An Improved Version of the Fuzzy Set Based Evolving Modeling with Multitask Learning

Amanda O. C. Ayres
School of Electrical and Computer Engineering
University of Campinas (Unicamp)
Campinas, SP, Brazil
https://orcid.org/0000-0002-6588-1940

Fernando J. Von Zuben
School of Electrical and Computer Engineering
University of Campinas (Unicamp)
Campinas, SP, Brazil
https://orcid.org/0000-0002-4128-5415

***Abstract*—This paper introduces two novel contributions to the online learning algorithm called Fuzzy set Based evolving Modeling with Multitask Learning (FBeM_MTL), the first algorithm in the literature to consider multitask learning in the context of data stream, adaptive and evolving systems. In this new version, the degree of intersection of the information granules is directly used to define a real-valued matrix representing the relationship among the learning tasks, responsible for defining the parameters of the consequent part of all functional IF-THEN fuzzy rules. Unlike the original FBeM_MTL, in this new version, we eliminated the need for the binarization of the matrix representing the connected rules, guiding to both performance improvement and reduction in the number of user-defined parameters. The second contribution is the adoption of the Weighted Least Squares (WLS) method to define the parameters of the consequent part of the rules, using the similarity measure between every pair of samples to the mean point to set their corresponding weights in the WLS problem. Computational experiments on time series prediction of weather temperature, rain precipitation, wind speed in eolian farms and stock exchange are used to validate the performance of this new version. When compared to the original FBeM_MTL and also to several other state-of-the-art evolving systems in the literature, our approach guides to competitive results using a reduced number of parameters.**

***Index Terms*—Evolving Fuzzy-Rule-Based Systems, Online Learning, Multitask Learning, Weighted Least Squares, Time Series Prediction.**

I. Introduction

In recent years, the large amount of data arriving at speeds never seen before has given rise to evolving systems, a field of Machine Learning (ML) where models are capable of self-adapting not only their parameters but also their internal structure in real-time. Notably, the evolving Fuzzy-Rule-Based (eFRB) systems are the evolving extension of the original Fuzzy-Rule-Based (FRB) systems [16], being first proposed by [2]. In this new approach, the learning mechanism for updating the information granules, each one acting as the antecedent of a fuzzy rule, is executed online in response to the input data stream. Besides, the parameters of the Takagi-Sugeno consequent part of the rules are automatically provided by solving linear regression problems [6].

Traditionally, each rule is responsible for its own learning in an independent way. Usually, no information is shared among them, thus neglecting the benefits of using Multitask Learning (MTL), which employs the joint treatment of multiple tasks [8], [26].

Although not vinculated to evolving systems, the first attempt in the literature to share knowledge among rules was [12], which assumes a low-dimensional subspace hiding the correlation information among all tasks. The parameters of those tasks are learned by an algorithm that is based on the ϵ-insensitive criterion and L2-norm penalty terms. On the other hand, the authors in [5] assumed, in an evolving approach, that the measure of the degree of intersection among the rules, readily extracted from their antecedents, more specifically from the evolving information granules, already provides a reliable and more parsimonious evolving model of the structural relationship supported by the rules. As a complement to the empirical loss, a regularization term was added to encode the structural relationship established by the rules. Among the several possibilities of incorporating the effect of those relationships as regularization constraints, the Sparse Structure-Regularized Learning with Least Squares Loss (Least SRMTL) was chosen. Despite its compactness, this strategy requires a routine to binarize the degree of relationship (restricted to the [0,1] interval) among the rules, which, besides demanding an extra parameter to the algorithm (values above a user-defined threshold will be interpreted as 1 and below such threshold will be converted to 0), results in a loss of information. Here, the effective degree of relationship will be directly used to model the structural dependencies among the rules.

An extensively used method to obtain the Takagi-Sugeno consequent parameters of each fuzzy rule is the Least Squares (LS) regression [20]. In this technique, all the samples designated to a particular rule receive the same relevance when calculating the consequent parameters, despite holding different membership degrees to the information granules. To the best of our knowledge, [13] is the only algorithm that employs the Weighted Least Squares (WLS) method [20] in the context of evolving systems, weighting each sample with the reverse of the distance to the center of the cluster to which it belongs. In this paper, we also employ the WLS method, but in the context

This work has been supported by grants from CNPq - Brazilian National Research Council, proc. #143455/ 2017-6 and proc. #307228/2018-5, and Fapesp proc. #2017/19397-9.

of the MTL approach introduced previously [5]. For that, an adaptation of the Accelerated Gradient method was developed [15], guiding to a version of a recursive WLS procedure.

The remaining sections of the paper are organized as follows: Section II presents the advances in the FBeM_MTL proposed in this paper, notably the new form to represent the connections among the rules and the adoption of the WLS method in the context of the MTL approach. Section III is dedicated to the computational experiments for the online prediction of weather temperature, rain precipitation, wind speed in eolian farms and S&P 500 Daily Closing Price. Whenever possible, a comparative analysis was performed considering several state-of-the-art evolving systems as contenders. Otherwise, the comparative analysis will solely involve the newly-proposed FBeM_MTL and the two extensions presented in this paper. Some concluding remarks and the further steps of the research compose Section IV.

II. An improved version of the Fuzzy set Based evolving Modeling with Multitask Learning

The new paradigm proposed by [5] introduces the MTL concept in the realm of eFRB systems. While the traditional approach used in the literature to update the Takagi-Sugeno consequent parameters does not share any information among the rules, authors in [5] state that what is learned by a rule can be helpful for other rules that have some behaviors in common. For this, the related rules are learned simultaneously by extracting and utilizing appropriate shared information across them.

Consider the training data set $\{\boldsymbol{x}^{i^{[h]}}, y^{i^{[h]}}\}_{h=1}^{N_i}$, where N_i is the quantity of data points assigned to rule R^i, $\boldsymbol{x}^{i^{[h]}} \in \mathbb{R}^n$ (n is the dimension of the input variable), and $y^{i^{[h]}} \in \mathbb{R}$. The matricial form of the input-output dataset associated with the i^{th} rule is expressed by:

$$X^i = \begin{bmatrix} 1 & \boldsymbol{x}^{i^{[1]}T} \\ 1 & \boldsymbol{x}^{i^{[2]}T} \\ \vdots & \vdots \\ 1 & \boldsymbol{x}^{i^{[N_i]}T} \end{bmatrix}, \quad \boldsymbol{y}^i = \begin{bmatrix} y^{i^{[1]}} \\ y^{i^{[2]}} \\ \vdots \\ y^{i^{[N_i]}} \end{bmatrix}. \tag{1}$$

In the new approach using MTL [5], the parameters of the Takagi-Sugeno consequent part $\boldsymbol{\theta^i} = [\theta_0^i, \ldots, \theta_j^i, \ldots, \theta_n^i]^T \in \mathbb{R}^{n+1}$, for each rule $R^i, i = 1, \ldots, r$, are obtained by solving the optimization problem in Equation (2):

$$\Theta^* = \arg\min_{\Theta} \sum_{i=1}^{r} \mathcal{L}(\boldsymbol{\theta^i}) + \Omega(\Theta), \tag{2}$$

where $\Theta = [\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^r] \in \mathbb{R}^{(n+1)\times r}$, $\mathcal{L}(\boldsymbol{\theta^i})$ is the empirical loss of rule R^i taking the input-output dataset and $\Omega(\Theta)$ is the regularization term that encodes the structural dependencies among the learning tasks.

The first contribution of this paper regards the first term of Equation (2), $\mathcal{L}(\boldsymbol{\theta^i})$. To calculate the prediction loss, we take into account the similarity between every sample $\boldsymbol{x}^{i^{[h]}}$ and the mean point of all the samples belonging to the i^{th} rule, as expressed by Equation (3):

$$s(\boldsymbol{x}^{i^{[h]}}, R^i) = 1 - \frac{1}{n}\sum_{j=1}^{n} |x_j^{i^{[h]}} - \frac{\sum_{h=1}^{N_i} x_j^{i\,[h]}}{N_i}|. \tag{3}$$

where $s(\boldsymbol{x}^{i^{[h]}}, R^i)$ can assume any value in the interval $[0, 1]$ whenever $x_j^{i\,[h]} \in [0, 1], \forall i, j, h$.

Let W^i be the diagonal matrix representing the similarity of each sample belonging to the i^{th} rule, as in Equation (4):

$$W^i = \mathrm{diag}(s(\boldsymbol{x}^{i^{[1]}}, R^i), s(\boldsymbol{x}^{i^{[2]}}, R^i), \ldots, s(\boldsymbol{x}^{i^{[N_i]}}, R^i)). \tag{4}$$

We consider this matrix W^i as the weights of the WLS method adopted in this paper. The prediction loss $\mathcal{L}(\boldsymbol{\theta^i})$ is calculated according to Equation (5). Instead of considering all samples with the same relevance, as was made before [5], now we adopt different penalty criteria according to the degree of similarity of the sample to the mean point of the rule.

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta^i}) &= \frac{1}{2}(\boldsymbol{y}^i - X^i\boldsymbol{\theta^i})^T W^i (\boldsymbol{y}^i - X^i\boldsymbol{\theta^i}) \\ &= \frac{1}{2}(\boldsymbol{y}^{i^T} W^i \boldsymbol{y}^i - 2\boldsymbol{\theta^i}^T X^{i^T} W^i \boldsymbol{y}^i + \boldsymbol{\theta^i}^T X^{i^T} W^i X^i \boldsymbol{\theta^i}) \end{aligned} \tag{5}$$

The further contribution of this paper concerns the second term of Equation (2), the regularization term encoding the dependencies established by the tasks, $\Omega(\Theta)$. Among the several assumptions on the structural dependencies leading to different regularization terms, in this paper, we will utilize a generalization of the Sparse Structure-Regularized Learning with Least Squares Loss (Least SRMTL) [27]. Unlike the original version of FBeM_MTL [5], the idea here is to define the regularization term by directly adopting the real-valued matrix representing the degree of connection among every pair of rules, and no more its binarized version. The advantage is twofold: the binarized version necessarily involves loss of information and demands an extra threshold to be properly defined by the user.

Consider the trapezoidal approach for the antecedent part of the rules as in [14] and let R^{i_1} be the rule whose antecedent part is defined as the fuzzy granule A^{i_1}, fully described by n quadruples $(l_j^{i_1}, \lambda_j^{i_1}, \Lambda_j^{i_1}, L_j^{i_1})$, $j = 1, \ldots, n$, and R^{i_2} be the rule whose antecedent part is defined as the fuzzy granule A^{i_2}, fully described by n quadruples $(l_j^{i_2}, \lambda_j^{i_2}, \Lambda_j^{i_2}, L_j^{i_2})$, $j = 1, \ldots, n$. The similarity measure between R^{i_1} and R^{i_2} is defined by [14]:

$$\begin{aligned} s(R^{i_1}, R^{i_2}) = & 1 - \frac{1}{4n}\sum_{j=1}^{n}(|l_j^{i_1} - l_j^{i_2}| + |\lambda_j^{i_1} - \lambda_j^{i_2}| \\ & + |\Lambda_j^{i_1} - \Lambda_j^{i_2}| + |L_j^{i_1} - L_j^{i_2}|). \end{aligned} \tag{6}$$

Calculating the similarity measure $s(R^{i_1}, R^{i_2})$ between every pair of rules $i_1, i_2 = 1, 2, \ldots, r$ (where r is the number of rules), we obtain the similarity matrix S of Equation (7), which may be directly interpreted as the pairwise dependencies among the rules [5]. One may note that S is symmetric and all the elements of S are in the interval [0, 1].

$$S = \begin{bmatrix} s(R^1,R^1) & s(R^1,R^2) & \dots & s(R^1,R^r) \\ s(R^2,R^1) & s(R^2,R^2) & \dots & s(R^2,R^r) \\ \vdots & \vdots & \ddots & \vdots \\ s(R^r,R^1) & s(R^r,R^2) & \dots & s(R^r,R^r) \end{bmatrix} \tag{7}$$

The approach introduced in this paper employs directly the information obtained by matrix S, thus avoiding the binarization routine adopted in [5]. Inspired by [27], we consider here a graph where each rule is a node, and an edge connects two nodes if their corresponding rules R^{i_1} and R^{i_2} are related, which happens every time that $s(R^{i_1}, R^{i_2}) > 0$. Let $\mathcal{E}$ be the set of edges, the edge k is represented as a vector $\boldsymbol{e}^k \in \mathbb{R}^r$ defined as follows: $\boldsymbol{e}^k_{i_1} = s(R^{i_1}, R^{i_2})$, $\boldsymbol{e}^k_{i_2} = -s(R^{i_1}, R^{i_2})$ and $\boldsymbol{e}^k_i = 0, i = 1, \dots, r, i \neq i_1, i \neq i_2$, if $s(R^{i_1}, R^{i_2}) > 0$. The complete graph is represented by matrix $G = [\boldsymbol{e}^1, \boldsymbol{e}^2, \dots, \boldsymbol{e}^{||\mathcal{E}||}] \in \mathbb{R}^{r \times ||\mathcal{E}||}$, where $||\mathcal{E}||$ is the cardinality of set $\mathcal{E}$ [5]. Supposing the absence of $s(R^{i_1}, R^{i_2}) = 0, i_1 = 1, \dots, r, i_2 = 1, \dots, r$ in matrix S, indicating that all rules are minimally interconnected, we have $||\mathcal{E}|| = r(r-1)/2$.

As an example, given the illustrative similarity matrix S of Equation (8), the corresponding matrix G representing the complete graph is expressed by Equation (9):

$$S = \begin{bmatrix} 1 & 0.15 & 0 & 0 & 0 \\ 0.15 & 1 & 0.25 & 0.12 & 0 \\ 0 & 0.25 & 1 & 0.35 & 0 \\ 0 & 0.12 & 0.35 & 1 & 0.20 \\ 0 & 0 & 0 & 0.20 & 1 \end{bmatrix}, \tag{8}$$

$$G = \begin{bmatrix} +0.15 & 0 & 0 & 0 & 0 \\ -0.15 & +0.25 & +0.12 & 0 & 0 \\ 0 & -0.25 & 0 & +0.35 & 0 \\ 0 & 0 & -0.12 & -0.35 & +0.20 \\ 0 & 0 & 0 & 0 & -0.20 \end{bmatrix}. \tag{9}$$

Equation (10) penalizes the Euclidean distance between all pairs of tasks connected in the graph as a direct proportion of their similarity measure [5], [27]:

$$\begin{aligned} ||\Theta G||_F^2 = & 0.15||\boldsymbol{\theta}^1 - \boldsymbol{\theta}^2||_2^2 + 0.25||\boldsymbol{\theta}^2 - \boldsymbol{\theta}^3||_2^2 + \\ & 0.12||\boldsymbol{\theta}^2 - \boldsymbol{\theta}^4||_2^2 + 0.35||\boldsymbol{\theta}^3 - \boldsymbol{\theta}^4||_2^2 + \\ & 0.20||\boldsymbol{\theta}^4 - \boldsymbol{\theta}^5||_2^2 \end{aligned} \tag{10}$$

where $||\cdot||_F^2$ is the squared Frobenius norm and $||\cdot||_2^2$ is the squared l_2-norm.

Finally, the regularization term adopted in this paper, $\Omega(\Theta)$, implementing the multitask perspective and generalized from the Sparse Structure-Regularized Learning with Least Squares Loss [27], is expressed by Equation (11):

$$\Omega(\Theta) = \rho||\Theta G||_F^2, \tag{11}$$

where the parameter ρ may be determined by a grid [25] or random search [7]. The term $\Omega(\Theta)$ forces the related tasks (informed in matrix G) to exhibit a low Euclidean distance between the corresponding pair of columns of matrix Θ [5]. The final expression of the optimization problem is given by Equation (12):

$$\Theta^* = \arg\min_{\Theta} \sum_{i=1}^{r} \frac{1}{2}(\boldsymbol{y}^i - X^i\boldsymbol{\theta^i})^T W^i (\boldsymbol{y}^i - X^i\boldsymbol{\theta^i}) + \rho||\Theta G||_F^2. \tag{12}$$

In the original version of the MTL approach proposed in [5], there were two additional regularization terms that implemented a kind of elastic net regularization [28] and forced the elements of matrix Θ to approach zero whenever the impact on the empirical loss was reduced. In the generalized version proposed by this paper, however, they were discarded without compromising the quality of the results and leading to the elimination of two additional user-defined parameters of the algorithm. In effect, the affine functions at the Takagi-Sugeno consequent part of the rules act locally and are hyperplanes, making the sub-models already regularized, thus reducing the potential benefit of those additional regularization terms.

To solve the problem of Equation (12), we adapted the accelerated gradient method [15] available on the MALSAR package [27]. The corresponding gradient of the objective function $J(\Theta)$ in Equation (12), associated with the parameters Θ, is defined by:

$$\frac{\partial J(\Theta)}{\partial \Theta} = \begin{bmatrix} \frac{\partial J_1(\boldsymbol{\theta^1})}{\partial \boldsymbol{\theta^1}} & \frac{\partial J_2(\boldsymbol{\theta^2})}{\partial \boldsymbol{\theta^2}} & \dots & \frac{\partial J_r(\boldsymbol{\theta^r})}{\partial \boldsymbol{\theta^r}} \end{bmatrix} + 2\rho\Theta G G^T, \tag{13}$$

where $\frac{\partial J_i(\boldsymbol{\theta^i})}{\partial \boldsymbol{\theta^i}}$ is defined as:

$$\frac{\partial J_i(\boldsymbol{\theta^i})}{\partial \boldsymbol{\theta^i}} = X^{i^T} W^i X^i \boldsymbol{\theta^i} - X^{i^T} W^i \boldsymbol{y^i}, \quad i = 1, \dots, r. \tag{14}$$

III. Computational Experiments

To compare the performance of this improved version of FBeM_MTL against (i) the original proposal in [5], (ii) popular ML algorithms, and (iii) several recently proposed evolving systems, four sets of computational experiments regarding weather forecasts around the world and stock exchange prediction were considered. They are described in the following sections. Evidently, other domains of application could have been considered, not restricted to time series prediction, such as adaptive filtering in signal processing and many other online regression problems.

In all experiments, the algorithms are evaluated using the root mean square error, whose formula is expressed in Equation (15). H is the total quantity of available data points, $y^{[h]}$ is the original value at instant h, and $\hat{y}^{[h]}$ is the prediction provided by the model at instant h.

$$RMSE = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(y^{[h]} - \hat{y}^{[h]})^2} \quad (15)$$

The values provided by all time series were normalized in the interval $[0, 1]$, considering the training dataset and according to Equation (16):

$$\overline{x}^i = \frac{x^i - \min([x^1, x^2, \ldots])}{\max([x^1, x^2, \ldots]) - \min([x^1, x^2, \ldots])}. \quad (16)$$

A. Precipitation prediction

The goal of this first experiment is to predict one step ahead monthly precipitation in several locations exhibiting diverse climate behavior in South America [1]. The dataset is provided by GPCC (Global Precipitation Climatology Centre) [22] and contains data from 1917 to 2016. Only the series that did not contain missing values were considered, which resulted in the 86 geographical locations indicated in Figure 1, with 1200 data points each.

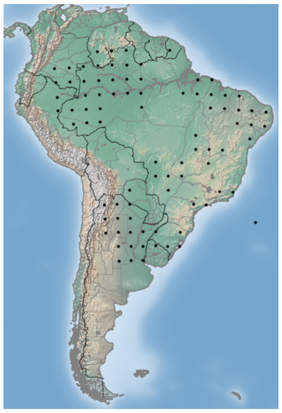


Fig. 1: Geographical locations of the precipitation time series (taken with permission from [1])

The number of lag variables selected is 2, after performing an autocorrelation analysis of the time series. The original dataset is used for training (to select the best hyperparameters), while the test is conducted with the flipped time series, as suggested by [11]. The hyperparameters of both the original and the improved version of FBeM_MTL are $\beta = 0.3$, $h_r = 12$, $\eta = 2$ and $\rho = 1000$, specified after a grid search strategy [25] taking the training dataset and considering the following candidate values: $\beta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $h_r \in \{3, 6, 12, 24\}$, $\eta \in \{0, 2, 5\}$ and $\rho \in \{10, 100, 1000, 10000\}$. A time window ($N$) of the last 24 data points was used in the solution of Equation (12), based on the seasonal behavior of the time series. For the original FBeM_MTL, there were three extra parameters that were eliminated in the improved version proposed here: $th_\sigma = 0.8$ and $\rho_2 = \rho_3 = 0$.

Table I provides the performance of all FBeM versions: the ancient FBeM [14] (batch version, without the MTL approach), the original FBeM_MTL [5] and the two new versions proposed in this paper. FBeM_MTL 2.0 refers to the version with the generalization of the Least SRMTL method, which besides using the real-valued matrix G and eliminating the binarization routine, reduces the number of hyperparameters of the algorithm. FBeM_MTL 2.1 builds on the improvements achieved in version 2.0 and introduces the WLS method to obtain the parameters of the Takagi-Sugeno consequent part of the rules. The average number of rules ($\overline{Rules}$) and root mean square error ($\overline{RMSE}$) are analyzed. In this first set of experiments, no other contender was considered, given that those datasets have never been explored before in the available literature of evolving systems. This new scenario will then attest that the relative behavior among the original and the two new versions of FBeM_MTL will be preserved no matter the particular nature of the climate time series. On the other hand, the remaining experiments will include additional contenders and the focus of analysis will be expanded.

TABLE I: One step ahead prediction performance for the time series of the precipitation in several locations of South America

Model	$\overline{Rules}$	$\overline{RMSE}$
FBeM	8.2390	0.1902
FBeM_MTL	8.2390	0.1803
FBeM_MTL 2.0	8.2390	0.1800
FBeM_MTL 2.1	8.2390	0.1800

As one may note, the number of rules follows the same trajectory for all the FBeM family, since the procedure to obtain the information granules remains unchanged. For the RMSE, the improvement in performance from FBeM to FBeM_MTL has already been observed in [5]. The two versions of the improved FBeM_MTL are responsible for an incremental reduction in the prediction error.

B. Temperature prediction

This second experiment intends to predict one step ahead of monthly temperatures in a weather time series for the cities Death Valley, Ottawa and Lisbon. It was first conceived by [14] and further extended by [5] for the MTL approach. The same representative contenders selected previously by [14] and [5] were used to properly evaluate the performance of the two improved versions of FBeM_MTL.

The hyperparameters of both the original and improved version of FBeM_MTL are $\beta = 0.7$, $h_r = 48$, $\eta = 0.5$, $\rho = 1$ and the last 12 observations of the time series, as in [14] and [5]. A time window (N) of the last 12 data points was used in the solution of Equation (12), corresponding to a year of observations. For the original FBeM_MTL, there were three extra parameters: $th_\sigma = 0.5$ and $\rho_2 = \rho_3 = 0$. The parameters of the other contenders follow the prescriptions of [14].

Table II presents the performance of the two improved versions of FBeM_MTL against several existing contenders in the literature, according to the parameters and results reported by [14] and [5]. The best values are highlighted in bold for each column.

TABLE II: One step ahead prediction performance for the time series of the temperature at Death Valley, Ottawa and Lisbon

Model	Death Valley		Ottawa		Lisbon	
	Rules	*RMSE*	Rules	*RMSE*	Rules	*RMSE*
MA*	-	0.167	-	0.162	-	0.141
SWMA*	-	0.083	-	0.081	-	0.071
MLP*	20	0.064	20	0.084	20	0.108
eTS*	**5**	0.086	8	0.084	7	0.094
xTS*	**5**	0.086	11	0.085	7	0.092
DENFIS*	13	0.068	23	0.086	27	0.094
FBeM_MTL*	8	**0.037**	**6**	0.049	7	0.051
FBeM_MTL 2.0	8	**0.037**	**6**	0.049	7	**0.050**
FBeM_MTL 2.1	8	**0.037**	**6**	**0.048**	7	**0.050**

* results obtained by [14] and [5]

The two improved versions of FBeM_MTL are capable of, while keeping the reduced number of rules of its predecessor, obtaining an incremental reduction in the *RMSE* measure when compared with its original version, outperforming all the other methods.

C. Wind speed prediction for eolian farms

This third experiment, first conceived by [5], consists of predicting the wind speed at the three largest wind farms in the United States: Alta Wind Energy Center, Roscoe Wind Farm and Shepherds Flat Wind Farm [23], whose locations are presented in Figure 2. For each wind farm, five well-distributed turbines were evaluated on an hourly time window basis during the year of 2012. The same evolving state-of-the-art prediction algorithms utilized by [5] as contenders were adopted.

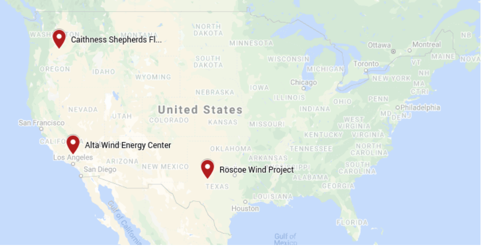


Fig. 2: Geographical locations of the wind energy farms

The hyperparameters of all the FBeM_MTL versions are $\beta = 0.9$, $h_r = 1344$, $\eta = 10$, $\rho = 100$ and the last 2 observations of the time series, as prescribed in [5]. A time window (N) of the last 24 data points was used in the solution of Equation (12), which is equivalent to the last day of wind speed observations. For the original FBeM_MTL, there were three extra parameters: $th_\sigma = 0.7$ and $\rho_2 = \rho_3 = 0$. The parameters of the other contenders follow the prescriptions of [5].

We compared statistically the *RMSE* obtained by the evolving algorithms employing the Friedman test [10], with $p = 0.05$ as the threshold. Whenever the null hypothesis is rejected, the Finner *posthoc* test is applied [9], with the same threshold, to verify the statistical advantage in a pairwise comparison. Table III presents the resulting statistical comparison. The table provides information about the rank of each algorithm, the average *RMSE* ($\overline{RMSE}$), the number of algorithms statistically better than the evaluated algorithm (#<) and the number of algorithms statistically worse than the evaluated algorithm (#>).

TABLE III: Ranking of the statistical comparison for the *RMSE*

Model	**rank**	$\overline{RMSE}$	**#<**	**#>**
FBeM_MTL 2.1	2.867	0.0455	0	4
ePL-KRLS	3.267	0.0459	0	4
FBeM_MTL 2.0	3.300	0.0456	0	4
FBeM_MTL	3.833	0.0457*	0	3
ePL	4.267	0.0507*	0	3
eTS	5.600	0.0616*	3	1
eTS-KRLS	6.467	0.0507*	5	1
eTS+	6.600	0.0509*	5	1
eTS-LS-SVM	8.800	0.0769*	8	0

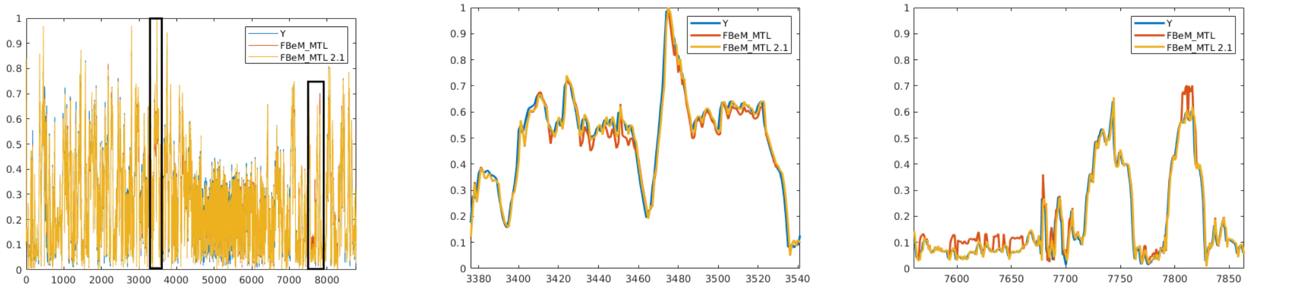
* results obtained by [5]

The rank is used to sort the rows of Table III. All the FBeM_MTL versions are highlighted in bold to facilitate the comparison with the other algorithms. The original FBeM_MTL version obtained the fourth position in the table, being statistically superior to the last three algorithms (column #>). With the improvements embedded by version 2.0 (the removing of the need for the binarization routine), we achieved a result slightly better, but yet behind the performance of ePL-KRLS [24], an evolving algorithm which combines, in the same model, the participatory learning paradigm with an adaptive method based on kernels for time series prediction. It was with the adoption of the WLS method in version 2.1, however, that FBeM_MTL was able to obtain the best performance compared to the other contenders, being statistically superior to the last four algorithms.

Figure 3 presents the predictions for the Site 36679 of Alta Wind Energy Center, comparing the performances of the original FBeM_MTL and the extended FBeM_MTL 2.1. In the general view of Figure 3a, the overall performance seems to be equivalent between the two versions. On the other hand, in Figures 3b and 3c, one may note the more accurate prediction provided by FBeM_MTL 2.1.

To illustrate the behavior of this new, improved version of FBeM_MTL, Figure 4 presents the evolution of the algorithm along time for the Site 36278 of Alta Wind Energy Center. The first row (Figures 4a to 4c) shows the distribution of the information granules (in fact, a top view of the trapezoidal shapes) at specific time instants and the other two rows, which correspond to the same time instants of row 1, refer to the contributions introduced in this paper: the second row (Figures 4d to 4f) depicts the operation of the generalized version of the Least SRMTL method, which considers directly the degree of connection among the pairs of rules without the need for any binarization process, and the third row (Figures 4g to 4i) illustrates the weighting policy of the WLS approach.

In row 2, the graphs' nodes are numbered and their coordinates are calculated as the center of the core region of the granules. The width of the edges is proportional to

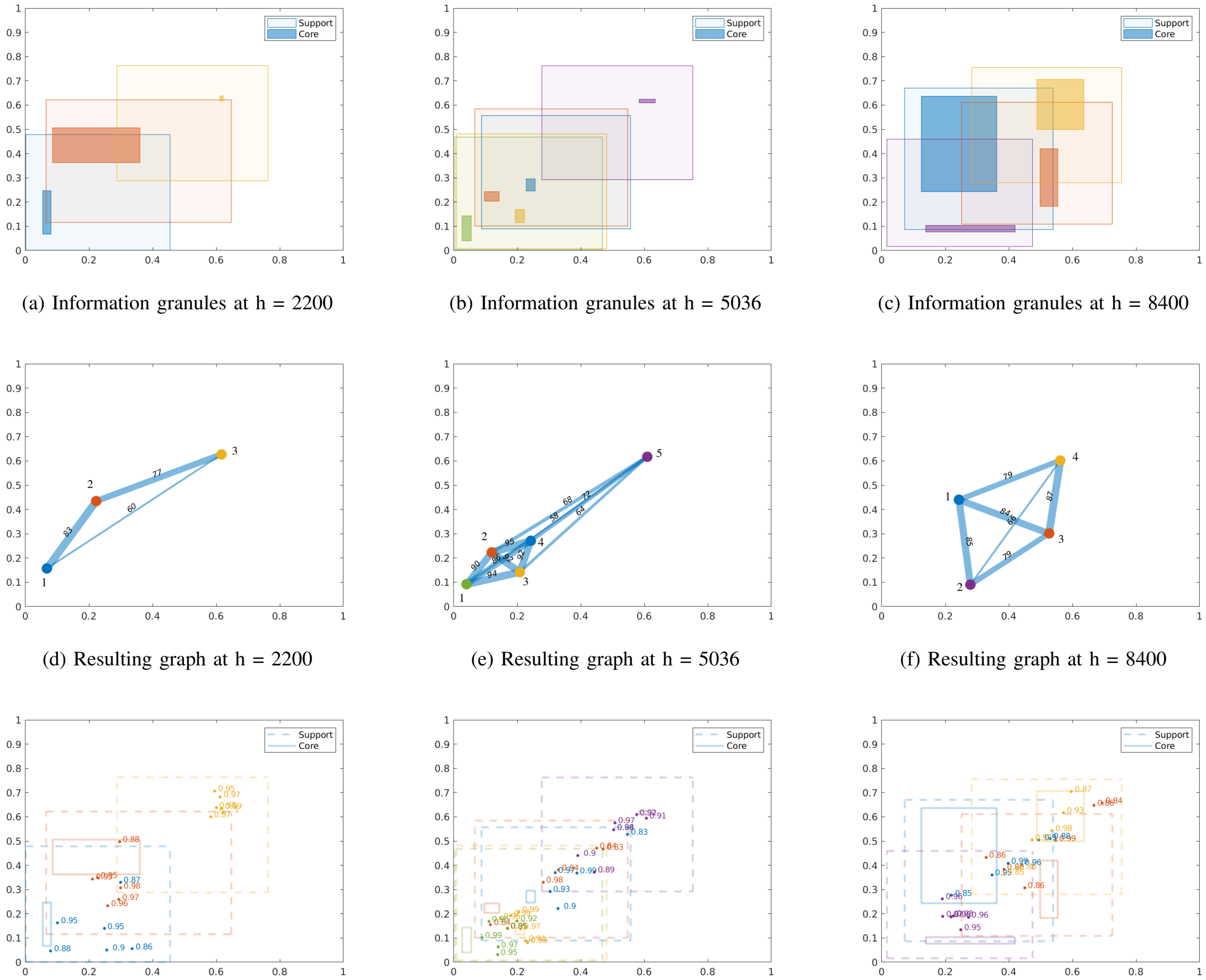


(a) General view

(b) First zoom

(c) Second zoom

Fig. 3: Comparison of the prediction performance for the Site 36679 of Alta Wind Energy Center



(a) Information granules at h = 2200

(b) Information granules at h = 5036

(c) Information granules at h = 8400

(d) Resulting graph at h = 2200

(e) Resulting graph at h = 5036

(f) Resulting graph at h = 8400

(g) Some samples and their weights at h = 2200

(h) Some samples and their weights at h = 5036

(i) Some samples and their weights at h = 8400

Fig. 4: Evolution of the behavior of the improved FBeM_MTL along time for the Site 36278 of Alta Wind Energy Center. Another version of Figures 4a, 4b and 4c were already published in [5]

the similarity between each pair of connected rules. The corresponding values, which are indicated above the edges, were multiplied by 100 to turn the graphs more visible.

At the time instant $h = 2200$, Figure 4a presents three scattered granules, with a central granule connected more intensively with the other two, as Figure 4d shows. The resulting G matrix is presented in Equation (17a). At the time instant $h = 5036$, Figure 4b indicates the creation of new information granules, which are more specialized in specific regions. They are concentrated more in the lower left region of the figure, with the exception of one more distant granule. As a consequence of the overlapping of granules in this area, Figure 4e shows nodes with a more intense connection rate, which is also observed in the resulting matrix G of Equation (17b). Finally, at $h = 8400$, Figure 4c exhibits the granules at the end of the time series. They are more equally spaced, which reflects in the connection pattern for the nodes of Figure 4f and in the matrix G of Equation (17c).

In row 3, we plot some of the samples belonging to the information granules with their corresponding weight, calculated by Equation (3). As one may note, to samples located far from the denser region lower weights are assigned and, as a consequence, they contribute less in the prediction loss given by Equation (5).

D. Online Prediction of S&P 500 Daily Closing Price

This fourth experiment consists of predicting the S&P 500 Daily Closing Price, a commonly employed time series used in the ML community and available publicly. The period analyzed comprises the dates 03.01.1950 and 12.03.2009, totalizing 14,893 data points. As recommended by [11], the original dataset is used for training, and the flipped time series is applied for testing. The output of the prediction model is calculated considering the last five observations of the time series, according to Equation (18):

$$\hat{x}(h+1) = f(x(h-4), x(h-3), x(h-2), x(h-1), x(h)). \tag{18}$$

Table IV presents the performance of all FBeM_MTL versions compared to several contenders in the literature. The hyperparameters are set to $\beta = 0.05$, $h_r = 50$, $\eta = 2$ and $\rho = 1$, after a grid search strategy [25] taking the training dataset and considering the following candidate values: $\beta \in \{0.01, 0.05, 0.1\}$, $h_r \in \{10, 50, 100\}$, $\eta \in \{0, 2, 5\}$ and $\rho \in \{0, 1, 10\}$, respectively. A time window (N) of the last 4 data points was considered. For the original FBeM_MTL, there were three extra parameters: $th_\sigma = 0.8$ and $\rho_2 = \rho_3 = 0$.

TABLE IV: Online Prediction of S&P 500 Daily Closing Price

	No. of rules (AVG.)	NDEI
PANFIS [17]	4	0.09
GENEFIS [18]	2	0.07
eT2RFNN [19]	2	0.04
Simpl_eTS [3]	7	0.04
eTS [4]	14	0.04
SEFS [11]	2(1.2835)	0.0182
FBeM_MTL [5]	1(2.0872)	0.0203
FBeM_MTL 2.0	1(2.0872)	0.0205
FBeM_MTL 2.1	1(2.0872)	0.0203

Figure 5 shows the prediction of FBeM_MTL 2.1 for the test data. Although the extended versions were not able to improve the performance of FBeM_MTL for this particular case, notice that FBeM_MTL 2.1 operates with three less user-defined parameters. FBeM_MTL and FBeM_MTL 2.1 achieved the second-best result, just behind SEFS [11], an approach that uses online training errors to automatically update the threshold parameter that controls the number of rules. The volatile nature of S&P 500 seems to make this time series more suitable for SEFS.

IV. Concluding remarks

This paper presents an improved version of the newly proposed Fuzzy set Based evolving Modeling with Multitask Learning (FBeM_MTL) [5], which showed, by the first time in the literature of evolving systems, the benefits of using multitask learning to share knowledge when determining the Takagi-Sugeno consequent parameters of the online composition of fuzzy IF-THEN rules. Given that each rule has a subset of linked samples, we extended the original Least Squares Regularized Optimization to a Weighted Least Squares (WLS) Regularized Optimization, with each weight being inversely proportional to the distance of the corresponding sample to the center of mass of the linked rule. Intuitively, it raises the impact of samples located at denser areas when compared with samples situated further away. Besides that, we also

$$G_{2200} = \begin{bmatrix} +0.83 & +0.60 & 0 \\ -0.83 & 0 & 0.77 \\ 0 & -0.60 & -0.77 \end{bmatrix} \tag{17a}$$

$$G_{5036} = \begin{bmatrix} +0.90 & +0.94 & +0.86 & +0.58 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.90 & 0 & 0 & 0 & +0.92 & +0.95 & +0.68 & 0 & 0 & 0 \\ 0 & -0.94 & 0 & 0 & -0.92 & 0 & 0 & +0.92 & +0.64 & 0 \\ 0 & 0 & -0.86 & 0 & 0 & -0.95 & 0 & -0.92 & 0 & +0.72 \\ 0 & 0 & 0 & -0.58 & 0 & 0 & -0.68 & 0 & -0.64 & -0.72 \end{bmatrix} \tag{17b}$$

$$G_{8400} = \begin{bmatrix} +0.85 & +0.84 & +0.79 & 0 & 0 & 0 \\ -0.85 & 0 & 0 & +0.79 & +0.66 & 0 \\ 0 & -0.84 & 0 & -0.79 & 0 & +0.87 \\ 0 & 0 & -0.79 & 0 & -0.66 & -0.87 \end{bmatrix} \tag{17c}$$

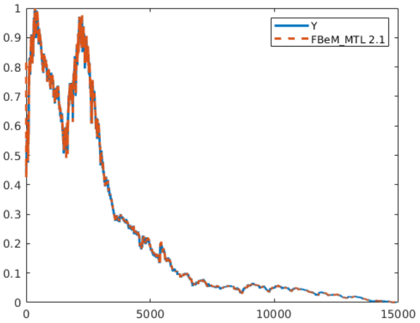


Fig. 5: Prediction of the S&P 500 Daily Closing Price

developed a generalization of the regularization term that encodes the structural dependencies among the rules. Unlike the original approach, which considered a binary connection between two rules, in this new formulation we directly resorted to the degree of intersection of the information granules to define a real-valued matrix representing a fuzzy relationship among the rules, avoiding the loss of information that the binarization routine brings, and also eliminating a threshold hyperparameter that drives the binarization and should be properly defined by the user. Lastly, after detecting a low impact in the final solution, we decided to eliminate two elastic net-like regularization terms from the optimization problem, thus guiding to the reduction in the number of hyperparameters to be set by the user and bringing compactness to the model, without compromising performance.

To analyze the impact of the contributions proposed by the paper, we conducted a series of four experiments in the challenging fields of weather forecast—more specifically on time series prediction of precipitation, temperature and wind speed for eolian farms—and of stock exchange prediction. We compared the performance of the two extended versions of FBeM_MTL with the original version and also, whenever possible, with several existing algorithms in the literature. The experiments evidenced the benefits of the contributions presented in this paper.

As future work, we plan to extend the MTL approach to deal with a regularized version of generalized linear models, such as kernel regression [21], in replacement to the traditional regularized linear regression.

References

[1] N. S. Aguiar, "A multitask learning approach to automatic threshold selection in Pareto distributions," Master's thesis, University of Campinas, 2019.

[2] P. Angelov and R. Buswell, "Evolving rule-based models: A tool for intelligent adaptation," in *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, vol. 2. IEEE, 2001, pp. 1062–1067.

[3] P. Angelov and D. Filev, "Simpl_ets: A simplified method for learning evolving takagi-sugeno fuzzy models," in *The 14th IEEE International Conference on Fuzzy Systems, 2005. FUZZ'05.* IEEE, 2005, pp. 1068–1073.

[4] P. P. Angelov and D. P. Filev, "An approach to online identification of takagi-sugeno fuzzy models," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 1, pp. 484–498, 2004.

[5] A. O. C. Ayres and F. J. Von Zuben, "Multitask learning applied to evolving fuzzy-rule-based predictors," *Evolving Systems*, pp. 1–16, 2019.

[6] R. D. Baruah and P. Angelov, "Evolving fuzzy systems for data streams: a survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 6, pp. 461–476, 2011.

[7] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. 2, pp. 281–305, 2012.

[8] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[9] H. Finner, "On a monotonicity problem in step-down multiple test procedures," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 920–923, 1993.

[10] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.

[11] D. Ge and X.-J. Zeng, "A self-evolving fuzzy system which learns dynamic threshold parameter by itself," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 8, pp. 1625–1637, 2018.

[12] Y. Jiang, F.-L. Chung, H. Ishibuchi, Z. Deng, and S. Wang, "Multitask TSK fuzzy system modeling by mining intertask common hidden structure," *IEEE Transactions on Cybernetics*, vol. 45, no. 3, pp. 534–547, 2015.

[13] N. K. Kasabov and Q. Song, "Denfis: dynamic evolving neural-fuzzy inference system and its application for time-series prediction," *IEEE transactions on Fuzzy Systems*, vol. 10, no. 2, pp. 144–154, 2002.

[14] D. Leite, R. Ballini, P. Costa, and F. Gomide, "Evolving fuzzy granular modeling from nonstationary fuzzy data streams," *Evolving Systems*, vol. 3, no. 2, pp. 65–79, 2012.

[15] Y. Nesterov, "Gradient methods for minimizing composite functions," *Mathematical Programming*, vol. 140, no. 1, pp. 125–161, 2013.

[16] W. Pedrycz and F. Gomide, *Fuzzy systems engineering: toward human-centric computing*. John Wiley & Sons, 2007.

[17] M. Pratama, S. G. Anavatti, P. P. Angelov, and E. Lughofer, "Panfis: A novel incremental learning machine," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 1, pp. 55–68, 2013.

[18] M. Pratama, S. G. Anavatti, and E. Lughofer, "Genefis: toward an effective localist network," *IEEE Transactions on Fuzzy Systems*, vol. 22, no. 3, pp. 547–562, 2013.

[19] M. Pratama, J. Lu, E. Lughofer, G. Zhang, and M. J. Er, "An incremental learning of concept drifts using evolving type-2 recurrent fuzzy neural networks," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 5, pp. 1175–1192, 2016.

[20] C. R. Rao, H. Toutenburg, C. Heumann *et al.*, *Linear models and generalizations: least squares and alternatives*. Springer Science & Business Media, 2007.

[21] J. D. A. Santos and G. A. Barreto, "A regularized estimation framework for online sparse LSSVR models," *Neurocomputing*, vol. 238, pp. 114–125, 2017.

[22] U. Schneider, A. Becker, P. Finger, A. Meyer-Christoffer, B. Rudolf, and M. Ziese, "Gpcc full data reanalysis version 7.0: Monthly land-surface precipitation from rain gauges built on gts based and historic data," Boulder CO, 2016. [Online]. Available: https://doi.org/10.5065/D6000072

[23] T. T. Tran and A. D. Smith, "Evaluation of renewable energy technologies and their potential for technical integration and cost-effective use within the US energy sector," *Renewable and Sustainable Energy Reviews*, vol. 80, pp. 1372–1388, 2017.

[24] R. Vieira, R. Ballini, and F. Gomide, "Kernel evolving participatory fuzzy modeling for time series forecasting," in *Proceedings of IEEE World Congress on Computational Intelligence (WCCI)*. IEEE, 2018, pp. 157–165.

[25] Z. B. Zabinsky, *Stochastic adaptive search for global optimization*. Springer Science & Business Media, 2013, vol. 72.

[26] Y. Zhang and Q. Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.

[27] J. Zhou, J. Chen, and J. Ye, "User's Manual MALSAR: Multi-tAsk Learning via StructurAl Regularization," Arizona State University, Tech. Rep., 2012.

[28] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.