2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)
July 6-11, 2014, Beijing, China

Fuzzy Rule-Based Ensemble for Time Series Prediction: The Application of Linguistic Associations Mining

Martin Štěpnička, Lenka Štěpničková, and Michal Burda

***Abstract*— As there are many various methods for time series prediction developed but none of them generally outperforms all the others, there always exists a danger of choosing a method that is inappropriate for a given time series. To overcome such a problem, distinct ensemble techniques, that combine more individual forecasts, are being proposed.**

In this contribution, we employ the so called fuzzy rule-based ensemble. This method is constructed as a linear combination of a small number of forecasting methods where the weights of the combination are determined by fuzzy rule bases based on time series features such as trend, seasonality, or stationarity. For identification of fuzzy rule base, we use linguistic association mining. An exhaustive experimental justification is provided.

I. Introduction

TIME SERIES PREDICTION is an important tool for support of individual and organizational decision making. It has a wide practical use in economy, industry, demography, and other areas of application. The time series is usually given as a finite sequence $y_1, y_2, \ldots, y_T$ of real numbers and the task is to predict future values $y_{T+1}, y_{T+2}, \ldots, y_{T+h}$ where h denotes so called *forecasting horizon*.

There are many different methods for this task that are nowadays widely used in practice, let us recall e.g. well-known Box-Jenkins methodology [1] which consists of autoregressive and moving average models, techniques based on a decomposition of a given time series into trend, seasonal and cyclic components, or exponential smoothing methods.

Further, a notable number of works aiming at fuzzy approach to time series analysis and prediction has been published. For instance, a study presenting Takagi-Sugeno rules [2] in the view of the Box-Jenkins methodology [3] or the works dealing with the linguistic approach [4], [5] have been published. Analogously, various neuro-fuzzy approaches, which lie on the border between neural networks, Takagi-Sugeno models, and evolving fuzzy systems, are very often successfully used [6], [7].

Unfortunately, there is no single forecasting method that generally outperforms any other. Thus, there is a danger of choosing a method which is inappropriate for a given time series. Note that even searching for methods, that outperform any other for narrower specific subsets of time series, has not been successful yet, see e.g. [8], where the authors stated: "Although forecasting expertise can be found in the literature, these sources often fail to adequately describe conditions under which a method is expected to be successful".

Michal Burda, Martin Štěpnička and Lenka Štěpničková are with the Institute for Research and Applications of Fuzzy Modeling, CE IT4Innovations, University of Ostrava, Ostrava, Czech Republic (email: {Michal.Burda, Martin.Stepnicka, Lenka.Stepnickova}@osu.cz).

This work was supported by the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070). Furthermore, we acknowledge the partial support of project KONTAKT II – LH12229 of MŠMT ČR.

A. Ensembles

In order to eliminate the risk of choosing an inappropriate method, distinct *ensemble techniques* (*ensembles* in short) have been designed and successfully applied. The main idea of ensembles consists in an appropriate combination of more forecasting methods. Typically, an ensemble technique is constructed as a linear combination of individual ones. It can be described as follows. Let us assume that we are given a set of M individual methods and let for a given times series $y_1, y_2, \ldots, y_T$ and a given forecasting horizon h, the j-th individual method provides us with the following prediction:

$$\hat{y}^{(j)}_{T+1}, \hat{y}^{(j)}_{T+2}, \ldots, \hat{y}^{(j)}_{T+h}, \quad j = 1, \ldots, M.$$

Then the ensemble forecast is given by the following formula:

$$\hat{y}_{T+i} = \frac{1}{\sum_{j=1}^{M} w_j} \cdot \sum_{j=1}^{M} w_j \cdot \hat{y}^{(j)}_{T+i}, \quad i = 1, \ldots, h,$$

where $w_j \in \mathbb{R}$ is a weight of the j-th individual method. These weights are usually normalized, that is, $\sum_{j=1}^{M} w_j = 1$.

Let us recall that it was perhaps Bates and Granger [9] who firstly showed significant gains in accuracy through combinations. Another early work by Newbold and Granger [10] combined various time series forecasts and compared the combination against the performance of the individual methods. They showed that for set of forecasts, a linear combination of these forecasts achieved a forecast error variance smaller than the individual forecasts. They found that the better combining procedures did produce an overall forecast superior to individual forecasts on the majority of tested time series.

How to combine methods, i.e. how to determine appropriate weights, is still a relatively open question. For instance, Makridakis et al. [11] show that taking a simple average outperforms taking a weighted average method combination. In other words, the so called "equal-weights combining" [12], that is an arithmetic mean, is a benchmark that is hard to beat and finding appropriate non-equal weights rather leads to a random damage of the main averaging idea that is behind the robustness and accuracy improvements.

B. Motivation for the Suggested Approach

Although the equal-weights ensemble performs as accurately as mentioned above, there are works that promisingly

978-1-4799-2072-3/14/$31.00 ©2014 IEEE

show the potential of more sophisticated approaches. We recall Lemke and Gabrys [13] that described an approach using meta-learning for time series forecasting based on the features of time series such as: measure for the strength of the trend, standard deviation, skewness, etc. Given time series were clustered using the k-means algorithm. Individual methods were ranked according to their performance on each cluster and then three best methods for each cluster were selected. For a given new time series, the closest cluster was determined and the given three best methods were combined.

It should be stressed that this approach performed very well on sufficiently large set of time series. For us, this approach is one of the main motivations because it demonstrates that there exists a dependence between time series features and a performance of a forecasting method.

The second major motivation stems from the so called *Rule-Based Forecasting* (RBF) developed by Collopy and Armstrong [8], [12]. It is an expert system that uses domain knowledge to combine forecasts from various forecasting methods. Using IF-THEN rules, RBF determines what weights to give to the forecasts.

We follow the main ideas of rule-based forecasting [8] and of using time series features [13] to obtain an interpretable and understandable model.

II. Fuzzy Rule-Based Ensemble

As mentioned above, RBF uses the rules to determine weights [8]. However, only few of these rules are directly used to set up weights. Most of them set up rather a specific model parameters, e.g. the smoothing factors of the Brown's exponential smoothing with trend. Moreover, in antecedents, the rules very often use properties that are not crisp but rather vague, e.g. expressions such as: "last observation is unusual; trend has been changing; unstable recent trend" etc., see [12]. For such cases, using crisp rules, that are either fired or not (with nothing between), seems to be less natural than using fuzzy rules. Similarly, the use of crisp consequents such as: "add 10% to the weight; subtract 0.4 from beta; add 0.1 to alpha" etc. [12], seems to be less intuitive than using vague expressions that are typical for fuzzy rules.

A. General Structure of the Model

Therefore, our goal was to propose a method that uses fuzzy rules instead of crisp rules in order to capture the omnipresent vagueness in the expressions; to use only quantitative features (no domain knowledge) in the antecedent variables which enable to fully automatize the method; to use only individual forecasting method weights as the consequent variables [14], [15]. The result of such motivated investigation is the Fuzzy Rule-Based Ensemble (FRBE) that is schematically illustrated in Figure 1.

The FRBE method uses a single *linguistic description*, i.e. fuzzy rule base with *evaluative linguistic expressions* [16], to determine a weight of each forecasting method based on transparent and interpretable rules, such as:

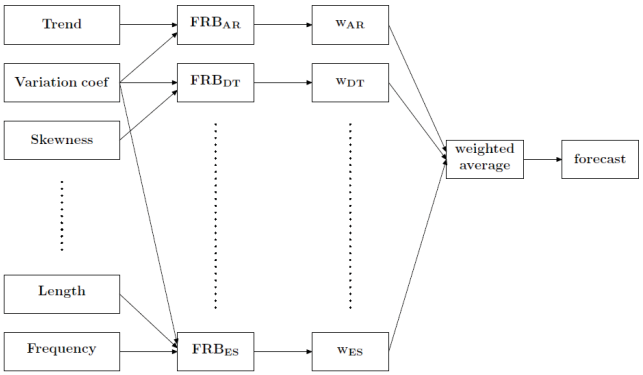

Fig. 1. A Structure of the FRBE method.

"**IF** *Strength of Seasonality is Small* **AND** *Coefficient of Variation is Roughly Small* **THEN** *Weight of the j-th method is Big*."

After an appropriate inference method is applied (see Section II-B) in order to obtain fuzzy output, a defuzzification method is employed and thus, a crisp result (weight of a particular method) is determined.

Based on experiments and previous publications [13], the following features were considered in introductory studies [14], [15], [17] as well as in this paper: *strength of trend, strength of seasonality, length of the time series, skewness, kurtosis, coefficient of variation, stationarity, and frequency*.

Based on listed features, the inference mechanism sets weights to the following forecasting methods in our ensemble: *seasonal Autoregressive Integrated Moving Average* (ARIMA), *Decomposition Techniques* (DT), *Exponential Smoothing* (ES), *Random Walk process* (RW), and *Random Walk process with a drift* (RWd). For details about these methods, we only refer to the relevant literature [1], [18], [19].

In this paper, motivated by the promising results published in [17], we stemmed from them and proceeded similarly yet with a wider perspective regarding the implementation and experimental justification.

B. Components of the Model

In order to estimate (set up) a particular value of the weight of each forecasting method with help of the fuzzy rules, an appropriate fuzzy inference mechanism has to be employed. As mentioned above, the FRBE method employs linguistic descriptions, i.e. fuzzy rule bases with so called evaluative linguistic expressions. These are expressions of natural language that are based on the expressions of the basic trichotomy `Small (Sm)`, `Medium (Me)`, and `Big (Bi)`. The expressions of the basic trichotomy may be modified using linguistic hedges either with narrowing or widening effect, see Table I.

Such linguistic expressions have their theoretical model of semantics based on intension, context, and extension, which is in detail described in the referred literature [16]. For the purpose of this contribution, it is sufficient to mention

TABLE I
LINGUISTIC HEDGES AND THEIR ABBREVIATIONS.

Narrowing Effect	Widening Effect
very (Ve)	more or less (ML)
significantly (Si)	roughly (Ro)
extremely (Ex)	quite roughly (QR)

that extensions, that model the meaning in a given context $[v_L, v_R]$, are fuzzy sets that are depicted in Figure 2. One may see the influence of the modifiers on the shape of the extensions.

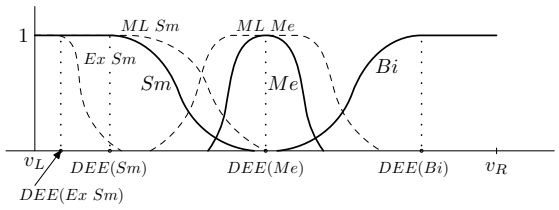

Fig. 2. Shapes of extensions (fuzzy sets) of evaluative linguistic expressions.

If a fuzzy rule base is viewed as a linguistic description, and thus uses the above recalled evaluative linguistic expressions with their model of semantics, one can neither model the rules (and consequently the whole description) as a conjunction of implicative rules nor as a disjunction of conjunctions (Mamdani-Assilian model). The used expressions, mainly their full overlapping, require a specific inference method – *Perception-based Logical Deduction* (PbLD) [20]. This method models each fuzzy rule

$$\mathcal{R}_i := \mathsf{IF}\ \mathtt{X}\ \text{is}\ \mathcal{A}_i\ \mathsf{THEN}\ \mathtt{Y}\ \text{is}\ \mathcal{B}_i,$$

by a fuzzy relation R_i on $X \times Y$ given as follows:

$$R_i(x, y) = A_i(x) \to_{\text{Ł}} B_i(y), \quad x \in X, y \in Y$$

where $\to_{\text{Ł}}$ is the Łukasiewicz implication [21] given by $a \to_{\text{Ł}} b = 1 \wedge (1 - a + b)$. For the sake of clarity, let us note that `X, Y` denote the so called linguistic variable that take values from a set of linguistic expressions, these linguistic expressions are modelled by fuzzy sets (extensions) on given universes (contexts) X, Y, and finally, $x \in X$ and $y \in Y$.

However, unlike in the case of implicative rules, the rules are not aggregated conjunctively. The PbLD uses a specific algorithm (perception) that chooses only some rules to be used in the inference. These are the most specific among the most fired rules. And only the outputs obtained based on these fuzzy rules are aggregated by the intersection at the final stage. For details regarding the algorithm, we refer to [22], [23].

Finally, the inferred output is defuzzified. This is done by the *Defuzzification of Evaluative Expressions* (DEE) that has been designed specifically for the outputs of the PbLD inference mechanism. In principle, this defuzzification is a combination of *First-Of-Maxima* (FOM), *Mean-Of-Maxima* (MOM) and *Last-Of-Maxima* (LOM) that are applied based on the classification of the inferred output fuzzy set. Particularly, if the inferred fuzzy set is of the type `Small`, the LOM is applied; if the inferred output is of the type `Medium`, the MOM is applied; and finally, if the inferred output is of the type `Big`, the FOM is applied, see Figure 2. In the case of the FRBE method, the defuzzification DEE is applied after the inference, so that the deduced weights $w_{AR}, w_{DT}, \ldots, w_{ES}$ displayed in Figure 1 are already crisp numbers.

C. Fuzzy Rule Base Identification

The last missing point is the identification of the linguistic descriptions. This may be done by distinct approaches. One could expect a deep applicable expert knowledge, however, neither our experience nor the experience of others confirms this expectations. Let us once more refer to the observation of Armstrong, Collopy, and Adya in [8], already recalled in Section I.

Because of the missing reliable expert knowledge, we focus on data-driven approaches that may bring us the interpretable knowledge hidden in the data.

However, before we apply any data-mining technique, we have to clarify how we interpret the weights in the data. Naturally, the individual method weights should be proportionally higher if a given method is supposed to provide lower forecasting error and vice-versa. Thus, it is natural to put

$$w_j = 1 - acc_j$$

where acc_j denotes an appropriate normalized forecasting error of the j-th method. Now, any appropriate data-mining technique may be applied in order to determine the dependence between features and the weight of each method.

III. FUZZY GUHA – LINGUISTIC ASSOCIATIONS MINING

In this paper, we employ the so called linguistic associations mining [24] for the fuzzy rule base identification. This approach, mostly known as mining association rules [25], was firstly introduced as GUHA method [26], [27]. It finds distinct statistically approved associations between attributes of given objects. Particularly, the GUHA method deals with Table II where $o_1, \ldots, o_n$ denote objects, $X_1, \ldots, X_m$ denote independent boolean attributes, Z denotes the dependent (explained) boolean attribute, and finally, symbols a_{ij} (or a_i) $\in \{0, 1\}$ denote whether an object o_i carries an attribute X_j (or Z) or not.

TABLE II
STANDARD GUHA TABLE.

	X_1	...	X_m	Z
o_1	a_{11}	...	a_{1m}	a_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
o_n	a_{n1}	...	a_{nm}	a_n

The original GUHA allowed only boolean attributes to be involved [28]. Since most of the features of objects are measured on the real interval, standard approach assumed to binarize the attributes by a partition of the interval into subintervals, see Example 3.1.

The goal of the GUHA method is to search for linguistic associations of the form

$$\mathtt{A}(X_1, \ldots, X_p) \simeq \mathtt{B}(Z)$$

where A, B are predicates containing only the connective AND and $X_1, \ldots, X_p$ for $p \leq m$ are all variables occurring in A. The A, B are called the *antecedent* and *consequent*, respectively. Generally, for the GUHA method, the well-known four-fold table is constructed, see Table III.

TABLE III
CLASSICAL GUHA FOUR-FOLD TABLE.

	B	not B
A	a	b
not A	c	d

Symbol a, in Table III, denotes the number of positive occurrences of A as well as B; b is the number of positive occurrences of A and of negated B, i.e. of 'not B'. Analogous meaning have the numbers c and d. For our purposes, only numbers a and b are important.

The relationship between the antecedent and consequent is described by the so called *quantifier* $\simeq$. There are many quantifiers that characterize validity of the association in data [27]. For our task, we use the so called *binary multitudinal quantifier* $\simeq := \sqsubset_r^\gamma$. This quantifier is taken as true if

$$\frac{a}{a+b} > \gamma \qquad \text{and} \qquad \frac{a}{n} > r,$$

where $\gamma \in [0,1]$ is a *degree of confidence* and $r \in [0,1]$ is a *degree of support*.

Example 3.1: For example, let us consider Table IV.

TABLE IV
EXAMPLE OF GUHA TABLE. $\mathrm{BMI}_{\leq 25}$ DENOTES BODY-MASS-INDEX LOWER OR EQUAL TO 25, $\mathrm{BMI}_{>25}$ DENOTES THE SAME INDEX ABOVE 25, $\mathrm{Chol}_{>6.2}$ DENOTES CHOLESTEROL HIGHER THAN 6.2 AND $\mathrm{BP}_{>130/90}$ DENOTES BLOOD PRESSURE HIGHER THAN 130/90. OBJECTS o_i ARE PARTICULAR PATIENTS.

	$\mathrm{BMI}_{\leq 25}$	$\mathrm{BMI}_{>25}$	$\mathrm{Chol}_{>6.2}$	$\mathrm{BP}_{>130/90}$
o_1	1	0	0	0
o_2	0	1	1	1
o_3	0	1	0	1
o_4	1	0	0	0
o_5	0	1	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
o_n	0	0	1	1

Depending on the chosen confidence and support degrees, the GUHA method could generate e.g. the following linguistic association:

$$\mathtt{A}(\mathrm{BMI}_{>25}, \mathrm{Ch}_{>6.2}) \simeq \mathtt{B}(\mathrm{BP}_{>130/90}),$$

which could be read as: "Body mass index higher than 25 AND cholesterol higher than 6.2 are associated with blood pressure higher than 130/90."

The chosen confidence and support degrees ensure that the association occurs in the given data in sufficiently high percentage (confidence) and sufficiently often (support).

In many situations, including ours, the fuzzy variant of the GUHA method [24], [29] seems to be more appropriate. We adopt the variant firstly used in [17] and described in detail in [30] that directly uses theory of evaluative linguistic expressions. Then the attributes are not boolean but vague such as $\mathrm{BMI}^{\mathrm{ExBi}}$, $\mathrm{BMI}^{\mathrm{MLBi}}$, $\mathrm{Chol}^{\mathrm{VeBi}}$ etc. With canonical adjectives Small, Medium, Big and eight different linguistic hedges including the empty one, we may define 24 fuzzy sets for every quantitative variable. The values a_{ij} (or a_i) are elements of the interval $[0,1]$ that express membership degrees to these fuzzy sets. Example of such a fuzzy GUHA table is provided in Table V.

TABLE V
EXAMPLE OF FUZZY GUHA TABLE. (COMPARE WITH TABLE IV.)

	$\mathrm{BMI}^{\mathrm{ExSm}}$	...	$\mathrm{Chol}^{\mathrm{ExBi}}$	$\mathrm{BP}^{\mathrm{ExSm}}$	...	$\mathrm{BP}^{\mathrm{ExBi}}$
o_1	0.5	...	0	1	...	0
o_2	0.8	...	0	0.4	...	0
o_3	0	...	0.1	0	...	0.4
o_4	0	...	0.4	0	...	0.3
o_5	0.6	...	0	1	...	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
o_n	0	...	0	0.5	...	0

The four-fold table analogous to Table III is constructed also for the used fuzzy variant of the method. The difference is that the numbers a, b, c, d are not summations of 1s and 0s, but summations of membership degrees of data into fuzzy sets representing the antecedent A, and consequent B, or their complements, respectively. Naturally, the fact, that antecedent A as well as consequent B hold simultaneously, leads to the natural use of a *t-norm* [31]. In our case, we use the Gödel t-norm that is the operation of minimum. For example, if an object o_i belongs to a given antecedent in a degree 0.7 and to a given consequent in a degree 0.6, the value that enters the summation equals to $\min\{0.7, 0.6\} = 0.6$. Summation of such values over all the objects equals to the value a in Table III, the other values from the table are determined analogously. The rest of the ideas of the method remain the same.

By using fuzzy sets, we generally get more precise results, and, more importantly, we avoid undesirable threshold effects [32]. The further advantage is that the method searches for implicative associations that may be directly interpreted as fuzzy rules for the PbLD inference system.

In our case, for each individual forecasting method, we have transformed the training data set of time series with their normalized features into a table similar to Table VI.

The rest of this section deals with ARIMA method. Of course, the same process has been applied for all the other forecasting methods in our ensemble.

TABLE VI

TRANSFORMED TRAINING DATA SET FOR THE ARIMA FORECASTING METHOD.

	Φ_1^{ExSm}	...	Φ_q^{ExBi}	$W_{\mathrm{AR}}^{\mathrm{ExSm}}$	...	$W_{\mathrm{AR}}^{\mathrm{ExBi}}$
TS_1	0.9	...	0.7	0	...	0.9
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
TS_n	0.1	...	0.2	0.8	...	0

Objects $\mathrm{TS}_1, \ldots, \mathrm{TS}_n$ in Table VI are the time series from the training set; $\Phi_1, \ldots, \Phi_q$ are normalized features of given time series. Note that there are significantly more columns in this part of Table VI because each evaluative linguistic expression leads to a single column for a single feature Φ_i, i.e. for the expression `ExSm`, there are q columns: $\Phi_1^{\mathrm{ExSm}}, \ldots, \Phi_q^{\mathrm{ExSm}}$, where q denotes the number of features. Once more, let us recall that we construct 24 linguistic expressions.

Symbol W_{AR} stands for the weight (inverted accuracy) of the ARIMA method, and again, there are as many columns in this part of the Table VI as there exist so many evaluative linguistic expressions, i.e. twenty-four in the chosen setting.

The fuzzy GUHA then combinatorically generates hypotheses that are immediately statistically either declined or confirmed as linguistic associations based on the chosen quantifier parameters, see Example 3.2.

Example 3.2: Our fuzzy GUHA approach provided us with the following implicative hypothesis:

$$\mathtt{A}(Season^{\mathrm{ExBi}}, Kurt^{\mathrm{QRSm}}) \ \sqsubset_r^{\gamma} \ \mathtt{B}(W_{\mathrm{AR}}^{\mathrm{Bi}})$$

where $Season$ denotes the normalized *strength of seasonality* and $Kurt$ denotes the *kurtosis*, that was confirmed on the following confidence degree and support degree, respectively:

$$\gamma = 0.65, r = 0.18\,.$$

Such a confirmed association may be viewed, and thus directly interpreted, as the following fuzzy rule:

"**IF** Strength of Seasonality is `Extremely Big` **AND** Kurtosis is `Quite Roughly Small` **THEN** Weight of the ARIMA method is `Big`."

For our purposes, we set up the thresholds for $\gamma = 0.65$ and $r = 0.05$.

Note that the above described application of the fuzzy GUHA method generates linguistic description determining the weight of a single method – in our example of the ARIMA method. Thus, the method, including the transformation of training data set into a table similar to Table VI, has to be applied as many times as is the number of methods (and consequently of the linguistic descriptions). In our case, this led to the fivefold use of the method as we deal with five individual methods.

IV. IMPLEMENTATION

To develop and validate the model, we have used 2829 time series from the M3 data set repository that contains 3003 time series from the M3-Competition [33]. We have omitted timeseries with other than yearly, quarterly, and monthly frequencies.

M3 set of time series serves as a generally accepted benchmark database provided by the authority of the International Institute of Forecasters. The time series are of 5 categories: Microeconomy, Macroeconomy, Industry, Finance, Demography.

This selected data set was divided into two distinct sets simply by putting time series with even or odd IDs into a *training* or *testing set*, respectively, see Table VII.

TABLE VII

SPLIT OF DATA INTO TRAINING AND TESTING SET. TOTAL TRAINING SET SIZE: 1414, TOTAL TESTING SET SIZE: 1415

Source	Training (Testing) Set		
	Monthly	Quarterly	Yearly
Demographic	55 (56)	28 (29)	123 (122)
Finance	73 (72)	38 (38)	29 (29)
Industry	167 (167)	42 (41)	51 (51)
Macro	156 (156)	168 (168)	41 (42)
Micro	237 (237)	102 (102)	73 (73)
Other	26 (26)	0 (0)	5 (6)
Total	714 (714)	378 (378)	322 (323)

The training set was used for an identification of our model, that is, for generation of our fuzzy rule base. The testing set was used for testing whether the determined knowledge encoded in the fuzzy rules works generally also for time series "not seen" by the rule base generation algorithm.

All forecasts were computed with the R software, version 3.0.2, and package `forecast` [34]. We have chosen the most often used forecasting methods: *seasonal Autoregressive Integrated Moving Average* (ARIMA), *Decomposition Techniques* (DT), *Exponential Smoothing* (ES), *Random Walk process* (RW) and *Random Walk process with a drift* (RWd).

These methods were executed with fully automatic parameter selection and optimization which made possible to concentrate the investigation purely on the combination technique. Moreover, their arithmetic mean (AM), that represents the equal weights ensemble method, was also determined and used as a valid benchmark.

There are many accuracy measures that are used to analyze the performance of the various forecasting methods. However, very popular measures such as *Mean Absolute Error* or *(Root) Mean Squared Error* are inappropriate for comparison across more time series because they are scale-dependent. Therefore, we use *Symmetric Mean Absolute Percentage Error* (SMAPE) that is scale-independent and thus, appropriate in order to compare methods across different time series [35]. This accuracy measure is defined as follows:

$$\mathrm{SMAPE} = \frac{1}{h} \sum_{t=T+1}^{T+h} \frac{|y_t - \hat{y}_t|}{(|y_t| + |\hat{y}_t|)/2} \times 100\%.$$

Let a given time series $y_1, y_2, \ldots, y_T$ is of the frequency F, i.e. $F = 1, 4, 12$, for yearly, quarterly, and monthly time series, respectively. Then the features used to predict weights

of the forecasting methods are defined as follows. The normalized *frequency* is given by the reciprocal value of F, i.e., it is given as 1/12, 1/4, and 1 in case of the monthly, quarterly, and yearly time series, respectively. The normalized *length of the time series* is given by $\min(T/100, 1)$ where T denotes the number of known time lags. The *skewness* is given as

$$skewness = \min\left(\frac{1}{6}\cdot\left(\frac{m_3}{m_2^{3/2}}+3\right), 1\right),$$

where $m_i = \frac{1}{T}\sum_{t=1}^{T}(y_t-\bar{y})^i$ and $\bar{y}$ is the arithmetic mean of the given time series $\{y_t\}_{t=1}^{T}$. The *kurtosis* is given as

$$kurtosis = \min\left(\frac{m_4}{10m_2^2}, 1\right),$$

with m_i given as above. The *coefficient of variation* is given as $CV = \min(s_y/\bar{y}, 1)$, where s_y is the standard deviation of $\{y_t\}_{t=1}^{T}$. The *strength of trend* is given by $(1-p)$, where p is a p-value of the statistical test of the null hypothesis $H_0 : \beta_1 = 0$, where β_1 is a slope parameter of the linear regression model:

$$y_t = \beta_0 + \beta_1 t + \varepsilon_i, \quad \text{for } t \in \{1, 2, \ldots, T\}.$$

The *strength of seasonality* is given by

$$1 - \min(p_2, p_3, \ldots, p_F),$$

where p_i for $i \in \{2, 3, \ldots, F\}$ is the p-value of the test of the null hypothesis $H_0 : \beta_i = 0$, where β_i is the coefficient of the linear regression model

$$y_t = \beta_0 + \beta_1 t + \beta_2 x_{t,2} + \beta_3 x_{t,3} + \ldots + \beta_F x_{t,F} + \varepsilon_i$$

for $t \in \{1, 2, \ldots, T\}$ and $x_{t,j} \in \{0, 1\}$ is an artificial variable such that $x_{t,j} = 1$ if $t \bmod F = j \bmod F$. Finally, the *stationarity* is given by $(1-p)$, where p is a p-value of the Augmented Dickey–Fuller Test of stationarity.

As you can see, many of the features have slightly different definition than expected by statisticians. The reason is that we need them to be *normalized* to the interval $[0, 1]$. Therefore, for instance, although a traditional definition of kurtosis is $\frac{m_4}{m_2^2} - 3$, we use $\min\left(\frac{m_4}{10m_2^2}, 1\right)$ to obtain a reasonable number in the interval $[0, 1]$ etc. Our future research will address the normalization of features more deeply.

V. Results

As mentioned above, the associations generated by GUHA method are implicative. Thus, they may be directly interpreted as fuzzy rules. Due to the large amount of such generated rules, a redundancy removal and size reduction algorithms were applied on these rules. The first process consists in an automatic detection and deletion of redundant rules based on a rather complicated and sophisticated, yet fully theoretically justified algorithm, see [22], [23]. After the application of the redundancy detection algorithm, the number of rules was significantly reduced, although not sufficiently in case of some huge fuzzy rule bases. Anyhow, the redundant rules are those that have to be deleted first. Only after the redundancy removal, a heuristic size reduction and simplification algorithm was applied again on the redundancy-free rule bases, for results see Table VIII.

TABLE VIII
Number of Rules Generated by the Fuzzy GUHA Method and Number of Rules After Post-Processing.

Method	Number of Rules After Application of Algorithms		
	Fuzzy GUHA	Redundancy Removal	Size Reduction
ARIMA	11206	9904	37
DT	63	29	10
ES	2244	1968	30
RW	153	49	14
RWd	2579	1941	45

In order to judge its performance, the fuzzy rule-based ensemble was applied on 1415 time series from the testing set. Table IX shows that arithmetic mean and standard deviation of SMAPE forecasting errors over all testing time series is better for fuzzy rule-based ensemble than any individual forecasting method. Moreover, the equal-weights, i.e. arithmetic mean (AM), has been outperformed as well.

TABLE IX
Average and Standard Deviation of the SMAPE Forecasting Errors.

Method	Error Average	Error Std.Dev.
ARIMA	14.58	16.77
DT	23.58	29.36
ES	14.31	16.44
RW	16.53	17.20
RWd	16.63	19.97
AM	14.73	16.88
FRBE	**13.93**	**15.47**

Although the improvement does not seem significant, it is evident that the fuzzy rule-based ensemble performs very well even against the equal-weights combining, i.e. a procedure that has performed well in prior studies.

To indicate superiority of our method, a statistical test of significance has been performed. Namely, we have performed Wilcoxon signed rank test with continuity correction for the null hypothesis that median of the random variable $(\mathrm{SMAPE}_{\mathrm{AM}} - \mathrm{SMAPE}_{\mathrm{FRBE}})$ equals to zero, with the non-zero equality alternative hypothesis. The null hypothesis was rejected in the standard significance level $\alpha = 0.05$. Particularly, the obtained p-value was less than 2.2×10^{-16}.

Let us stress that the victory has been reached not only in the accuracy but also in the robustness (standard deviation of the SMAPE forecasting errors, see Table IX), which is perhaps even more important w.r.t. the goals of ensemble methods.

To compare variances of $\mathrm{SMAPE}_{\mathrm{AM}}$ and $\mathrm{SMAPE}_{\mathrm{FRBE}}$, the F-test was performed. As a result, null hypothesis of ratio of variances being equal to 1 was rejected with p-value lower than 0.001. Also comparison of error variance of the FRBE method with error variance of all the individual methods (with p-values adjusted for multiple comparisons) indicate statistically significant differences with adjusted p-values lower than 0.02.

This investigation also confirms that there is really a dependence between time series features and success of forecasting method. This fact is good motivation to continue in this topic.

$\mathcal{R}_1$: **IF** *length* is *QRMe* **AND** *trendStrength* is *QRBi* **AND** *kurtosis* is *QRSm* **AND** *varcoef* is *QRMe* **AND** *stationarity* is *QRMe* **THEN** w_{DT} is *RoSm*.
$\mathcal{R}_2$: **IF** *length* is *QRMe* **AND** *trendStrength* is *Bi* **AND** *seasonStrength* is *QRMe* **AND** *stationarity* is *QRMe* **THEN** w_{DT} is *RoSm*.
$\mathcal{R}_3$: **IF** *trendStrength* is *VeBi* **AND** *skewness* is *QRMe* **AND** *stationarity* is *RoMe* **THEN** w_{DT} is *RoSm*.
$\mathcal{R}_4$: **IF** *length* is *QRMe* **AND** *trendStrength* is *VeBi* **AND** *kurtosis* is *QRSm* **AND** *varcoef* is *QRMe* **AND** *stationarity* is *QRMe* **THEN** w_{DT} is *MLSm*.
$\mathcal{R}_5$: **IF** *length* is *RoMe* **AND** *trendStrength* is *Bi* **AND** *varcoef* is *QRMe* **AND** *stationarity* is *QRMe* **THEN** w_{DT} is *MLSm*.
$\mathcal{R}_6$: **IF** *trendStrength* is *QRBi* **AND** *seasonStrength* is *QRSm* **AND** *skewness* is *QRMe* **THEN** w_{DT} is *QRSm*.
$\mathcal{R}_7$: **IF** *length* is *QRMe* **AND** *seasonStrength* is *QRMe* **AND** *stationarity* is *QRMe* **THEN** w_{DT} is *QRSm*.
$\mathcal{R}_8$: **IF** *length* is *RoMe* **AND** *varcoef* is *QRMe* **AND** *stationarity* is *QRMe* **THEN** w_{DT} is *QRSm*.
$\mathcal{R}_9$: **IF** *skewness* is *QRMe* **AND** *stationarity* is *RoMe* **THEN** w_{DT} is *QRSm*.
$\mathcal{R}_{10}$: **IF** *length* is *MLMe* **AND** *trendStrength* is *QRBi* **AND** *seasonStrength* is *QRSm* **THEN** w_{DT} is *QRSm*.

Fig. 3. Complete exemplary post-processed rule base for the Decomposition Techniques (DT) method.

In order to emphasize the linguistic nature of the approach, we provide readers with one of the linguistic descriptions generated by the fuzzy GUHA method in Figure 3. Because of the small number of generated rules, we choose the linguistic description that set up the weight of the DT method. The fuzzy rules symbolically displayed in Figure 3 can be easily read as conditional sentences of natural language. For example, let us take fuzzy rule $\mathcal{R}_9$, which may be read as follows:

"**IF** time series skewness is quite roughly medium **AND** its stationarity is roughly medium **THEN** the weight of Decomposition Techniques is quite roughly small."

Please note that the obtained rule base (Fig. 3) contains rules with consequents containing variants of "small" weight only. That means, the rule base captures circumstances of small weight of the DT method. For inputs that do not meet antecedent of any of the determined rules, no rule is fired, the output is constantly equal to one on the whole output domain and the defuzzified output equals precisely to its middle.

Recall, that we have chosen the weight to be proportional to the expected method accuracy and thus the weight and the accuracy may be freely replaced. This makes the rule even more interpretable, which underlines the goal of our approach.

VI. Critical Discussion and Future Directions

The obtained results showed an improvement in the accuracy as well as in the standard deviation of the accuracy that confirms the improvement in the sense of "robustness".

Let us now open a short critical discussion related to the results and the approach, generally. Undoubtedly, the results confirm some sort of improvement. One could surely express objections to the too slight improvement and also to the too difficult and technologically demanding approach. Both objections have to be taken seriously as they have reasonable cores.

Related to the first objection, we have to stress that we have tested the improvement in accuracy not only compared to the arithmetic mean but also compared to all the individual methods (with p-value adjustment for multiple comparisons).

The suggested FRBE method was found significantly better in median error than DE, RW and RWd. For ARIMA and Exponential Smoothing, the null hypothesis could not be rejected. The null hypothesis of the variance F-test was rejected in case of all the individual methods with no exception.

This means that the results provide maybe slight yet statistically significant improvement. This is not that much surprising, having in mind the extremely high number of time series in the testing set. However, this is nothing against the validity of the results. Vice-versa, the bigger the testing set, the better for the experimental justification.

One should also note the interesting fact that two of the individual methods, particularly ARIMA and ES, provided better results than the arithmetic mean AM. This is rather unusual observation that should not occur. It is a conclusion of the fact that the ensemble was composed of only two well-established methods, two naive methods, and one method (DT) that was significantly outperformed by all the others including the naive ones.

The choice of this unbalanced composition of the ensemble significantly reduced the positive influences of any possible combination, including the standard equal weights (AM) method.

This may be viewed even positively. The AM combination of two well-performing methods with one significantly worse method and two naive methods is nearly as good as the two well-working methods. The sophisticated FRBE even outperformed all the methods in error variance. This underlines the potential of such ensembles, especially of the FRBE.

Of course, the choice of individual methods is a crucial step. The way out of this problem clearly lies in a different composition of individual methods in the ensemble. On the other hand, too many methods might be counterproductive.

In order to avoid the trial-error approach, a stochastic optimization task will be implemented on a high performance computer in order to find the optimal setting of all "bricks" building the FRBE. This does not relate only to the individual methods, but also to the features itself, and their normalization.

For example, the used $(1-p)$-values (*strength of trend, strength of seasonality, stationarity*) lie in the $[0,1]$-interval and thus, are not further normalized anymore. However, $(1-p)$-values around 0.7 or 0.8 are extremely low from the statistical point of view, as p-value around 0.2 usually does not allow to reject null hypothesis. But within the standard context $[0,1]$, the values around 0.8 are found rather big. Narrowing the interval of p-values and consequently the

derived features given by $(1-p)$-values seems to be necessary. Nevertheless, the particular realization of the narrowed normalization is again an open question that may be solved within the more general optimization task performed by the stochastic optimization implemented on a supercomputer. This foreshadows the future direction of our research.

Regarding the second objections, let us stress that the difficulty appears only in the construction phase. In the final phase, that is planned to be reached, we expect to have a rather simple (from a user point of view) tool that will automatically determine a given time series features, use the pre-determined fuzzy rules to set-up weights of individual methods, perform individual method forecasts, combine them according to the determined weights, and finally, provide a user with a single accurate yet robust forecast. The actual study shows the potential to reach this task rather soon as the constructed ensembles demonstrated that: i) even the arithmetic mean that includes also an unreliable method may be comparable with the good ones; ii) FRBE composed of the same methods is sophisticated enough to even outperform the good methods though, 3 out of 5 individual methods used in FRBE should not be preferably used. The potential of FRBE composed only of well-performing methods and based on optimized (and optimally normalized) features can be only guessed at this phase of investigation and it is our goal number one in future investigations.

REFERENCES

[1] G. Box and G. Jenkins, *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day, 1976.

[2] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 15, pp. 116–132, 1985.

[3] J. Aznarte, J. Benítez, and J. Castro, "Smooth transition autoregressive models and fuzzy rule-based systems: Functional equivalence and consequences," *Fuzzy Sets and Systems*, vol. 158, pp. 2734–2745, 2007.

[4] V. Novák, M. Štěpnička, A. Dvořák, I. Perfilieva, V. Pavliska, and L. Vavříčková, "Analysis of seasonal time series using fuzzy approach," *International Journal of General Systems*, vol. 39, pp. 305–328, 2010.

[5] M. Štěpnička, A. Dvořák, V. Pavliska, and L. Vavříčková, "A linguistic approach to time series modeling with the help of the f-transform," *Fuzzy sets and systems*, vol. 180, pp. 164–184, 2011.

[6] G. Leng, T. McGinnity, and G. Prasad, "An approach for on-line extraction of fuzzy rules using a self-organising fuzzy neural network," *Fuzzy sets and systems*, vol. 150, pp. 211–243, 2005.

[7] H. J. Rong, N. Sundararajan, G. B. Huang, and P. Saratchandran, "Sequential adaptive fuzzy inference system (safis) for nonlinear system identification and prediction," *Fuzzy Sets and Systems*, vol. 157, pp. 1260–1275, 2006.

[8] J. S. Armstrong, M. Adya, and F. Collopy, "Rule-based forecasting using judgment in time series extrapolation," in *Principles of Forecasting: A handbook for reasearchers and practitioners*, J. S. Armstrong, Ed. Boston/Dordrecht/London: Kluwer Academic Publishers, 2001.

[9] J. M. Bates and C. W. J. Granger, "Combination of forecasts," *Operational Research Quarterly*, vol. 20, pp. 451–468, 1969.

[10] P. Newbold and C. W. J. Granger, "Experience with forecasting univariate time series and combination of forecasts," *Journal of the Royal Statistical Society Series a-Statistics in Society*, vol. 137, pp. 131–165, 1974.

[11] S. Makridakis, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler, "The accuracy of extrapolation (time-series) methods - results of a forecasting competition," *Journal of Forecasting*, vol. 1, pp. 111–153, 1982.

[12] F. Collopy and J. S. Armstrong, "Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations," *Management Science*, vol. 38, pp. 1394–1414, 1992.

[13] C. Lemke and B. Gabrys, "Meta-learning for time series forecasting in the nn gc1 competition," in *Proc. 16th IEEE Int. Conf. on Fuzzy Systems*, Barcelona, 2010, pp. 2258–2262.

[14] D. Sikora, M. Štěpnička, and L. Vavříčková, "Fuzzy rule-based ensemble forecasting: Introductory study," in *Synergies of Soft Computing and Statistics for Intelligent Data Analysis*, ser. Advances in Intelligent Systems and Computing, vol. 190. Springer-Verlag, 2013, pp. 379–387.

[15] ——, "On the potential of fuzzy rule-based ensemble forecasting," in *International Joint Conference CISIS'12 - ICEUTE'12 - SOCO'12 SPECIAL SESSIONS*, ser. Advances in Intelligent Systems and Computing, vol. 189. Springer-Verlag, 2013, pp. 487–496.

[16] V. Novák, "A comprehensive theory of trichotomous evaluative linguistic expressions," *Fuzzy Sets and Systems*, vol. 159, no. 22, pp. 2939–2969, 2008.

[17] L. Štěpničková, M. Štěpnička, and D. Sikora, "Fuzzy rule-based ensemble with use linguistic associations mining for time series prediction," in *Proc. 8th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-2013)*. Milano: Atlantic Press, 2013, pp. 408–415.

[18] J. D. Hamilton, *Time Series Analysis*. New Jersey: Princeton University Press, 1994.

[19] S. Makridakis, S. Wheelwright, and R. Hyndman, *Forecasting: methods and applications*. USA: John Wiley & Sons, 2008.

[20] V. Novák, "Perception-based logical deduction," in *Computational Intelligence, Theory and Applications*, ser. Advances in Soft Computing, B. Reusch, Ed. Berlin: Springer, 2005, pp. 237–250.

[21] V. Novák, I. Perfilieva, and J. Močkoř, *Mathematical Principles of Fuzzy Logic*. Boston: Kluwer Academic Publishers, 1999.

[22] A. Dvořák, M. Štěpnička, and L. Vavříčková, "Redundancies in systems of fuzzy/linguistic if-then rules," in *Proc. 7th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-2011) and LFA-2011*, ser. Advances in Intelligent Systems Research. Paris: Atlantic Press, 2011, pp. 1022–1029.

[23] L. Štěpničková, M. Štěpnička, and A. Dvořák, "New results on redundancies of fuzzy/linguistic if-then rules," in *Proc. 8th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-2013)*. Milano: Atlantic Press, 2013, pp. 400–407.

[24] J. Kupka and I. Tomanová, "Some extensions of mining of linguistic associations," *Neural Network World*, vol. 20, pp. 27–44, 2010.

[25] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th Int. Conf. on Very Large Databases*. Chile: AAAI Press, 1994, pp. 487–499.

[26] P. Hájek, "The question of a general concept of the GUHA method," *Kybernetika*, vol. 4, pp. 505–515, 1968.

[27] P. Hájek and T. Havránek, *Mechanizing hypothesis formation: Mathematical foundations for a general theory*. Berlin/Heidelberg/New York: Springer-Verlag, 1978.

[28] P. Hájek, M. Holeňa, and J. Rauch, "The GUHA method and its meaning for data mining," *Journal of Computer and Systems Sciences*, vol. 76, pp. 34–48, 2010.

[29] V. Novák, I. Perfilieva, A. Dvořák, Q. Chen, Q. Wei, and P. Yan, "Mining pure linguistic associations from numerical data," *International Journal of Approximate Reasoning*, vol. 48, pp. 44–22, 2008.

[30] M. Burda, P. Rusnok, and M. Štěpnička, "Mining linguistic associations for emergent flood prediction adjustment," *Adnavces in Fuzzy Systems*, 2013, , DOI: 10.1155/2013/131875.

[31] E. P. Klement, R. Mesiar, and E. Pap, *Triangular Norms*, ser. Trends in Logic. Dordrecht: Kluwer Academic Publishers, 2000, vol. 8.

[32] T. Sudkamp, "Examples, counterexamples, and measuring fuzzy associations," *Fuzzy Sets Systems*, vol. 149, no. 1, pp. 57–71, 2005.

[33] S. Makridakis and M. Hibon, "The m3–competition: results, conclusions and implications," *International Journal of Forecasting*, vol. 16, pp. 451–476, 2000.

[34] R. J. Hyndman, G. Athanasopoulos, S. Razbash, D. Schmidt, Z. Zhou, Y. Khan, and C. Bergmeir, *forecast: Forecasting functions for time series and linear models*, 2013, r package version 4.06. [Online]. Available: http://CRAN.R-project.org/package=forecast

[35] R. Hyndman and A. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, pp. 679–688, 2006.