Available online at www.sciencedirect.com
ScienceDirect
Mathematics and Computers in Simulation 162 (2019) 18–30
MATHEMATICS AND COMPUTERS IN SIMULATION
www.elsevier.com/locate/matcom

Original articles

Time series forecasting for stock market prediction through data discretization by fuzzistics and rule generation by rough set theory

Shanoli Samui Pal, Samarjit Kar*

Department of Mathematics, NIT Durgapur, Durgapur 713209, West Bengal, India

Received 11 April 2017; received in revised form 21 June 2018; accepted 3 January 2019
Available online 21 January 2019

Abstract

Data discretization is a preprocessing technique to mine essential information from the pool of information. It is also essential to generate rules from the processed data after mining information. In this paper, a hybrid approach is proposed to forecast time series of stock price by using data discretization based on fuzzistics (Mendel, 2007 [24]; Liu and Mendel, 2008), where cumulative probability distribution approach (CPDA) is used to get the intervals for the linguistic values. First order fuzzy rule generation and reduction of rule sets by rough set theory have been performed. Thereafter, forecasting of the time series data is computed from defuzzification using reduced rule base and its historical evidences. Proposed approach is applied on stock index closing price of three time series data (BSE, NYSE, and TAIEX) as experimental data sets and the results show that the method is more effective than its counter parts.

Keywords: Time series forecasting; Fuzzistics; First order fuzzy logic; Rough set based rule reduction; Stock market data

1. Introduction

Stock market is unpredictable as there are several complex factors influencing its ups and downs. Therefore the trend of the series is also affected by those factors and by their non-linear relationship. In the stock market forecasting, the technical analysis is one of the traditional methods applied by investors for decision making. There are some models predicting the distribution of stock returns [23], predicting the stock price index [25,35], providing volatility measure [6,12]. All these models are different types of regression models assuming some mathematical distribution. This distributions are not always followed by realistic stock market time series data.

Nowadays, several soft computing approaches like evolutionary algorithms [34], artificial neural networks [16,36], fuzzy logic [15,21], rough set theory [10] and their hybridization [27,28,35] have been developed and perform well in forecasting of stock markets. Back propagation neural network had been used to find the fuzzy relationship in fuzzy time series [16]. A hybridized genetic algorithm and neural network model had been developed to predict stock price index [25]. Caia [7] proposed a hybrid GA model based on fuzzy time series and genetic algorithm (FTSGA) on Taiwan stock exchange capitalization weighted stock index (TAIEX) as experimental data set and concluded that the model improved the accuracy. Teoh et al. [35] proposed a hybrid model based on multi-order fuzzy time series

* Corresponding author.
E-mail addresses: shanoli.pal@gmail.com (S.S. Pal), samarjit.kar@maths.nitdgp.ac (S. Kar).

https://doi.org/10.1016/j.matcom.2019.01.001

by using rough sets theory to obtain fuzzy logical relationship from time series and an adaptive expectation model to improve forecasting accuracy on TAIEX and National Association of Securities Dealers Automated Quotations (NASDAQ) experimental data sets. Pai and Lin [26] developed a hybrid ARIMA and support vector machine model for stock price forecasting.

Though time series data are collected in particular time interval, it seems continuous for long time period. So, discretization is helpful in time series data to find important information within predicting attributes or conditional attributes. The discretization methods are classified as endogenous (unsupervised) versus exogenous (supervised), local versus global, parametrized versus non-parametrized, and hard versus fuzzy depending on its processing nature [18]. In last classification, the hard methods discretize the intervals at the cutting point where as fuzzy methods discretize intervals by overlapping endpoints [33].

There are several work available in the literature on data discretization [11,38]. Liu and Mendel [22] developed type-2 fuzzistics to decide interval endpoint data for a word from group of subjects. They made a MATLAB M-file for this interval approach (IA) [1]. In this paper, we have tried to apply this T2-fuzzistics methodology to obtain IT2 FS models for linguistics. Stock exchange time series are analyzed and forecasted by using data discretization, rule generation and adaptive forecasting technique. Cumulative probability distribution along with interval approach in linguistics's intervals are used as data discretization technique. After deciding the range of linguistics, Gaussian fuzzy membership is assigned to each predicting attribute for which first order fuzzy logical relationships are obtained. Then, rough set theory is used to generate rules which are used for defuzzification. Subsequently, adaptive forecasting is applied on defuzzified results which is eventually evaluated on three stock market time series. Finally, theses results are compared with the work where CPDA is used for data discretization.

The remainder of this paper is organized as follows. Preliminary discussion on related topics is given in Section 2. Proposed approach is presented in Section 3. We present the descriptions about the data sets in Section 4. Results and discussions have been provided in Section 5. Finally, the conclusions and future research directions are presented in Section 6.

2. Preliminaries

This section presents some basic concepts of fuzzy set, higher order fuzzy set, and fuzzistics.

2.1. Type 1 fuzzy set

It is an extension of classical set, where the elements have varying degrees of membership. A fuzzy set [37] (or type 1 fuzzy set) allows its members/ elements to have different membership degree in [0, 1].

A fuzzy set A on a set is of the form $A = (x, y_A(x)) : x \in X$, where $y_A : X \to [0, 1]$ denote the degree of membership of element $x \in X$.

2.2. Type-2 fuzzy set

Fuzzy set is a collection of objects associated with membership function that characterizing the objects by grades from [0, 1]. Depending on the membership function higher order fuzzy sets like interval type-2, generalized type-2 fuzzy sets are defined.

General type-2 fuzzy sets are extensions of type-1 fuzzy sets where uncertainty is measured on the membership function of type-1 fuzzy sets. For interval type-2 fuzzy sets of pairs (x, y_x), where y_x is the membership of x within the range $0 \leq y_x \leq 1$ is an interval considering lower and upper membership function of y_x, which is defined as

$$\tilde{A} = \int_{x \in X} \int_{y_x \in [\underline{y}_x, \bar{y}_x]} 1/(x, y_x) \tag{1}$$

where $\mu(x, y_x) = 1$. In case of general type-2 fuzzy set, $\mu(x, y_x) = \mu_x$, where μ_x is the membership of (x, y_x) within the range $0 \leq \mu_x \leq 1$, which is defined as below.

$$\tilde{A} = \int_{x \in X} \int_{y_x \in [\underline{y}_x, \bar{y}_x]} \mu_x/(x, y_x) \tag{2}$$

y_x and μ_x are the primary and secondary memberships of x respectively.

In case of both the general and interval type-2 fuzzy sets, the membership functions are three dimensional, the only difference between them is that the secondary membership value of the interval type-2 membership function is always equal to 1.

2.3. Fuzzistics

Fuzzistics is applied in encoding the words of CWW (computing with words) engine [24] in which the objects of computation are words and propositions collected from a natural language. Here, fuzzistics is used to fulfill another aspect to decide the linguistic intervals. In case of stock price, numerical data are collected over some period, the time series seems continuous. Data discretization is helpful in such cases which is achieved here by assigning linguistic values to the time series. The methodology by assigning linguistic values can make robust forecasting when the historical data are incorrect which is another advantage [8]. Liu and Mendel [22] proposed type-2 fuzzistics methodology to obtain interval type-2 fuzzy sets for words collected from a group of subject. An interval approach (IA) is used to determine the footprint of uncertainty (FOU) of each word in the group of subject. The data having inherent uncertainties are described satisfactorily as intervals. IA consists of two parts - (1) the data part : the interval endpoint data are preprocessed, after that data statistics are calculated for the processed intervals, (2) the fuzzy set (FS) part : the data are used to decide the foot print of uncertainty (FOU) of the word i.e., whether it would be a left-shoulder, an interior or a right-shoulder.

2.3.1. Data part

From a given set of information, suppose, a set of intervals $[a^{(i)}, b^{(i)}],\ i = 1, \ldots, n$ are collected and following two steps have been used to process the data.

1. **Data preprocessing :**
   In this step, there are four stages : (1) bad data processing, (2) outliers processing, (3) tolerance-limit processing, and (4) reasonable-interval processing. After processing, some intervals are discarded among n intervals. Suppose, there are m number of interval endpoints for each word.
2. **Computation of data statistics for each preprocessed intervals :**
   A probability distribution is assigned to each of the m surviving data intervals. Statistics are computed for each interval using the probability and the interval endpoints. Data statistics S_i for each surviving interval is given below :

$$S_i = (m_Y^{(i)}, \sigma_Y^{(i)}), \quad i = 1, \ldots, m \tag{3}$$

   where $m_Y^{(i)} = \dfrac{a^{(i)} + b^{(i)}}{2}$ is the mean and $\sigma_Y^{(i)} = \dfrac{b^{(i)} - a^{(i)}}{\sqrt{12}}$ is the standard deviation. These data statistics are used in FS part of the IA to get the parameters of type-1 membership function (T1 MF).

2.3.2. FS part

Following nine steps are used in FS part.

1. **Selection of a T1 FS model :**
   Here, a symmetrical triangle interior T1 MF, a left-shoulder T1 MF, or a right-shoulder T1 MF [22] are used. Consider the mean and variance of the assumed uniform probability distribution to map the interval of data to a T1 MF.
2. **Establish FS uncertainty measures :**
   Consider the probability distribution of x as

$$f(x) = \frac{\mu_A(x)}{\int_{a_{MF}}^{b_{MF}} \mu_A(x)dx} \tag{4}$$

   where $x \in [a_{MF}, b_{MF}]$, a_{MF} and b_{MF} denote the left-end and right-end of the T1 MF. In case of shoulder MFs, a_{MF} and b_{MF} do not cover the entire span of MF, where the uncertainty lies.
3. **Compute uncertainty measures for T1 FS models :**
   The mean and standard deviations for three possible structures namely interior triangle, left-shoulder, and right-shoulder T1 MF are computed.

4. **Compute general formulas for parameters of T1 FS models :**
   The parameters $a_{MF}^{(i)}$ and $b_{MF}^{(i)}$, $i = 1, 2, \ldots, m$ of three T1 FSs are computed by equating the mean and standard deviation of itself to the same of data interval respectively for each of the m remaining data intervals.
5. **Establish the nature of FOU :**
   This is FOU classification problem to decide which of the FOUs (interior, left-shoulder or right-shoulder) will be selected.
6. **Compute embedded T1 FSs :**
   Once the type of FOU has been decided for specific word, each of the word's remaining m data intervals are mapped into their respective T1 FSs. These T1 FSs are called embedded T1 FSs.
7. **Delete inadmissible T1 FSs :**
   It is possible that some of the m embedded T1 FSs are inadmissible, i.e., which are less than 0 or greater than 10, those are deleted.
8. **Compute an IT2 FS using the union :**
   IT2 FS of a word, denoted by $\tilde{A}$, is computed by taking the union of all admissible embedded T1 FSs.
9. **Compute mathematical model for FOU($\tilde{A}$) :**
   To compute mathematical model for FOU($\tilde{A}$), UMF($\tilde{A}$) and LMF($\tilde{A}$) are approximated in such a way that all the admissible embedded T1 FSs consist of FOU($\tilde{A}$).

2.4. Cumulative probability distribution approach (CPDA) for data discretization

CPDA is an important approach to discretize the data set in several hybrid forecasting models [9,32,35]. It helps in partitioning the universe of discourse (UOD) with unequal intervals. It calculates the inverse of the normal cumulative distribution function based on two parameters mean (μ) and standard deviation (σ) of the data set for a given probability p, where $p \in [0, 1]$. Lower and upper bound cumulative probabilities for each linguistic value are defined as follows:

$$p_{LB}^{1} = 0, \; p_{UB}^{n} = 1, \tag{5}$$

$$p_{LB}^{i} = \frac{(2i-3)}{2n} \quad (2 \le i \le n), \tag{6}$$

$$p_{UB}^{j} = \frac{j}{n} \qquad (1 \le j \le (n-1)), \tag{7}$$

where n denotes the number of linguistic values and i, j are the given order of linguistic value.

After getting the lower and upper bound cumulative probabilities for each linguistic value, corresponding endpoints for each linguistic value are calculated by finding inverse of the normal cumulative distribution function which are given below:

$$x_{LB}^{1} = \text{minimum of UOD} - \sigma, \tag{8}$$

$$x_{LB}^{i} = F^{-1}(p_{LB}^{i}|\mu, \sigma) = \{x : F(x|\mu, \sigma) = p_{LB}^{i}\} \tag{9}$$
$$\text{where } (2 \le i \le n),$$

$$x_{UB}^{j} = F^{-1}(p_{UB}^{j}|\mu, \sigma) = \{x : F(x|\mu, \sigma) = p_{UB}^{j}\} \tag{10}$$
$$\text{where } (1 \le j \le (n-1)),$$

$$x_{UB}^{n} = \text{maximum of UOD} + \sigma, \tag{11}$$

2.5. Rough set theory

Rough set theory (RST) was proposed by Pawlak [29]. It is motivated by practical needs to interpret, characterize, represent, and process in-discernibility of individuals. Any set of all indiscernible objects is called an elementary set and forms a basic granule of knowledge about the universe. Any union of elementary sets is referred to as a precise set; otherwise the set is rough. For example, if a group of patients is described by using several symptoms, many patients would share the same symptoms, and hence are indistinguishable. This forces to think a subset of the patients as one unit, instead of many individuals. Rough set theory provides a systematic method for representing and processing

vague concepts caused by in-discernibility in situations with incomplete information or a lack of knowledge. At least two views can be used to interpret this theory, operator-oriented view and set-oriented view.

Definition of Rough Set: Let X be the universe of discourse of non-empty finite set of objects and A be the non-empty finite set of attributes, then $S = (X, A)$ is said to be an information system.

For an information system $S = (X, A)$, and $B \subseteq A$, $T \subseteq X$ can be approximated based on the information contained in B into B-lower and B-upper approximation of T which are represented as $\underline{B}(T)$ and $\bar{B}(T)$, which are defined as $\underline{B}(T) = \{x|[x]_B \subseteq T\}$ and $\bar{B}(T) = \{x|[x]_B \cap T \neq 0\}$ respectively. The lower approximation consists of all objects that definitely belong to the set, and the upper approximation contains all objects that possibly belong to the set.

The difference between the upper and the lower approximation describes the boundary region, $BN_B(T) = \bar{B}(T) - \underline{B}(T)$, of the rough sets. If the boundary region is non-empty, then the set T is called "rough" with respect to knowledge in B, otherwise T is called a precise set.

Rough Set Exploration System (RSES) [2]: It is a tool set for analyzing data with the use of methods coming from Rough Set Theory. Learning from Examples Module, version 2 (LEM2) [2] algorithm is used to reduce the rule sets, known as rule induction algorithm. Rule induction is one of the most important techniques of machine learning. Since regularities hidden in data are frequently expressed in terms of rules, rule induction is one of the fundamental tools of data mining. Usually rules are the expressions of the form

$$\begin{aligned} &if(attr-1; value-1)\ and\ (attr-2; value-2)\ and\ ... \\ &and\ (attr-n; value-n)\ then\ (decision; value). \end{aligned} \tag{12}$$

There are two attributes attr0 and attr1 are used in each rule of the rule base. The rule (attr0 = 7) => (attr1 = 6) means if condition attribute attr0 is A_7 then decision attribute attr1 is A_6.

3. Proposed method

This section describes the general architecture of the proposed hybrid model based on fuzzistics and RST. Liu and Mendel [22] developed an algorithm to encode words into interval type-2 fuzzy sets using an interval approach. Based on a ten point scale, the end-points of an interval have been searched. Here, in case of time series data points, which are basically numerical values in fixed time intervals, we need to develop some methods to discretize time series data points in linguistic values. In this proposed method, we have applied type-2 fuzzistics methodology [22] with some modifications. Fig. 1 shows the flow diagram of the overall architecture of the model. The steps of the proposed approach are as follows.

Step 1: Find maximum, minimum, mean, and standard deviation of the time series data points.

Step 2: Considering 600 training data points and 100 testing data points in each time series. 600 data points are grouped into 30 groups consisting 20 data points each are described in following ways:

1. First 20 data points in first group, next 20 data points in second group, and so on.
2. Taking the $(an + 1)$th data points in first group, taking the $(an + 2)$th data points in second group, taking the $(an + 3)$th data points in third group, where $a = 21$ and $n = 1, 2, \ldots, 20$ respectively.
3. Taking the $(an + 1)$th data points in first group, $(an + 2)$th data points in second group, $(an + 3)$th data points in third group where $a = 21$ and $n = 1, 2, \ldots, 20$, respectively up to 10 groups, next group is started with $(20 * 10 + 1)$th data point up to consecutive 19 more data points, next group is started with $(20 * 11 + 1)$th data point up to consecutive 19 more data points, next group is started with $(20 * 12 + 1)$th data point up to consecutive 19 more data points, respectively.

Here, we have used the code developed by Liu and Mendel [22] to get endpoints for the linguistic values.

Step 3: Find maximum, minimum, mean, and standard deviation of each of the 30 groups. Then, CPDA is applied on each group to get the lower bound and upper bound for each linguistic value. Those lower and upper bounds are converted on a 10 point scale.

Step 4: Next, matlab code [22] is used to get the lower, mean, and upper centroid resulted due to embedded type-1 fuzzy sets which are mapped into interval type-2 fuzzy set. The code used here, is not able to give lower and upper centroid values all the time, it also gives single centroid value due to filled in situation occurred [22]. Lower and upper centroid values give the endpoints of each linguistics's interval. To remove the single centroid value, we have

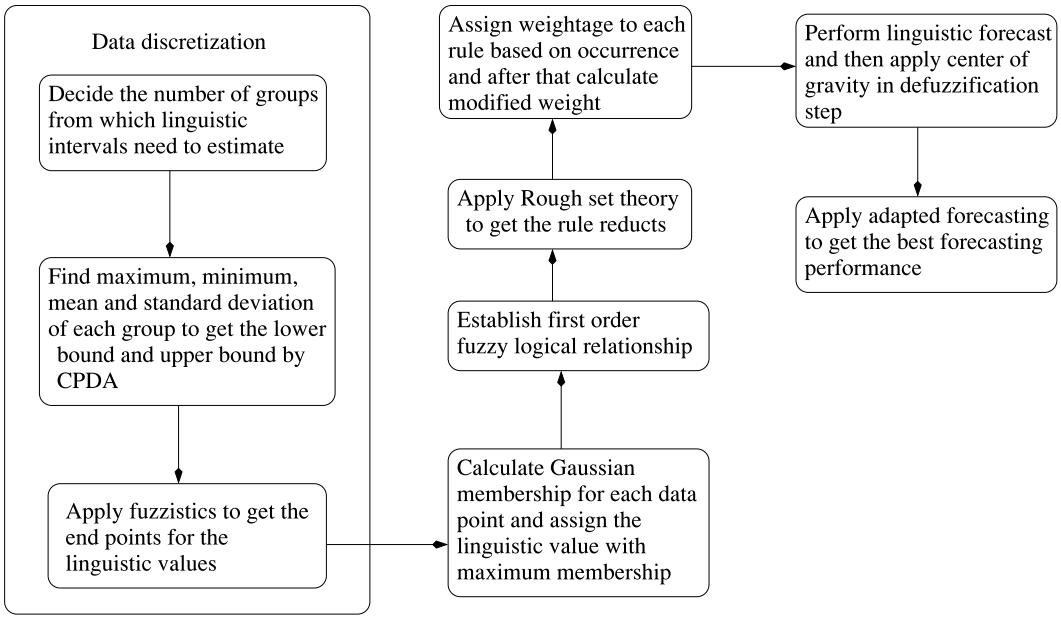

Fig. 1. Flowchart of the overall architecture of the proposed approach.

Table 1
Intervals (LB, UB), where LB: Lower bound and UB: Upper bound.

	Fuzzistics		CPDA	
	LB	UB	LB	UB
1	6869.0	7670.0	4812.1	11064.0
2	9395.0	11006.0	9779.3	12658.0
3	11806.0	12102.0	11946.0	13849.0
4	13238.0	13467.0	13279.0	14916.0
5	14527.0	14704.0	14389.0	15983.0
6	15827.0	15992.0	15443.0	17174.0
7	15997.0	18398.0	16553.0	18768.0
8	20485.0	21920.0	17886.0	24221.3

used the average of just previous centroid mean and single centroid mean as lower endpoint of itself. Similarly, we have used the average of single centroid mean and just following centroid mean to get the upper endpoint of itself. In some cases, we have got the endpoints of interval, which are non-overlapping. We have used those as it has no single centroid mean.

After computing the algorithm, we get the intervals on 10 point scale. Then, those intervals are transformed back to the original scale. Intervals (Lower bound, Upper bound) are shown in Table 1 for both the fuzzistics and CPDA.

Step 5: After getting the intervals for linguistic values, Gaussian membership function is calculated for each data point with respect to each linguistic value. Linguistic value with maximum membership is assigned to each data point.

Step 6: First order fuzzy logical relationship (FLR) is generated using assigned linguistic of current data point from assigned linguistic of just previous data point, i.e., $f(t-1) \rightarrow f(t)$, where $f(t-1)$ and $f(t)$ are the condition attribute and decision attribute respectively. All $f(t)$s are represented by assigned linguistics.

Step 7: Rough set theory (RST) is used to generate the FLR in group (FLG) and analyze the decision table generated from first order FLR. Rule reducts are shown in Table 2 for both the fuzzistics and CPDA. In Table 2, first row shows

$(attr0 = 7) \Rightarrow (attr1 = 7[313], 6[9], 5[17], 4[7])346,$

which means $A_7 \rightarrow A_7, A_6, A_5, A_4$. There are 346 rules from condition attribute A_7 to decision attributes A_7, A_6, A_5, A_4 of which $A_7 \rightarrow A_7$ occurs 313 times, $A_7 \rightarrow A_6$ occurs 9 times, $A_7 \rightarrow A_5$ occurs 17 times and

Table 2
Rule reducts.

RULES 7 (Fuzzistics)
(attr0 = 7) => (attr1 = 7[313], 6[9], 5[17], 4[7]) 346
(attr0 = 2) => (attr1 = 2[127], 1[2]) 129
(attr0 = 8) => (attr1 = 8[37]) 37
(attr0 = 5) => (attr1 = 7[17], 5[10]) 27
(attr0 = 4) => (attr1 = 4[13]) 13
(attr0 = 6) => (attr1 = 6[3]) 3
(attr0 = 3) => (attr1 = 3[10]) 10

RULES 8 (CPDA)
(attr0 = 6) => (attr1 = 6[86], 7[13], 5[11]) 110
(attr0 = 7) => (attr1 = 6[13], 7[85]) 98
(attr0 = 1) => (attr1 = 2[5], 1[96]) 101
(attr0 = 5) => (attr1 = 5[64], 4[10]) 74
(attr0 = 8) => (attr1 = 8[70]) 70
(attr0 = 4) => (attr1 = 5[10], 4[46], 3[6]) 62
(attr0 = 2) => (attr1 = 2[35]) 35
(attr0 = 3) => (attr1 = 3[16]) 16

Table 3
Number of rules (Fuzzistics)

Rules	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
A_1	0	0	0	0	0	0	0	0
A_2	2	127	0	0	0	0	0	0
A_3	0	0	10	0	0	0	0	0
A_4	0	0	0	13	0	0	0	0
A_5	0	0	0	0	10	0	17	0
A_6	0	0	0	0	0	3	0	0
A_7	0	0	0	7	17	9	313	0
A_8	0	0	0	0	0	0	0	37

$A_7 \rightarrow A_4$ occurs 7 times. For this rough set analysis, LEM2 algorithm [2] is used. Tables 3 and 4 show the number of rules from each linguistic to itself and all other linguistic for both fuzzistics and CPDA. Summation of occurrence of each rule is used as weights in following Step. That means, the rule $A_7 \rightarrow A_7$ occurs 313 times, so the weights of this rule is $\{313(313+1)/2\}$. Calculated weights corresponding to each rule are given in Tables 5 and 6 for fuzzistics and CPDA respectively.

Step 8: Modified weight on fuzzy time series is used based on the various recurrences of FLR. Computation of modified weights are described below. Suppose $A_i \rightarrow A_j, A_k, \ldots, A_p$ is a FLG where the weights are specified as $j = g_1, k = g_2, \ldots, p = g_n$, n is the number of attributes on right hand side of FLG. Then

$$\begin{aligned} W(t) &= [w_1\ w_2\ \ldots\ w_n] \\ &= \left[\frac{j}{(j+k+\cdots+p)}\frac{k}{(j+k+\cdots+p)}\cdots\frac{p}{(j+k+\cdots+p)}\right] \\ &= \left[\frac{g_1}{\sum_{i=1}^{n} g_i}\frac{g_2}{\sum_{i=1}^{n} g_i}\cdots\frac{g_n}{\sum_{i=1}^{n} g_i}\right] \end{aligned} \tag{13}$$

Modified weights corresponding to each rule are given in Tables 7 and 8 for fuzzistics and CPDA respectively.

Step 9: In this step, linguistic forecast is performed based on the rules and calculated weights. The steps are given below.

1. If the linguistic value of $(t-1)$th data point is A_i and there is only one rule from it, say $A_i \rightarrow A_j$ with weight W_j, then linguistic forecast for tth data point is A_j with weight W_j.

Table 4
Number of rules (CPDA)

Rules	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
A_1	96	5	0	0	0	0	0	0
A_2	0	35	0	0	0	0	0	0
A_3	0	0	16	0	0	0	0	0
A_4	0	0	6	46	10	0	0	0
A_5	0	0	0	10	64	0	0	0
A_6	0	0	0	0	11	86	13	0
A_7	0	0	0	0	0	13	85	0
A_8	0	0	0	0	0	0	0	70

2. If the linguistic value of $(t-1)$th data point is A_i and there are several rules from it, say $A_i \rightarrow A_j, A_k, A_m$ with weights W_j, W_k, W_m, then linguistic forecast for tth data point are A_j, A_k, A_m with weight W_j, W_k, W_m respectively.
3. If there is no matched FLR with $(t-1)$th data point, then $(t-1)$th data point is used to forecast the tth data point.

Step 10: In this step, first the linguistic forecasts are defuzzified by using center of gravity (COG) method and then adapted forecast has been performed as the movement of current stock price data highly depends on just previous or some of just previous stock price data.

The defuzzified linguistic forecasts from **Step 9** is obtained by:

$$\text{forecast}(t) = \sum_{i=1}^{n} \text{COG}(i) \times W_i, \tag{14}$$

where $\text{COG}(A_i)$ is the mean of the linguistic region A_i and W_i corresponding weight of A_i.

To get the adapted forecast, smoothing factor is used on just previous data point and current forecasted value. The equation is given below.

$$F(t) = f(t-1) + \alpha \times (\text{forecast}(t) - f(t-1)) \tag{15}$$

where $F(t)$ is the predicted value at tth time point, $f(t-1)$ represents closing price value at $(t-1)$th time point, α is the smoothing factor and forecast(t) is the defuzzified value of tth data point.

Step 11: For testing purpose, each data point of testing data sets is assigned with their linguistic value according to their maximum membership. Previously calculated normalized weight W_i is used for the first rule of the testing data sets and it is updated for the following rules as weights w_is are increased as the number of rules are increased. After getting the weights for first rule, current forecast and adapted forecast values are calculated and similar steps have been used.

4. Data description

The proposed methodology is applied to forecast three stock market time series data namely Bombay stock exchange (BSE) [3] sensex, New York stock exchange (NYSE) [4] composite index, and Taiwan stock exchange capitalization weighted stock index (TAIEX) [5].

1. BSE popular equity index – the BSE sensex – is India's mostly tracked stock market benchmark index. It has the largest number of companies listed and traded among all the exchanges in India. There are several work i.e., stock market analysis, forecasting models on BSE data [13,17].
2. NYSE is American stock exchange, hugely used in several works [14,20].
3. TAIEX is Taiwan stock exchange benchmark index. It is also hugely used in several works [7,35]

Total 1800 data points are collected from first day of September, 2007 to last day of December, 2014 from each of the historical data of three stock exchanges closing price index. 1800 data points are grouped into three time series namely BSE1, BSE2, BSE3 of BSE each of 600 data points. Similarly we follow for NYSE and TAIEX time series. So, there are total nine small time series, three from each BSE, NYSE, and TAIEX form as training data sets.

Table 5
Weights (Fuzzistics)

Weights	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	$\sum_{i=1}^{8} w_i$
w_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
w_2	3.00	8128.00	0.00	0.00	0.00	0.00	0.00	0.00	8131.00
w_3	0.00	0.00	55.00	0.00	0.00	0.00	0.00	0.00	55.00
w_4	0.00	0.00	0.00	91.00	0.00	0.00	0.00	0.00	91.00
w_5	0.00	0.00	0.00	0.00	55.00	0.00	153.00	0.00	208.00
w_6	0.00	0.00	0.00	0.00	0.00	6.00	0.00	0.00	6.00
w_7	0.00	0.00	0.00	28.00	153.00	45.00	49141.00	0.00	49367.00
w_8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	703.00	703.00

Table 6
Weights (CPDA)

Weights	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	$\sum_{i=1}^{8} w_i$
w_1	4656.00	15.00	0.00	0.00	0.00	0.00	0.00	0.00	4671.00
w_2	0.00	630.00	0.00	0.00	0.00	0.00	0.00	0.00	630.00
w_3	0.00	0.00	136.00	0.00	0.00	0.00	0.00	0.00	136.00
w_4	0.00	0.00	21.00	1081.00	55.00	0.00	0.00	0.00	1157.00
w_5	0.00	0.00	0.00	55.00	2080.00	0.00	0.00	0.00	2135.00
w_6	0.00	0.00	0.00	0.00	66.00	3741.00	91.00	0.00	3898.00
w_7	0.00	0.00	0.00	0.00	0.00	91.00	3655.00	0.00	3746.00
w_8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2485.00	2485.00

Table 7
$\mathrm{W_i} = \{w_i / \sum_{i=1}^{8} w_i\}$ (Fuzzistics)

W_i	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8
W_1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
W_2	0.000369	0.999631	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
W_3	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
W_4	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000
W_5	0.000000	0.000000	0.000000	0.000000	0.264423	0.000000	0.735577	0.000000
W_6	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000
W_7	0.000000	0.000000	0.000000	0.000567	0.003099	0.000912	0.995422	0.000000
W_8	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000

For testing data sets, where the BSE1 time series is ended, next 100 points are collected which forms the BSE1 testing data sets. Similarly, where the BSE2 and BSE3 time series are ended, next 100 points are collected which forms the BSE2 and BSE3 testing data sets respectively. Similarly NYSE1, NYSE2, NYSE3, TAIEX1, TAIEX2, and TAIEX3 testing data sets are collected.

5. Result discussion

In this section, we demonstrate the effectiveness of the proposed methodology on three stock exchange closing price data. The performance of the proposed methodology is compared with the following existing works.

1. Chen [8] proposed a fuzzy time series model by using simplified arithmetic operations to calculate forecasted result which performed efficiently comparison to complicated max–min composition of the conventional work proposed by Song and Chissom [30,31].
2. Lee et al. [19] developed a fuzzy time series model by assigning weights using chronological number to FLR belongs to the group of FLR. Also, the difference of actual data and mid point of the interval has been used to obtain the estimated value. The method is successfully applied on enrollment data of University of Alabama and University Teknologi Malaysia.

Table 8
$W_i = \{w_i / \sum_{i=1}^{8} w_i\}$(CPDA)

W_i	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8
W_1	0.996789	0.003211	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
W_2	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
W_3	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
W_4	0.000000	0.000000	0.018150	0.934313	0.047537	0.000000	0.000000	0.000000
W_5	0.000000	0.000000	0.000000	0.025761	0.974239	0.000000	0.000000	0.000000
W_6	0.000000	0.000000	0.000000	0.000000	0.016932	0.959723	0.023345	0.000000
W_7	0.000000	0.000000	0.000000	0.000000	0.000000	0.024293	0.975707	0.000000
W_8	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000

Table 9
Root mean square error of the four methods on three time series of BSE, NYSE, and TAIEX closing price.

Data sets	Chen [8]	Lee et al. [19]	Fuzzistics	CPDA [35]
BSE1	274.54	226.77	**176.54**	176.62
BSE2	200.68	**136.04**	136.25	136.57
BSE3	311.41	**270.82**	282.57	282.54
NYSE1	160.55	**103.90**	107.85	107.98
NYSE2	103.08	66.85	62.17	**62.16**
NYSE3	127.82	**81.27**	83.77	83.72
TAIEX1	95.42	**86.28**	87.20	87.21
TAIEX2	85.96	62.57	56.90	**56.87**
TAIEX3	89.14	72.94	64.97	**64.04**

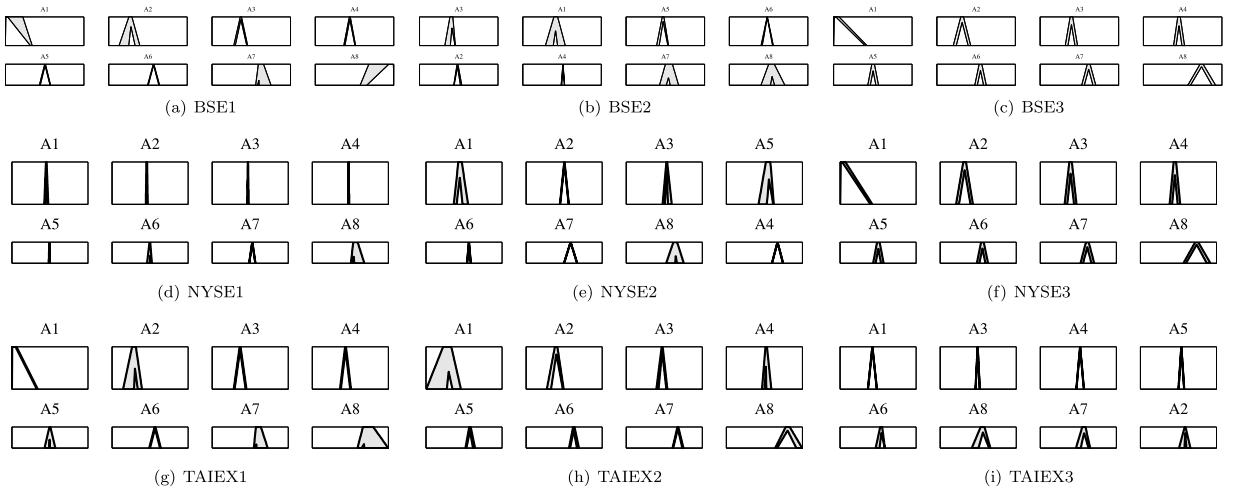

Fig. 2. FOUs for 8 linguistics for each data set.

3. Teoh et al. [35] used cumulative probabilistic distribution approach (CPDA) to decide the linguistic intervals and then establish first order fuzzy relation within consecutive data point. Rough set theory thereby was used for rule reduction and finally adapted forecast was used on defuzzified data.

Proposed methodology decides the linguistic intervals by using T2-fuzzistics methodology. In T2-fuzzistics methodology, we need a set of intervals for each linguistic value to obtain IT2 FS models. We have used CPDA to determine the set of intervals. Then, first order fuzzy relation is established within consecutive data points, rough set based rule reduction is utilized and finally, adapted forecast is used on defuzzified data.

Fig. 3. Performance of proposed approach together with other methods like Chen [8], Lee et al. [19], and Teoh et al. [35].

Fig. 2(a) – Fig. 2(i) show the FOUs of 8 linguistic values of each BSE1, BSE2, BSE3, NYSE1, NYSE2, NYSE3, TAIEX1, TAIEX2, and TAIEX3 respectively by using the method proposed by Liu and Mendel [22]. Fig. 3(a) – Fig. 3(f) show the actual and forecasted values of the proposed methods with the work by Chen [8], Lee et al. [19], and Teoh et al. [35]. It has been observed that the performance of the proposed method is better than the model proposed by Chen [8] and more or less similar to the result obtained by the work proposed by Lee et al. [19]

and Teoh et al. [35]. Table 9 shows the root mean square error (RMSE) of the actual and forecasted results for nine testing data sets using different time series methods.

6. Conclusion

In this paper, type-2 fuzzistics methodology is used as a data discretization process to obtain interval type-2 fuzzy sets for linguistic values. The fuzzistics is used as a data discretization process to get the intervals for each linguistic value. In the type-2 fuzzistics methodology, Liu and Mendel [22] collected intervals for each word from different sources. Here, we organize the time series data and applied the CPDA method for each linguistic from single source to generate the intervals. After deciding the intervals for each linguistic value, fuzzy sets and first order fuzzy logical relationships (FLR) are established. Rough set theory is applied on the originated FLR to generate reduced rule base. Then, normalized weights are assigned to each existing rule and COG defuzzification has been performed based on the assigned weights. We have applied adapted forecasting and then RMSE is calculated to check the accuracy of the forecasted time series with the actual time series.

We have compared the proposed method with the other existing methods. It is observed that proposed method performs better compared to the method by Chen [8] and more or less similar to the other two existing works [19,35].

In future, one can find other methodologies to collect the intervals for linguistics from single source to overcome the difficulties which we have faced for non-overlapping intervals and single centroid mean.

Acknowledgment

The first author would like to thank DST INSPIRE, India for their help and supports to sustain the work.

References

[1] http://sipi.usc.edu/mendel.
[2] <http://alfa.mimuw.edu.pl/rses/>.
[3] BSE data set, http://in.finance.yahoo.com/q/hp?s=^BSESN.
[4] NYSE data set, http://finance.yahoo.com/q/hp?s=^NYA+Historical+Prices.
[5] TAIEX data set, http://finance.yahoo.com/q/hp?s=^TWII+Historical+Prices.
[6] T. Bollerslev, Generalized autoregressive conditional heteroscedasticity, J. Econometrics 31 (1986) 307–327.
[7] Q. Caia, D. Zhanga, B. Wua, S.C.H. Leung, A novel stock forecasting model based on fuzzy time series and genetic algorithm, Procedia Comput. Sci. 18 (2013) 1155–1162.
[8] S.-M. Chen, Forecasting enrollments based on fuzzy time series, Fuzzy Sets and Systems 81 (1996) 311–319.
[9] J.S. Chen, Extracting classification rules based on a cumulative probability distribution approach, J. Zhejiang Univ., Sci. C (Comput. Electron.) 12 (2011) 379–386.
[10] C.H. Cheng, T.L. Chen, L.Y. Wei, A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting, Inform. Sci. 180 (2010) 1610–1629.
[11] M.R. Chmielewski, J.W. Grzymala-Busse, Global discretization of continuous attributes as preprocessing for machine learning, Int. J. Approx. Reason. 15 (1996) 319–331.
[12] R.F. Engle, Autoregressive conditional heteroscedasticity with estimator of the variance of United Kingdom inflation, Econometrica 50 (1982) 987–1008.
[13] S.S. Gangwar, S. Kumar, Probabilistic and intuitionistic fuzzy sets–based method for fuzzy time series forecasting, Cybern. Syst., Int. J. 45 (2014) 349–361.
[14] C.W.J. Granger, Forecasting stock market prices: Lessons for forecasters, Int. J. Forecast. 8 (1992) 3–13.
[15] K. Huarng, H.-K. Yu, A type 2 fuzzy time series model for stock index forecasting, Physica A 353 (2005) 445–462.
[16] K. Hurang, H.K. Yu, The application of neural networks to forecast fuzzy time series, Physica A 363 (2006) 481–491.
[17] B.P. Joshi, S. Kumar, Intuitionistic fuzzy sets based method for fuzzy time series forecasting, Cybern. Syst., Int. J. 43 (2012) 34–47.
[18] K. Kim, I. Han, Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index, Expert Syst. Appl. 19 (2000) 125–132.
[19] M.H. Lee, R. Efendi, Z. Ismail, Modified weighted for enrollment forecasting based on fuzzy time series, Matematika 25 (2009) 67–78.
[20] W. Leigh, R. Purvis, J.M. Ragusa, Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: A case study in romantic decision support, Decis. Support Syst. 32 (2002) 361–377.
[21] Y. Lin, Y. Yang, Stock markets forecasting based on fuzzy time series model, in: IEEE International Conference Intelligent Computing and Intelligent Systems, Nov. 2009, pp. 782-786..
[22] F. Liu, J.M. Mendel, Encoding words into interval type-2 fuzzy sets using an interval approach, IEEE Trans. Fuzzy Syst. 16 (2008) 1503–1521.
[23] D. Massacci, Predicting the distribution of stock returns: Model formulation, statistical evaluation, VaR analysis, and economic significance, J. Forecast. 34 (2015) 191–208.
[24] J.M. Mendel, Computing with words and its relationships with fuzzistics, Inform. Sci. 177 (2007) 988–1006.

[25] C. Nikolopoulos, P. Fellrath, A hybrid expert system for investment advising, Expert Syst. 11 (1994) 245–250.
[26] P.-F. Pai, C.-S. Lin, A hybrid ARIMA and support vector machines model in stock price forecasting, Omega 33 (2005) 497–505.
[27] S.S. Pal, S. Kar, Time series forecasting using fuzzy transformation and neural network with back propagation learning, J. Intell. Fuzzy Systems 33 (1) (2017) 467–477.
[28] S.S. Pal, T. Pal, S. Kar, An improvement in forecasting interval based fuzzy time series, in: Fuzzy Systems (FUZZ-IEEE) International Conference, 2014, pp. 1390-1394..
[29] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (1982) 341–356.
[30] Q. Song, B.S. Chissom, Forecasting enrollments with fuzzy time series, Part I, Fuzzy Sets and Systems 54 (1993) 1–9.
[31] Q. Song, B.S. Chissom, Forecasting enrollments with fuzzy time series, Part II, Fuzzy Sets and Systems 62 (1994) 1–8.
[32] C.H. Su, T.L. Chen, C.H. Cheng, Y.C. Chen, Forecasting the stock market with linguistic rules generated from the minimize entropy principle and the cumulative probability distribution approaches, Entropy 12 (2010) 2397–2417.
[33] R. Susmaga, Analyzing discretizations of continuous attributes given a monotonic discrimination function, Intell. Data Anal., Int. J. 1 (1997) 157–179.
[34] G.G. Szpiro, Forecasting chaotic time series with genetic algorithms, Phys. Rev. E 55 (3) (1997) 2557–2568.
[35] H.J. Teoh, C.H. Cheng, H.H. Chu, J.S. Chen, Fuzzy time series model based on probabilistic approach and rough set rule induction for empirical research in stock markets, Data Knowl. Eng. 67 (2008) 103–117.
[36] J.-Z. Wang, J.-J. Wang, Z.-G. Zhang, S.-P. Guo, Forecasting stock indices with back propagation neural network, Expert Syst. Appl. 38 (2011) 14346–14355.
[37] L.A. Zadeh, Fuzzy sets, Inf. Control 8 (1965) 338–353.
[38] D.A. Zighed, S. Rabaséda, R. Rakotomalala, FUSINTER: A method fot discretization of continuous attributes, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 6 (1998) 307–326.