. 23, No. 8, 2011, pp. 1230-1243.

[12] B. P. Joshi and S. Kumar, “A computational method for fuzzy time series forecasting based on difference parameters,” International Journal of Modeling, Simulation, and Scientific Computing, vol. 4, No. 1, 2013, pp. 1250023-1 – 1250023-12, doi: 10.1142/ S1793962312500237.

[13] H. Chu, T. Chen, C. Cheng and C. Huang, “Fuzzy dual-factor time-series for stock index forecasting,” Expert Systems with Applications, vol. 36, 2009, pp. 165–171.

[14] W. Qiu, X. Liu and H. Li, “A generalized method for forecasting based on fuzzy time series,” Expert Systems with Applications, vol. 38, 2011, pp. 10446–10453.

[15] S. R. Singh, “A computational method of forecasting based on fuzzy time series,” Mathematics and Computers in Simulation, vol. 79, 2008, pp. 539–554.

[16] S. Chen and K. Tanuwijaya, “Multivariate fuzzy forecasting based on fuzzy time series and automatic clustering techniques,” Expert Systems with Applications, vol. 38, 2011, pp. 10594–10605.

[17] L. A. Zadeh, "Fuzzy set," Information and Control, vol. 8, 1965, pp. 338–353.

[18] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning - part 1," Information Sciences, vol. 8, 1975, pp. 199–249.

[19] C. Cheng, G. Cheng and, J. Wang, “Multi-attribute fuzzy time series method based on fuzzy clustering,” Expert Systems with Applications, vol. 34, 2008, pp. 1235–1242.

[20] V. Barnett and T. Lewis, “Outliers in statistical data”, 3rd edition, 1994, NY: John Wiley & Sons.

[21] K. Tanuwijaya and S. Chen, “A new method to forecast enrollments using fuzzy time series and clustering techniques,” Proc. of the Eighth International Conference on Machine Learning and Cybernetics, jul. 2009, pp. 12-15.