CENTRE FOR INFORMATICS AND SYSTEMS
Adaptive Computation Group

Evolving Takagi-Sugeno Fuzzy Models

TECHNICAL REPORT

Version 1.0.2

José Victor Ramos and António Dourado

Coimbra, September 2003

Abstract

The approach proposed by Angelov for on-line learning of Takagi-Sugeno (TS) type models is described in this technical report. It is based on a novel learning algorithm that recursively updates TS model structure and parameters by combining supervised and unsupervised learning. The rule-base and parameters of the TS model continually evolve by adding new rules with more summarization power and by modifying existing rules and parameters. In this way, the rule-base structure is inherited and updated when new data become available.

The adaptive nature of these evolving TS models in combination with the highly transparent and compact form of fuzzy rules makes them a promising candidate for on-line modelling. The approach has significantly wider implications in a number of fields, including forecasting, fault detection, control, behavior modelling and knowledge extraction.

Key words: Takagi-Sugeno models, rule-base adaptation, fuzzy clustering, subtractive clustering, on-line learning.

Index

1. Introduction

In the last years significant attention has been given to data-driven techniques for generation of flexible models and among these techniques are fuzzy systems (Takagi-Sugeno and Mamdani models) and neural networks, which are universal approximators. In the 80's and early 90's most of the models design rely on the subjective expert knowledge with well known drawbacks since the efforts have been mostly directed to tuning and static optimization of rules generated by experts. The fact that nowadays huge quantities and forms of data exists give rise to the fast development of the data mining and related knowledge extraction techniques. These techniques, however, are still mostly applied do classification and off-line modelling.

At the present there are clear demands for effective approaches to design autonomous, self-developing and self-enriching systems, which at the same time should be flexible and robust. Their computational efficiency and compactness are prerequisites of a practical application. The problems of on-line applications are mostly related to the non-linear nature of the rule-base/network structure and computational expenses of the training technique, which hampers development of recursive, adaptive schemes. Some practical applications of fuzzy systems and fuzzy neural networks are called self-learning or adaptive, but they are rather self-adjusting and self-tuning since they, normally, suppose structure of the model to be fixed. Algorithms for on-line learning with self-constructing or evolving structure have been reported recently and independently for the fuzzy and neural network models.

Takagi-Sugeno models have recently become a powerful practical engineering tool for modelling of complex systems. They form a natural transition between conventional and rule-based systems by expanding and generalizing the concept of gain scheduling. While gain scheduling paradigm is based on the assumption of local approximation of a nonlinear system by a collection of linear models, the TS fuzzy models use the idea of linearization in a fuzzily defined region of the state space. Due to the fuzzy regions, the nonlinear system is decomposed into a multi-model structure consisting of linear models that are not necessarily independent. The TS fuzzy model representation often provides efficient and computationally attractive solutions to a wide range of problems introducing a powerful multiple model structure that is capable to approximate nonlinear dynamics, multiple operating modes and significant parameter and structure variations.

Evolving rule-based models use methods for learning TS fuzzy models from data are based on the idea of consecutive structure and parameter identification. Structure identification includes estimation of the focal points of the rules (antecedent parameters) by fuzzy clustering. With fixed antecedent parameters, the TS model transforms into a linear model. Parameters of the linear models associated with each of the rule antecedents are obtained by applying the recursive least-squares (RLS) method or the weighted recursive least-squares (wRLS) method.

For on-line learning of the TS fuzzy models it is necessary an on-line clustering method responsible for the model structure learning. Angelov proposed a new method inspired on the subtractive clustering algorithm that allows the recursive calculation of the information potential of the new data sample, which represents a spatial proximity measure used to define the focal points of the rules (antecedent parameters). Evolving rule-based models use the information potential of the new data sample as a trigger to update the rule-base, which ensures greater generality of the structural changes (Angelov and Filev, 2003).

The evolution mechanism is basically the following: if the information potential of the new data sample is higher than the potential of the existing rules a new focal point (rule) is created. If the new focal point is too close to a previously existing rule then the old rule is replaced by the new one. A new rule is generated only if there is significant new information present in the data. The appearance of a new rule indicates a region of the data space that has not been covered by the existing rules. This could be a new operating mode of the plant or reaction to a new disturbance.

The advantage of using the information potential instead of the distance to a certain rule centre only for forming the rule base is that the spatial information and history are not ignored, but are part of the decision whether to upgrade or modify the rule base. This interesting feature makes the approach potentially very useful as a tool for accumulation of knowledge (Angelov and Filev, 2003).

2. Identification of Takagi-Sugeno Fuzzy Models

Fuzzy model identification has its roots in the pioneering papers of Sugeno and his coworkers and is associated with the so called Takagi-Sugeno fuzzy models, a special group of rule-based models with fuzzy antecedents and functional consequents that follow from the Takagi-Sugeno-Kang reasoning method:

$$\Re_i : If\ x_1\ is\ X_{i1} \ldots and\ x_n\ is\ X_{in}\ then\ y_i = a_{i1}x_1 + \cdots + a_{in}x_n + b_i;\ i = 1,2,\ldots,R$$

where $\Re_i$ denotes the i^{th} fuzzy rule; R is the number of fuzzy rules; $\boldsymbol{x}$ is the input vector, $x = [x_1, x_2, \ldots, x_n]^T$; X_{ij} denotes the antecedent fuzzy sets, $j = 1,2,\ldots,n$; y_i is the output of the i^{th} rule; a_{ij} and b_i are parameters of the consequence.

The degree of firing of each rule, μ_i, is proportional to the level of contribution of the corresponding linear model to the overall output of the TS model. It is determined by the Gaussian law, which ensures the greatest possible generalization:

$$\mu_{ij} = e^{-\alpha\left\|x_j - x_i^*\right\|};\ i = 1,2,\ldots,R;\ \ j = 1,2,\ldots,n$$

where $\alpha = \frac{4}{r^2}$ and r is a positive constant, which defines the spread of the antecedent and the zone of influence of the i^{th} model (radius of the neighbourhood of a data point); x_i^* is the focal point of the i^{th} rule antecedent.

The firing level of the rules is defined as the Cartesian product or conjunction of respective fuzzy sets for this rule:

$$\tau_i = \mu_{i1}(x_1) \times \mu_{i2}(x_1) \times \cdots \times \mu_{in}(x_n) = \bigcap_{j=1}^{n} \mu_{ij}(x_j)$$

The TS model output is calculated by weighted averaging of individual rules contributions:

$$y = \sum_{i=1}^{R} \lambda_i y_i = \sum_{i=1}^{R} \lambda_i x_e^T \pi_i$$

where

$$\lambda_i = \frac{\tau_i}{\sum_{j=1}^{R} \tau_j}$$

is the normalized firing level of the i^{th} rule; y_i represents the output of the i^{th} linear model; $\pi_i = \begin{bmatrix} a_{i0} & a_{i1} & a_{i2} & \dots & a_{in} \end{bmatrix}^T$, $i = 1, \dots, R$ is the vector of the parameters of the i^{th} linear model; $x_e = \begin{bmatrix} 1 & x^T \end{bmatrix}^T$ is the expanded data vector.

Generally, the problem of identification of a TS model is divided into two sub-tasks:

- Learning the antecedent part of the model, which consists on the determination of the centres and spreads of the membership functions;
- Learning the parameters of the linear subsystems of the consequents.

2.1 Learning Rule Antecedents by Data Space Clustering

First sub-task can be solved by clustering the input-output data space $\left(z = \begin{bmatrix} x^T ; y \end{bmatrix}^T\right)$. The subtractive clustering method, fuzzy c-means and the Gustafson-Kessel clustering method are among well-established methods for learning the antecedent parameters off-line in a batch-processing learning mode when all the input-output data is available.

The procedure called subtractive clustering is an improved version of the so-called mountain clustering approach. It uses the data points as candidate prototype cluster centres. The capability of a point to be a cluster centre is evaluated through its potential, a measure of the spatial proximity between a particular point z_i and all other data points:

$$P_i = \frac{1}{TD}\sum_{j=1}^{TD} e^{-\alpha\left\|z_i - z_j\right\|^2} \; ; \; i = 1,...,TD$$

where P_i denote the potential of the *i*th data point, *TD* is the number of training data and $\alpha = \frac{4}{r_a^2}$. r_a is a positive constant, called radii.

The value of the potential is higher for a data point that is surrounded by a large number of close data points. Therefore, it is reasonable to establish such a point to be the centre of a cluster. The potential of all other data points is reduced by an amount proportional to the potential of the chosen data point and inversely proportional to the distance to this centre. The next centre is found also as the data point with the highest potential. The procedure is repeated until the potential of all data points is reduced below a certain threshold.

The procedure of the subtractive clustering includes the following steps (Chiu, 1994):

1. Initially, the data point with the highest potential is chosen to be the first cluster centre:

$$P_1^* = \max_{i=1}^{TD} P_i$$

where P_1^* denotes the potential of the first centre.

2. The potential of all other points are then reduced by an amount proportional to the potential of the chosen point and inversely proportional to the distance to this centre:

$$P_i = P_i - P_k^* e^{-\beta\left\|z_i - z_k\right\|^2} \; ; \; i = 1,...,N$$

where P_k^* denotes the potential of the *k*th centre; $k = 1,...,TD$. $\beta = \frac{4}{r_b^2}$, where r_b is a positive constant, determining the radius of the neighbourhood that will have measurable reductions in the potential because of the closeness to an existing centre. Recommended value of r_b is $r_b = 1.5 r_a$.

3. Two boundary conditions are defined: lower $\left(\underline{\varepsilon} * P^{ref}\right)$ and upper $\left(\overline{\varepsilon} * P^{ref}\right)$ thresholds, determined as a function of the maximal potential called reference potential $\left(P^{ref}\right)$. The values for $\underline{\varepsilon}$ and $\overline{\varepsilon}$ are respectively 0.15 and 0.5. A data point is chosen to be a new cluster centre, and respectively centre of a rule, if its potential is higher than the upper threshold. If its potential is minor than the lower threshold it will be definitely rejected.

4. If the potential of a point lies between the two boundaries, the shortest of the distances $\left(\delta_{\min}\right)$ between the new candidate to be a cluster centre $\left(z_k^*\right)$ and all previously found cluster centres is decisive. The following inequality expresses the trade-off between the potential value and the closeness to the previous centres:

$$\frac{\delta_{\min}}{r} + \frac{P_k^*}{P_1^*} \geq 1$$

This approach has been used for initial estimation of the antecedent parameters in fuzzy identification. It relies on the idea that each cluster centre is representative of a characteristic behaviour of the system. The resulting cluster centres are used as parameters of the antecedent parts defining the focal points of the rules of the model.

2.2 Learning Rule Consequents

For fixed antecedent parameters the second subtask, estimation of the parameters of the consequent linear models, can be transformed into a least squared problem. This is accomplished by eliminating the summation operation in the TS model output and replacing it with an equivalent vector expression of *y*:

$$y = \psi^T \theta$$

where $\theta = \left[\pi_1^T, \pi_2^T, \ldots, \pi_R^T\right]^T$ is a vector composed of the linear model parameters; $\psi = \left[\lambda_1 x_e^T, \lambda_2 x_e^T, \ldots, \lambda_R x_e^T\right]^T$ is a vector of the inputs that are weighted by the normalized firing levels of the rules.

For a given set of input-output data $\left(x_k^T, y_k\right)$, $k = 1, \ldots, TD$, the vector of linear parameters θ minimizing the objective function is:

$$J_G = \sum_{k=1}^{TD} \left(y_k - \psi_k^T \theta\right)^2$$

where $\psi_k = \left[\lambda_1(x_k)x_{ek}^T, \lambda_2(x_k)x_{ek}^T, \ldots, \lambda_R(x_k)x_{ek}^T\right]^T$ and $x_{ek} = \left[1, \quad x_k^T\right]^T$ can be estimated by the Recursive Least Squares (RLS) algorithm:

$$\hat{\theta}_k = \hat{\theta}_{k-1} + C_k\psi_k\left(y_k - \psi_k^T\hat{\theta}_{k-1}\right)$$

$$C_k = C_{k-1} - \frac{C_{k-1}\psi_k\psi_k^T C_{k-1}}{1+\psi_k^T C_{k-1}\psi_k};\ k = 1,\ldots,TD$$

with initial conditions and $C_0 = \Omega I$, where Ω is a large positive number; C is a $R(n+1)\times R(n+1)$ co-variance matrix; $\hat{\theta}_k$ is an estimation of the parameters based on *k* data samples.

3 On-Line Learning of Takagi-Sugeno Fuzzy Models

In on-line mode the training data is collected continuously, rather than being a fixed set. Some of the new data reinforce and confirm the information contained in the previous data. Other data, however, bring new information, which could indicate a change in the operating conditions, development of a fault or simply a more significant change in the dynamic of the process. They may posses enough new information to form a new rule or to modify an existing one (Angelov and Filev, 2003).

Real-time on-line applications are hampered by the need to recursive calculation of the model parameters. On-line learning of evolving TS fuzzy models includes on-line clustering under assumption of a gradual change of the rule-base and modified (weighted) recursive least squares.

3.1 On-Line Clustering Algorithm

The on-line clustering procedure starts with the first data point established as the focal point of the first cluster. Its coordinates are used to form the antecedent part of the fuzzy rule using Gaussian membership functions. Its potential is assumed equal to 1.

Starting from the next data point onwards the potential of the new data points is calculated recursively using a Cauchy type function of first order:

$$P_k\left(z_k\right) = \frac{1}{1+\frac{1}{(k-1)}\sum_{i=1}^{k-1}\sum_{j=1}^{n+1}\left(d_{ik}^j\right)^2};\ k = 2,3,\ldots$$

where $P_k(z_k)$ denotes the potential of the data point (z_k) calculated at time *k*; $d_{ik}^j = z_i^j - z_k^j$, denotes the projection of the distance between two data points, z_i^j and z_k^j, on the axis z^j (x^j for $j = 1,2,\ldots,n$ and on the axis *y* for $j = n+1$).

This function is monotonic and inversely proportional to the distance and enables recursive calculation, which is important for on-line implementation of the learning algorithm (Angelov and Filev, 2003).

Opposing to subtractive clustering there is not a specific amount subtracted from the highest potential, but update of all the potentials after a new data point is available on-line. The potential of the new data sample is recursively calculated as follows (Angelov and Filev, 2003):

$$P_k(z_k) = \frac{k-1}{(k-1)(\vartheta_k + 1) + \sigma_k - 2v_k}$$

where

$$\vartheta_k = \sum_{j=1}^{n+1} \left(z_k^j\right)^2 \; ; \; \sigma_k = \sum_{l=1}^{k-1} \sum_{j=1}^{n+1} \left(z_l^j\right)^2 \; ; \; v_k = \sum_{j=1}^{n+1} z_k^j \beta_k^j \; ; \; \beta_k^j = \sum_{l=1}^{k-1} z_l^j$$

Parameters ϑ_k and v_k are calculated from the current data point z_k, while β_k^j and σ_k are recursively updated as follows (Angelov and Filev, 2003):

$$\sigma_k = \sigma_{k-1} + \sum_{j=1}^{n+1} \left(z_{k-1}^j\right)^2 \; ; \; \beta_k^j = \beta_{k-1}^j + z_{k-1}^j$$

After the new data are available in on-line mode, they influence the potentials of the centres of the clusters $\left(z_l^*, l = 1,\ldots,R\right)$, which are respective to the focal points of the existing rules $\left(x_l^*, l = 1,\ldots,R\right)$. The reason is that by definition the potential depends on the distance to all data points, including the new ones.

The recursive formula for update the potential of the focal points of the existing clusters is derived from the recursive formula used to calculate the potential (Angelov and Filev, 2003):

$$P_k\left(z_l^*\right) = \frac{(k-1)P_{k-1}\left(z_l^*\right)}{k - 2 + P_{k-1}\left(z_l^*\right) + P_{k-1}\left(z_l^*\right)\sum_{j=1}^{n+1}\left(d_{k(k-1)}^j\right)^2}$$

where $P_k\left(z_l^*\right)$ is the potential of the cluster at time *k*, which is a prototype of the *l*th rule.

Potential of the new data point is compared to the updated potential of the centres. If the potential of the new data point is higher than the potential of the existing centres then the new data point is accepted as a new centre and a new rule is formed with a focal point based on the projection of this centre on the axis x.

If $P_k(z_k) > P_k(z_i^*);\ i = 1,...,R$ then $(R = R+1; x_R^* = x_k)$

If in addition to the previous condition the new data point is close to an old centre then the new data point replaces this rule (Angelov and Filev, 2003).

If $P_k(z_k) > P_k(z_l^*)$ and $\frac{\left\| z_k - \arg\min_{l=1}^{R} \left\| z_k - z_l^* \right\| \right\|}{radii} + \frac{\max_{l=1}^{R} P_k(z_l^*)}{P_k(z_k)} < 1$ then $(z_j^* = z_k);\ l = 1,...,R$

If none of the conditions is true it means the new data point has no relevance in terms of creation or modification of rules.
It should be noted that using the potential instead of the distance to a certain rule centre only for forming the rule-base results in rules that are more informative and a more compact rule-base. The reason is that the spatial information and the history are not ignored, but are part of the decision whether to upgrade or modify the rule-base.
The on-line clustering approach proposed by Angelov ensures an evolving rule-base by dynamically upgrading and modifying it while inheriting the bulk of the rules (R-1 rules are preserved even when a modification or an upgrade take place).

3.2 On-Line Recursive Estimation of Consequence Parameters

The problem of increasing size of the training data is handled by RLS for the globally optimal case and by wRLS for the locally optimal case. They, however, are based on the assumption of a constant/unchanged rule base, i.e. fixed antecedent parameters. Under this assumption, the optimization problems are linear in parameters.
In evolving TS fuzzy models, however, the rule-base is assumed to be gradually evolving. Therefore, the number of rules as well as the parameters of the antecedent part will vary, though the changes are normally significantly rarer than the time step. Because of this evolution, the normalized firing strengths of the rules (λ_i) will change. Since this affects all the data (including the data collected before time of the change) the straightforward application of the RLS or wRLS is not correct. A proper resetting of the covariance matrices and parameters initialization of the RLS is needed at each time a rule is added to and/or removed from the rule base (Angelov and Filev, 2003).

Angelov proposed to estimate the covariance matrices and parameters of the new $(R+1)^{th}$ rule as a weighted average of the respective covariance and parameters of the remaining *R* rules (Angelov and Filev, 2003). This is possible since the approach of rule-base innovation considered concerns one rule only; the other *R* rules of the rule base remain unchanged.

3.2.1 Global Parameter Estimation

The evolving TS model is used for on-line prediction of the output based on the past inputs:

$$\hat{y}_{k+1} = \psi_k^T \hat{\theta}_k ;\ k = 2,3,...$$

The following RLS procedure is applied:

$$\hat{\theta}_k = \hat{\theta}_{k-1} + C_k \psi_k \left(y_k - \psi_k^T \hat{\theta}_{k-1} \right),\ k = 2,3,...$$

$$C_k = C_{k-1} - \frac{C_{k-1}\psi_k \psi_k^T C_{k-1}}{1 + \psi_k^T C_{k-1} \psi_k}$$

with initial conditions

$$\hat{\theta}_1 = \left[\hat{\pi}_1^T, \hat{\pi}_2^T, \ldots, \hat{\pi}_R^T \right]^T = 0 ;\ C_1 = \Omega I$$

When a new rule is added to the rule-base, the RLS is reseted in the following way:

i) Parameters of the new rule are determined by weighted average of the parameters of the other rules. The weights are the normalized firing levels of the existing rules λ_i. The idea is to use the existing centers as a rule-base to approximate the initialization of the parameters of the new rule by a weighted sum. Parameters of the other rules are inherited from the previous step:

$$\hat{\theta}_k = \left[\hat{\pi}_{1(k-1)}^T, \hat{\pi}_{2(k-1)}^T, \ldots, \hat{\pi}_{R(k-1)}^T, \hat{\pi}_{R+1(k)}^T \right]^T$$

where

$$\hat{\pi}_{R+1(k)} = \sum_{i=1}^{R} \lambda_i \hat{\pi}_{i(k-1)}$$

ii) Co-variance matrices are reseted as:

$$C_k = \begin{bmatrix} \rho\zeta_{11} & \cdots & \rho\zeta_{1R(n+1)} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho\zeta_{R(n+1)1} & \cdots & \rho\zeta_{R(n+1)R(n+1)} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \Omega & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & \Omega \end{bmatrix}$$

where ζ_{ij} is an element of the co-variance matrix $\left(i = [1, R \times (n+1)]; j = [1, R \times (n+1)]\right)$; $\rho = \dfrac{R^2 + 1}{R^2}$ is a coefficient.

In this way, the part of the co-variance matrix associated with the new $(R+1)^{th}$ rule (last n + 1 columns and last n + 1 rows) is initialized as usual (with a large number, Ω) in its main diagonal and co-variance matrices respective for the rest of the rules (from 1 to *R*) are updated by multiplication of ρ. The rationale for this is that the correction of the co-variance matrices needs, to approximate the role the new, $(R+1)^{th}$ rule would have if it was in the rule-base from the beginning, can be presented by ρ (Angelov and Filev, 2003).

When a rule is replaced by another one, which has antecedent parameters close to the rule being replaced, then parameters and co-variance matrices are inherited from the previous time step (Angelov and Filev, 2003).

3.2.2 Local Parameter Estimation

The local parameter estimation is based on the wRLS:

$$\hat{\pi}_{ik} = \hat{\pi}_{ik-1} + c_{ik} x_{ek} \lambda_i(x_k)\left(y_k - x_{ek}^T \hat{\pi}_{ik-1}\right),\ k = 2,3,...$$

$$c_{ik} = c_{ik-1} - \frac{\lambda_i(x_k) c_{ik-1} x_{ek} x_{ek}^T c_{ik-1}}{1 + \lambda_i(x_k) x_{ek}^T c_{ik-1} x_{ek}};\ i = 1,...,R$$

with initial conditions $\hat{\pi}_1 = 0$; $c_{i1} = \Omega I$

In this case, the co-variance matrices are separated for each rule and have smaller dimensions $\left(c_{ik} \in R^{(n+1)\times(n+1)}; i = 1,...,R\right)$. Parameters of the newly added rule are determined as weighted average of the parameters of the rest *R* rules by:

$$\hat{\pi}_{R+1(k)} = \sum_{i=1}^{R} \lambda_i \hat{\pi}_{i(k-1)}$$

Parameters of the other *R* rules are inherited $\left(\pi_{ik} = \pi_{i(k-1)}; i = 1,..., R\right)$.

The co-variance matrix of the newly added rule is initialized by:

$$C_{R+1(k)} = \Omega I$$

The co-variance matrices of the rest *R* rules are inherited $\left(c_{ik} = c_{i(k-1)}; i = 1,..., R\right)$.

3.4 The Procedure for Rule-Base Evolution in TS Fuzzy Models

The recursive procedure for on-line learning of evolving TS models includes the following stages (Angelov and Filev, 2003):

Stage 1: Initialization of the rule-base structure (antecedent part of the rules);
Stage 2: At the next time step reading the next data sample;
Stage 3: Recursive calculation of the potential of each new data sample to influence the structure of the rule-base;
Stage 4: Recursive up-date of the potentials of old centres taking into account the influence of the new data sample;
Stage 5: Possible modification or up-grade of the rule-base structure based on the potential of the new data sample in comparison to the potential of the existing rule centres (focal points);
Stage 6: Recursive calculation of the consequent parameters;
Stage 7: Prediction of the model output for the next time step.

The execution of the algorithm continues for the next time step from stage 2, Figure 1. It should be noted that the first output to be predicted is $\hat{y}_3$. In the following is briefly reminded the essential of each stage.

Stage 1: Initialization of the rule-base structure

The rule base can contain only one single rule, based, for example, on the first data sample.

$$k = 1;\ R = 1;\ x_1^* = x_k;\ P_1\left(z_1^*\right) = 1$$

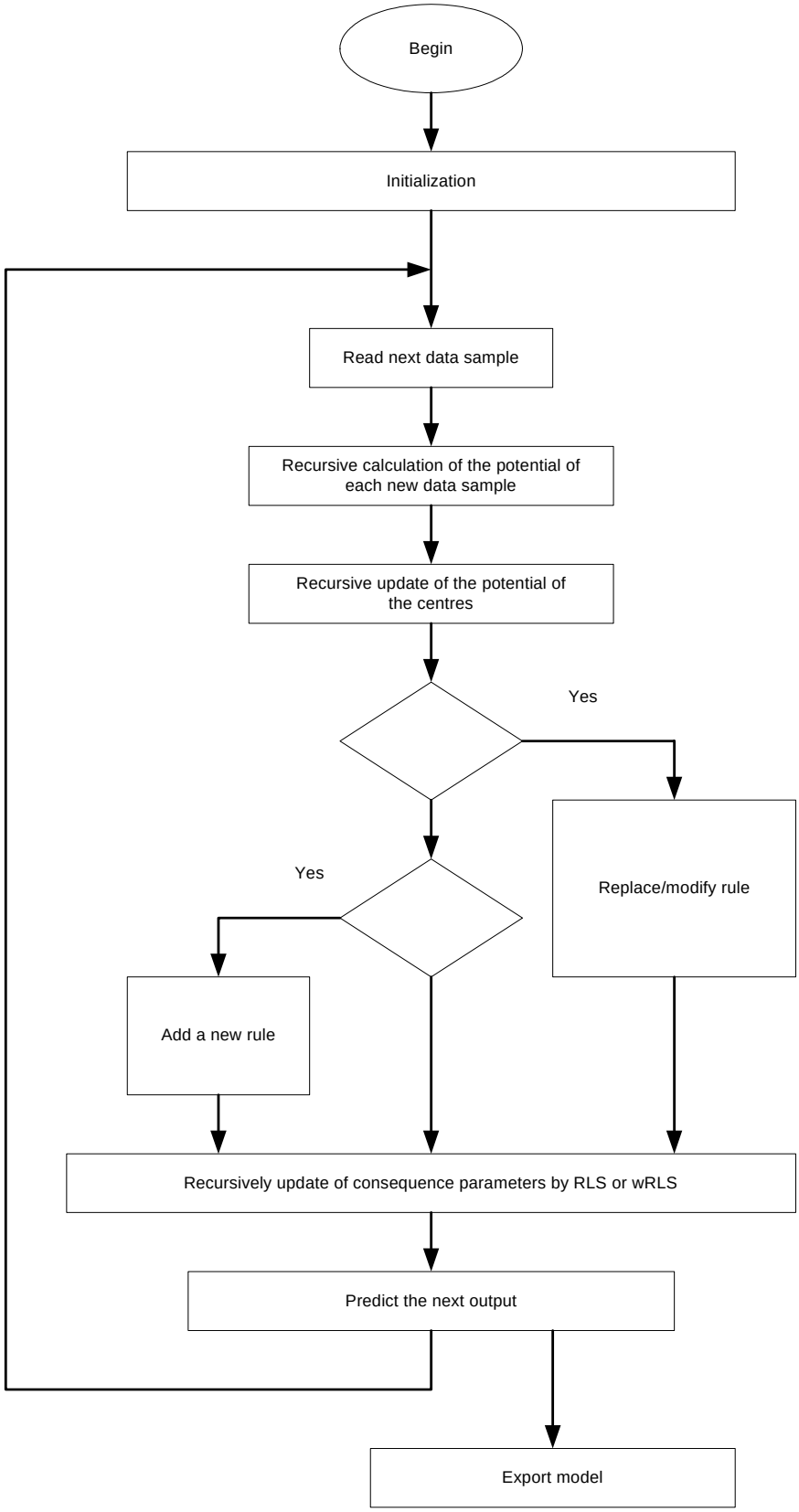

Figure 1: Recursive procedure for on-line learning of evolving TS fuzzy models (Angelov and Filev, 2003).

$\theta_1 = \pi_1 = 0$; $C_1 = \Omega I$

where z_1^* is the first cluster centre; x_1^* is a focal point of the first rule being a projection of z_1^* on the axis x.

The rule-base can be initialised by existent expert knowledge or based on off-line identification approaches. In this case:

$R = R^{ini}$; $P_1(z_i^*) = 1$; $i = 1,...,R^{ini}$

where R^{ini} denotes the number of rules defined initially off-line.

Stage 2: Reading the next data sample

At the next time step $(k = k+1)$ the new data sample (z_k) is collected.

Stage 3: Recursive calculation of the potential of each new data sample

The potential of each new data sample is recursively calculated. The use of already calculated values for σ_k and β_k^j leads to significant time and calculation savings. At the same time, they have accumulated information regarding the spatial proximity to all previous data.

Stage 4: Recursive up-date of the potential of old centres

The potentials of the focal points (centres) of the existing clusters/rules are recursively updated.

Stage 5: Possible modification or up-grade of the rule-base structure

The potential of the new data sample is compared to the updated potential of existing centres and a decision whether to modify or upgrade the rule-base is taken.

Stage 6: Recursive calculation of the consequent parameters

Parameters of the consequence are recursively updated by RLS, for globally optimal parameters, or by wRLS, for locally optimal parameters. In the first case the cost function is minimized, which guarantees globally optimal values of the parameters, while in the second case the locally weighted cost function is minimized and locally meaningful parameters are obtained.

Stage 7: Prediction of the output

The output for the next time step, $(k+1)$, is calculated by:

$$\hat{y}_{k+1} = \psi_k^T \hat{\theta}_k ;\ k = 2,3,...$$

The algorithm continues from stage 2 by reading the next data sample at the next time step.

Using the approach proposed by Angelov a transparent, compact and accurate model can be found by rule base evolution based on experimental data with the simultaneous recursive estimation of the fuzzy set parameters.
It is interesting to note that the rate of upgrade with new rules does not lead to an excessively large rule base. The reason for this is that the condition for the new data point to have higher potential than the focal points of all existing rules is a hard requirement. Additionally, the possible proximity of a candidate centre to the already existing focal points leads to just a replacement of the existing point, i.e. modification of its coordinates without enlarging the rule-base size (Angelov and Filev, 2003).

4. Experimental Results

The approach proposed by Angelov is tested on a benchmark problem: the Mackey-Glass chaotic time series prediction. The data set has been used as a benchmark example in the areas of fuzzy systems, neural networks and hybrid systems. The chaotic time series is generated from the Mackey-Glass differential delay equation defined by:

$$\dot{x}(t) = \frac{0.2x(t-\tau)}{1+x^{10}(t-\tau)} - 0.1x(t)$$

The aim is using the past values of *x* to predict some future value of *x*. We assume $x(0) = 1.2$, $\tau = 17$ and the value of the signal 85 steps ahead is predicted, $x(t+85)$, based on the values of the signal at the current moment, 6, 12 and 18 steps back, $[x(t-18), x(t-12), x(t-6), x(t)]$.

The following experiment was conducted: 3000 data points, for $t = 201:3200$, are extracted from the time series and used as training data, Figure 2; 500 data points, for $t = 5001:5500$, are used as testing (validation) data, Figure 3.

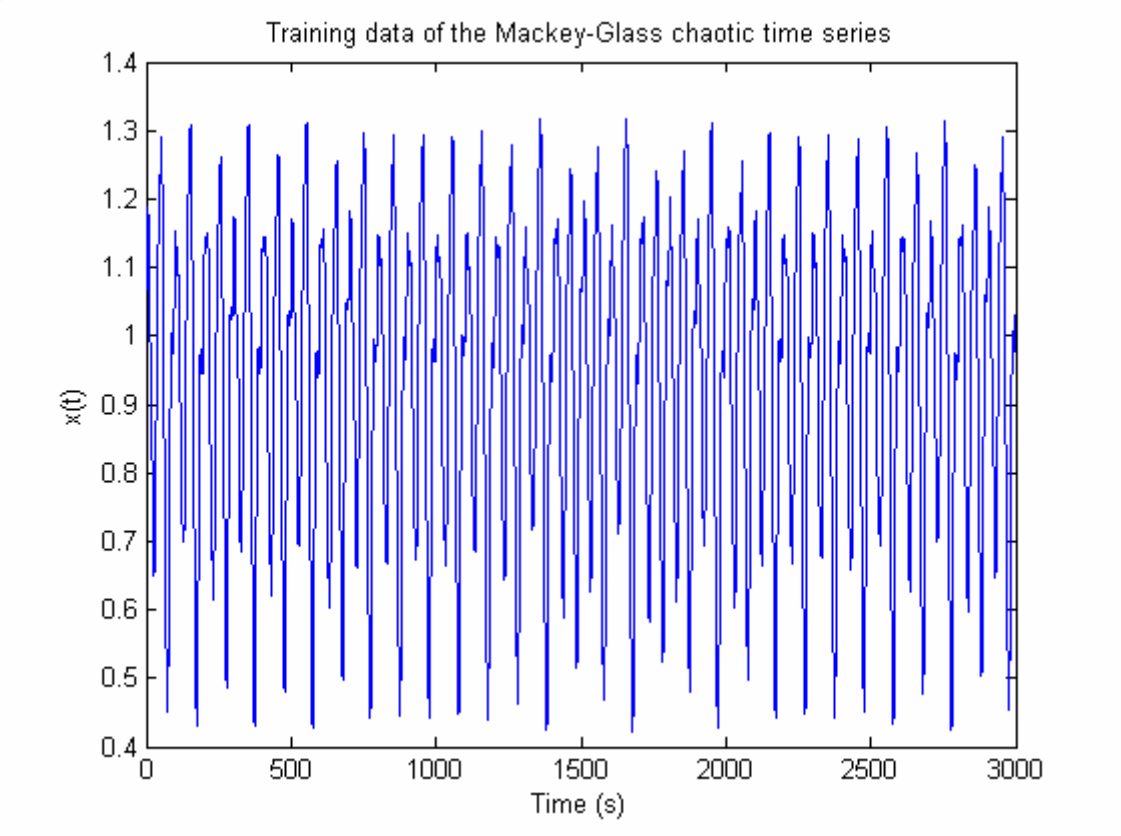

Figure 2: Training data.

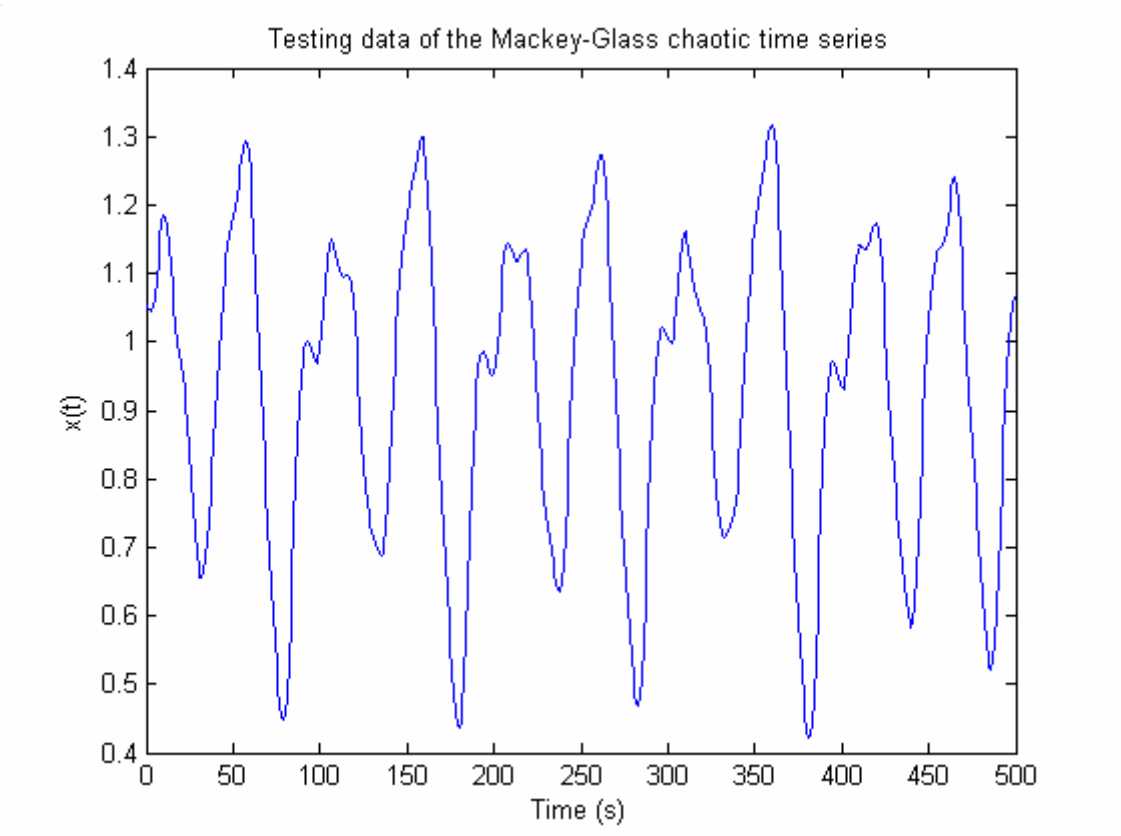

Figure 3: Testing (validation) data.

All the data is putted in the same file, i.e. training data plus testing data. The data set has 3500 data samples and the learning mechanism is always active, even for the testing data.
To evaluate the performance of the models we use the RMSE and the NDEI (Non-Dimensional Error Index), defined as the ratio of the root mean square error over the standard deviation of the target data.

$$NDEI = \frac{RMSE}{std(y(t))}$$

The values of the performance measures will be calculated separately for the training and testing data.

4.1 Parameters Influence

In this algorithm it is necessary to specify the following parameters:

- Radii (r_a)
- Omega (Ω)
- Ro (ρ)

The first parameter, radii, has a strong influence in the structure of the model since it directly affects the number of rules and consequently the performance and complexity of the models. The experiments confirm that in general as the constant radii increases the number of rules created decreases. Too large values of radii lead to averaging and too small values lead to over-fitting. Values between 0.3 and 0.5 can be recommended. In the studies we carry out smaller and larger values of radii were considered in order to better understand its influence in the rule creation/modification process.
Parameter Omega has influence on the estimation of the consequence parameters. A small value for Ω means that we have some confidence in the initialization parameters of the new rule consequents. A bigger value expresses less confidence in the initialization and inherently it is given a better adaptation capability to the method. In all the scenarios studied the value for parameter Ω is 750.
Parameter Ro also influences the estimation of the consequence parameters. When a new rule is added the existent elements of the co-variance matrix are updated by multiplication of ρ in order to allow the method to make the necessary adjustments. The goal of this action is quite the same when we use recursive estimation with forgetting factor, except that in this case the elements of the co-variance matrix are

updated only when a new rule id added. In all the scenarios studied the coefficient ρ is given by $\rho = \frac{R^2 + 1}{R^2}$, where *R* is the number of rules (Angelov and Filev, 2003).

4.2 Conditions for rule innovation/modification

The conditions for creation of rules and modification or up-date of rules have a strong influence in the algorithm behaviour and performance. There are several conditions that can be used to create or modify rules and it is not easy to adjust the definitions for a specific problem. The generic conditions proposed by Angelov are the following:

If < Condition 1 > and < Condition 2 > **Then** *modify rule*

Else If < Condition 3 > **Then** *create new rule*

Condition 1: the potential of the new data point is higher than some value
Condition 2: the new data point is close to an old centre
Condition 3: most of the times the same as Condition 1

We will consider several scenarios where different conditions will be tested. Scenarios A, B and C, as we will see, were proposed by Angelov while scenarios D, E, F and G are new. Scenarios A and B are presented only to better understand the subsequent scenarios since these are only valid for previous versions of the approach, the off-line version and the version with a moving window (Angelov, 2002).

In this study precision and generalization capabilities of the models obtained are important but it is also important the complexity and transparency of the models. The first aspect is measured by the RMSE and NDEI for training and testing data. The second one is measured by the number of rules and the location of the membership functions.

Scenario A (Angelov and Filev, 2003).

Condition 1: the potential of the new data point is higher than certain threshold $\underline{\varepsilon}$, i.e. $P_k(z_k) > \underline{\varepsilon}$;

Condition 2: the new data point is close to an old centre, given by $\frac{d_{\min}}{radii} + \frac{P(z_i^*)}{P(z_k)} < 1$;

Condition 3: the potential of the new data point is higher than certain threshold $\overline{\varepsilon}$, i.e. $P_k(z_k) > \overline{\varepsilon}$.

In a more formal way we have:

If $P_k(z_k) > \underline{\varepsilon}$ and $\frac{d_{\min}}{radii} + \frac{P(z_i^*)}{P(z_k)} < 1$ **Then** $z_j^* = z_k$
Else If $P_k(z_k) > \overline{\varepsilon}$ **Then** $R = R+1; z_R^* = z_k$

The values of the thresholds are determined as:

$$\underline{\varepsilon} = 0.15 * \max_{i=1}^{R} P_i^* \; ; \; \overline{\varepsilon} = 0.5 * \max_{i=1}^{R} P_i^*$$

The upper threshold $\overline{\varepsilon}$ ensures that only data samples with potential above half of the best are accepted. Its value influences the number of rules present in the final model. The value of 50% of the maximal potential showed a good balance between the model complexity and precision on a number of examples.
The lower threshold $\underline{\varepsilon}$ defines a grey zone where the data points are checked on the proximity to the existing centres. It influences the number of replacements of centres, which will take place. Alternatively, both thresholds could be equal to the mean potential of all rules:

$$\underline{\varepsilon} = \overline{\varepsilon} = \frac{1}{R}\sum_{i=1}^{R} P_i^*$$

It must be stated again that this scenario can only be used off-line.

Scenario B

Condition 1: the potential of the new data point is higher than the mean potential of the existing centres, i.e. $P_k(z_k) > meanP_k(z_i^*)$;

Condition 2: the new data point is close to an old centre, given by $\frac{d_{\min}}{radii} + \frac{P(z_i^*)}{P(z_k)} < 1$;

Condition 3: same as Condition 1.

In a more formal way we have:

If $P_k(z_k) > meanP_k(z_i^*)$ and $\frac{d_{\min}}{radii} + \frac{P(z_i^*)}{P(z_k)} < 1$ **Then** $z_j^* = z_k$
Else If $P_k(z_k) > meanP_k(z_i^*)$ **Then** $R = R+1; z_R^* = z_k$

Condition 1 is very flexible which means the number of rules will increase rapidly. This could make sense for some problems but most of the times we want a rule base with the minimum number of rules possible.

Scenario C

Condition 1: the potential of the new data point is higher than the potential of the existing centres, i.e. $P_k(z_k) > P_k(z_i^*)$;

Condition 2: the new data point is close to an old centre, given by $\frac{d_{\min}}{radii} + \frac{P(z_i^*)}{P(z_k)} < 1$;

Condition 3: same as Condition 1.

In a more formal way we have:

If $P_k(z_k) > P_k(z_i^*)$ and $\frac{d_{\min}}{radii} + \frac{P(z_i^*)}{P(z_k)} < 1$ **Then** $z_j^* = z_k$
Else If $P_k(z_k) > P_k(z_i^*)$ **Then** $R = R + 1; z_R^* = z_k$

We conduct a study for different values of radii and Table 1 summarizes some of the information obtained.

Radii	New rules	Modified rules	RMSE Training	RMSE Validation	NDEI Training	NDEI Validation
0.1	19	0	0.08755	0.08894	0.38667	0.39262
0.2	19	0	0.08688	0.08779	0.38372	0.38754
0.3	19	0	0.08622	0.08676	0.38081	0.38299
0.4	19	0	0.08547	0.08670	0.37749	0.38272
0.5	19	0	0.08508	0.08760	0.37580	0.38668
0.6	19	0	0.08512	0.08877	0.37593	0.39187
0.7	19	0	0.08530	0.08979	0.37675	0.39635

Table 1: Results from scenario C.

The behaviour of the algorithm is not the expected, since the number of rules is always the same for different values of radii. The same happens for the number of modified rules, which is zero in all the situations. The reason for this is because Condition 2 is practically impossible and therefore no data point that verifies Condition 1 verifies also Condition 2.

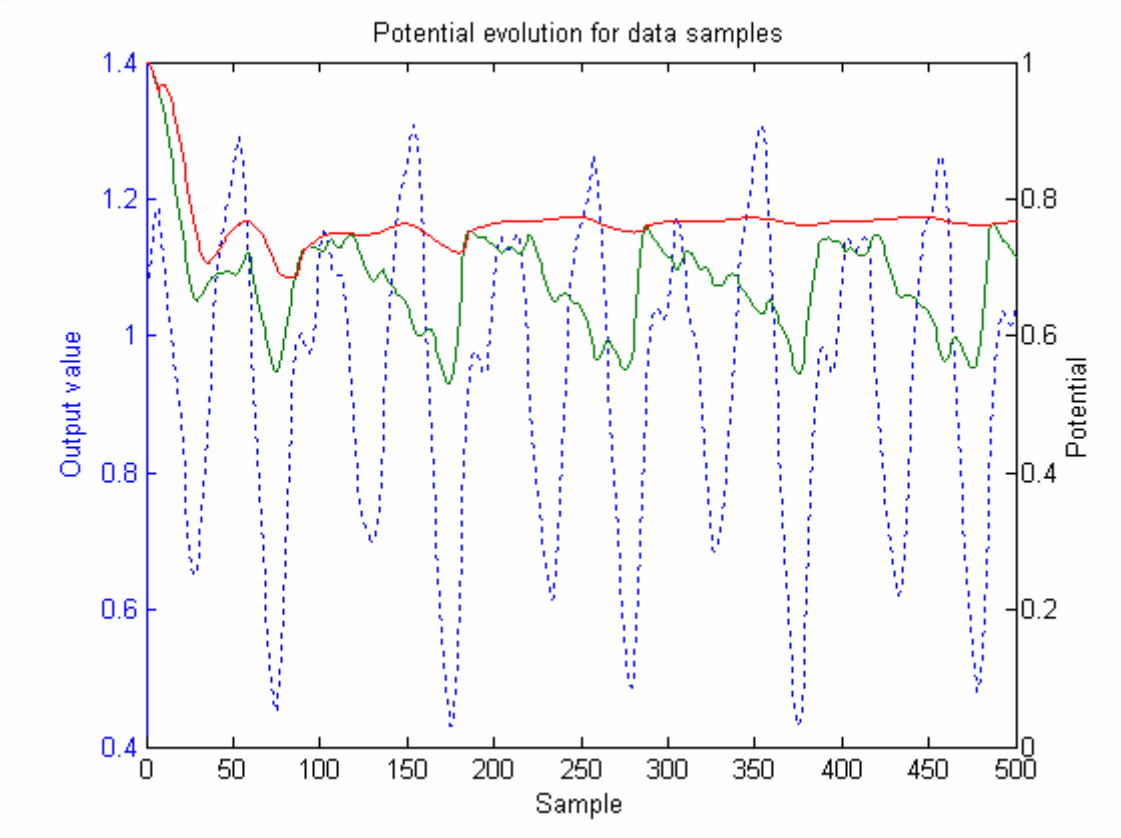

Figure 4: Evolution of the potential for the first 500 samples.

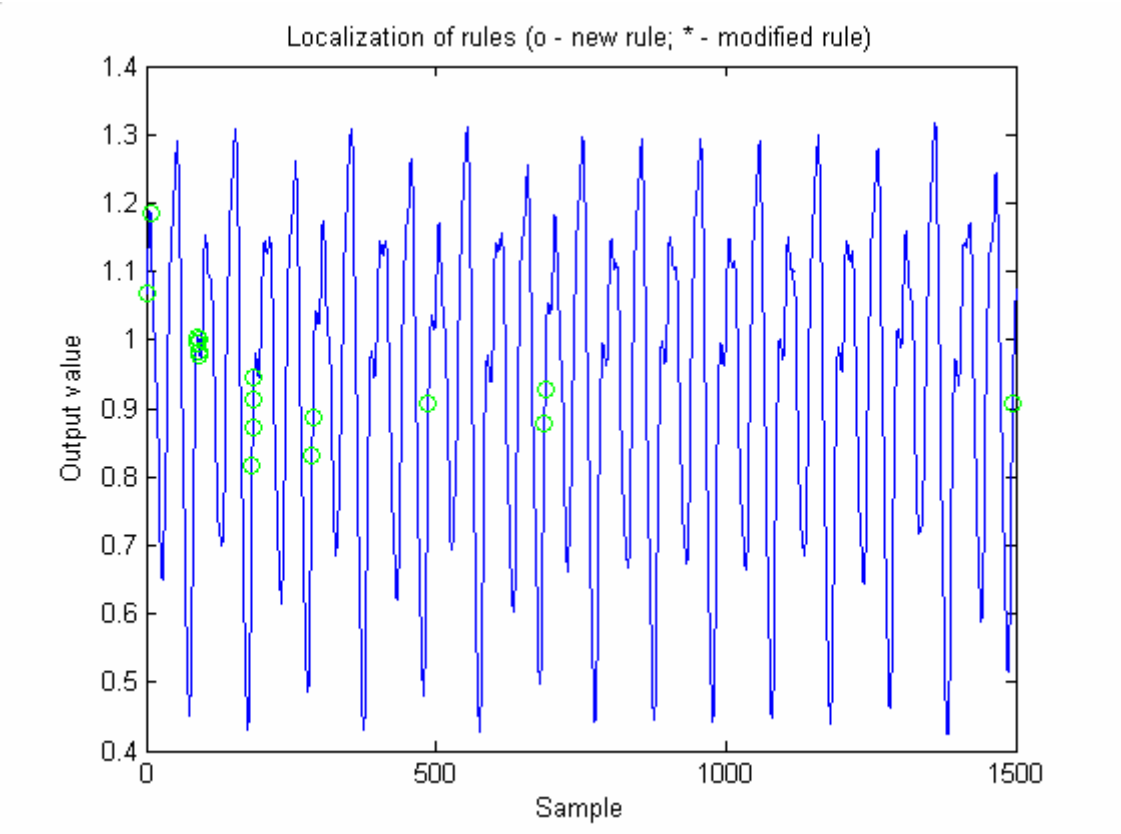

Figure 5: Rule base evolution.

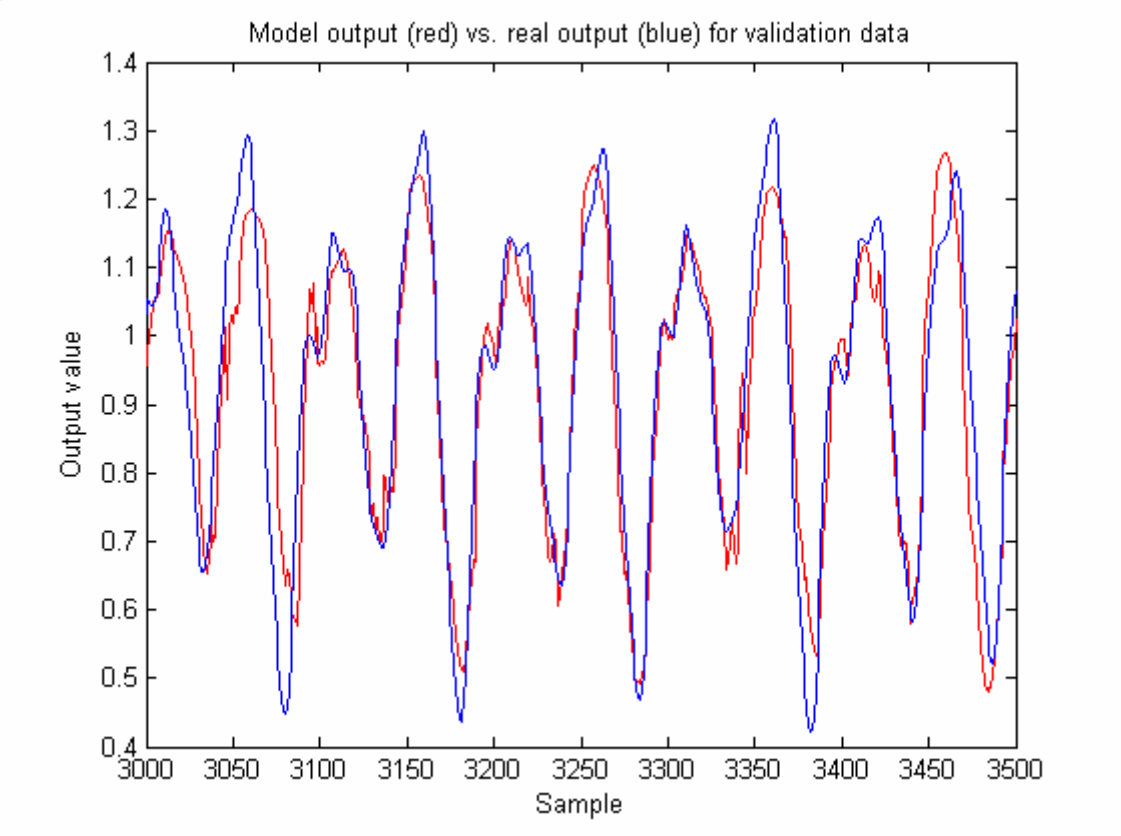

Figure 6: Model output for the testing data.

In Figure 4, 5 and 6 is presented more detailed information for a particular value of the parameter radii, radii = 0.4. Figure 4 presents the evolution of the potential, recursively calculated, for the first 500 samples and Figure 5 the samples that give origin to new rules.

Figure 6 shows the model output for the testing data and the real output of the time series. The model output follows most of the time the real output but for the regions of extrema the behaviour of the model is not so good.

In the following are described the new scenarios proposed in this study.

Scenario D

Condition 1: the potential of the new data point is higher than the potential of the existing centres, i.e. $P_k(z_k) > P_k(z_i^*)$;

Condition 2: the new data point is close to an old centre, given by $\frac{d_{\min}}{radii} < 0.5$;

Condition 3: same as Condition 1.

In a more formal way we have:

If $P_k(z_k) > P_k(z_i^*)$ and $\frac{d_{\min}}{radii} < 0.5$ **Then** $z_j^* = z_k$
Else If $P_k(z_k) > P_k(z_i^*)$ **Then** $R = R+1; z_R^* = z_k$

In the following it is explained the deduction of Condition 2, $\frac{d_{\min}}{radii} < 0.5$.

Let's consider the condition to create a new centre in the subtractive clustering algorithm when the potential of the data point lies between the two boundaries defined by the lower and upper thresholds, i.e.:

$\underline{\varepsilon} P_1^* < P_k^* < \bar{\varepsilon} P_1^*$:

where P_1^* is the potential of data point x_1^*, the one with the biggest potential, and P_k^* is the potential of point x_k^* which can be a accepted as a centre if the condition below is true.

If $\frac{d_{\min}}{r_a} + \frac{P_k^*}{P_1^*} \geq 1$ then accept point as a cluster centre.

We know that $0.15 < \frac{P_k^*}{P_1^*} < 0.5$, since $\underline{\varepsilon} = 0.15$ and $\bar{\varepsilon} = 0.5$.

For data point x_k^* being possibly accepted as a centre the term $\frac{d_{\min}}{r_a}$ has to be greater than 0.5, i.e.:

$\frac{d_{\min}}{r_a} > 0.5$, or $d_{\min} > 0.5 * r_a$

The meaning of this condition is that the points that are under the influence of an existing centre can be selected as new centres. More precisely, the points for which the potential lies between the two boundaries and the distance to the existing cluster centres is superior to 0.5 the radius are possibly accepted as new centres. If $\frac{d_{\min}}{r_a} \geq 0.85$ then we can say for sure the point will be accepted as a new centre.

In the on-line clustering algorithm the condition proposed by Angelov for the modification of a rule is:

$$\frac{d_{\min}}{radii} + \frac{P(z_i^*)}{P(z_k)} < 1, \text{ with } P(z_k) > P(z_i^*)$$

This condition is false for almost all data points, if not for all, for which $P(z_k) > P(z_i^*)$ because the second term of the condition is very close to 1. This brings the need for the definition of a new condition to determine if a data point is close to an existing centre.

Following the same principle used in the subtractive clustering algorithm we eliminate the second term of the condition and obtain:

$$\frac{d_{\min}}{radii} < 1$$

However, this condition allows the modification of the clusters centres for any data point that is under the influence of a cluster. In practice all points that verify the condition $P(z_k) > P(z_i^*)$ will also verify the previous condition and consequently no new cluster centres are created.

To avoid this we can use a condition similar to the one used in subtractive clustering, i.e. the ratio between $d_{\min}$ and *radii* in this case should be lesser than 0.5, i.e.:

$$\frac{d_{\min}}{radii} < 0.5$$

or $d_{\min} < 0.5 * radii$

This condition allows a good compromise for the points that have a potential higher than the potential of the existing centres to originate new centres or modify the existing ones.

We conduct a study for different values of radii and Table 2 summarizes some of the information obtained.

In this scenario the algorithm presents a good behaviour since as radii increases the number of created rules decreases and the number of modified rules increases.

Radii	New rules	Modified rules	RMSE Training	RMSE Validation	NDEI Training	NDEI Validation
0.1	15	4	0.09011	0.08964	0.39798	0.39569
0.2	9	12	0.09610	0.09301	0.42444	0.41059
0.3	6	20	0.09847	0.09483	0.43492	0.41862
0.4	6	27	0.09622	0.09392	0.42497	0.41460
0.5	4	22	0.10562	0.10325	0.46648	0.45577
0.6	4	33	0.09970	0.10255	0.44036	0.45268
0.7	3	49	0.10586	0.10405	0.46752	0.45930

Table 2: Results from scenario D.

For the particular case of radii = 0.4 we present the evolution of the potential, the rule base evolution and the model output for the testing data set.

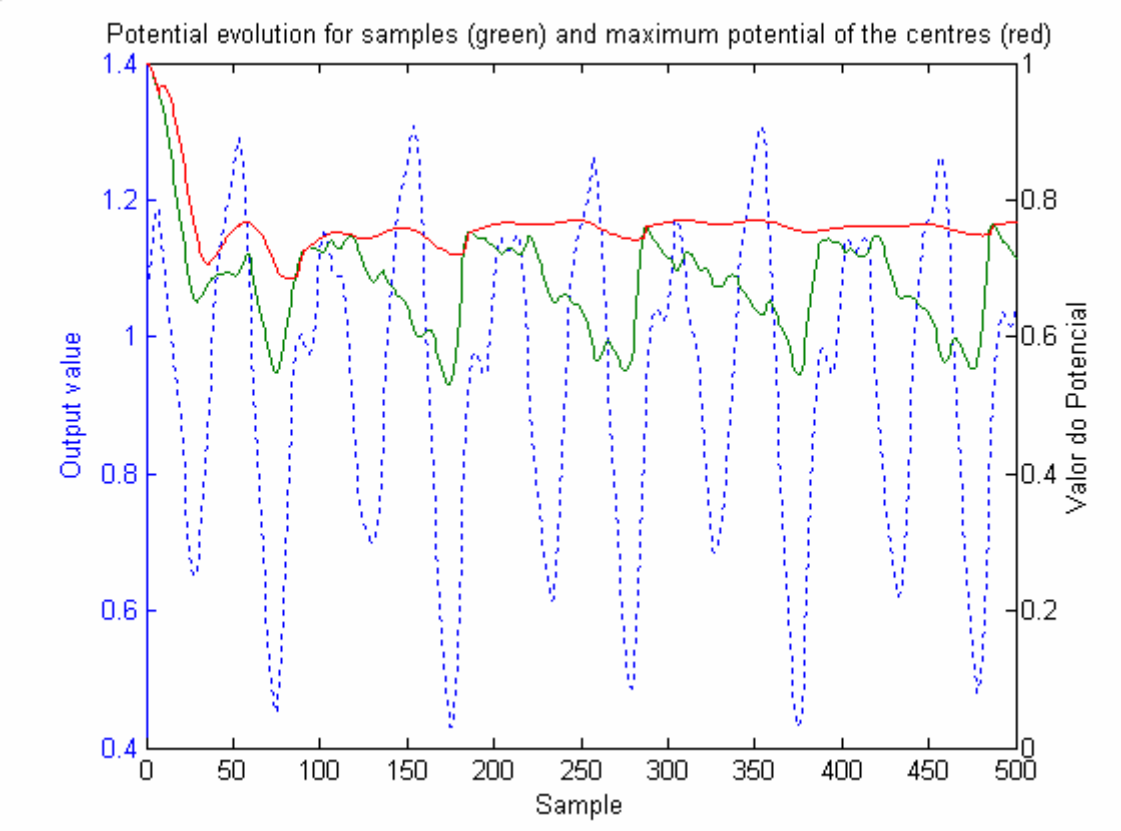

Figure 7: Evolution of the potential for the first 500 samples.

Figure 7 presents the evolution of the potential for each sample and the maximum potential for the existing centres. All these values are recursively calculated.
The analysis of the graphic shows that the condition $P_k(z_k) > P_k(z_i^*)$ is very strong since the number of data points that verify this condition is quite small.

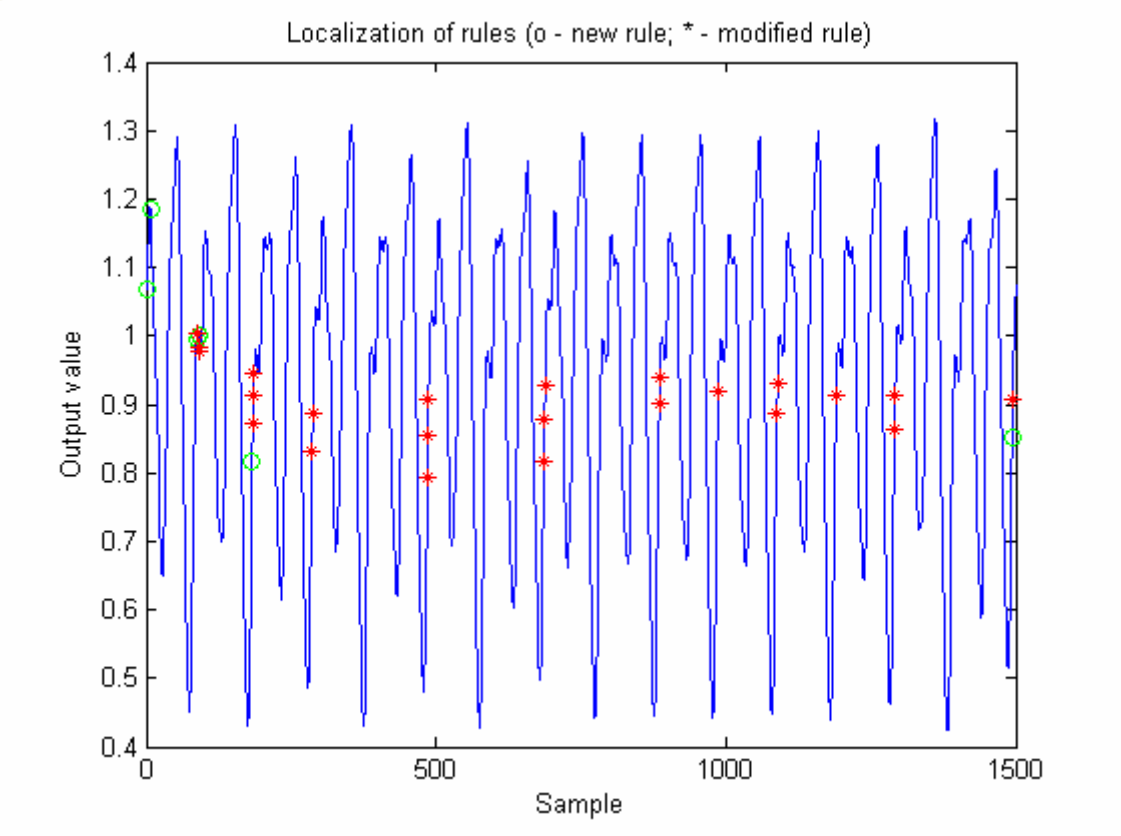

Figure 8: Rule base evolution.

Figure 8 shows for the output the instants when rules are created or modified, i.e. the rule base evolution. The samples that originate new rules were the following: 1, 7, 87, 90, 183 and 1494. Samples 88, 89, 92, 93, 184, 185, 186, 286, 287, 485, 486, 487, 686, 687, 688, 885, 886, 987, 1088, 1089, 1190, 1291, 1292, 1495, 2086, 2891 and 3495 correspond to situations where a rule is modified. With the conditions used in this scenario the rules are created in the beginning of the learning process, with the exception of rule 6 and after that only 4 rules are modified.

In Figure 9 is presented the model output for the testing data set. The model output shows that the performance is not quite good, particularly for the regions of extrema, which means the model has good prediction capabilities for some regions but for others the behaviour is poor.

This fact can also be confirmed by the location of the fuzzy sets of the input variables. Most of the fuzzy sets are concentrated only on some parts of the universe of discourse while some parts are not covered. In order to improve the model performance a better distribution of the fuzzy sets must be guaranteed, particularly for the regions that are not covered.

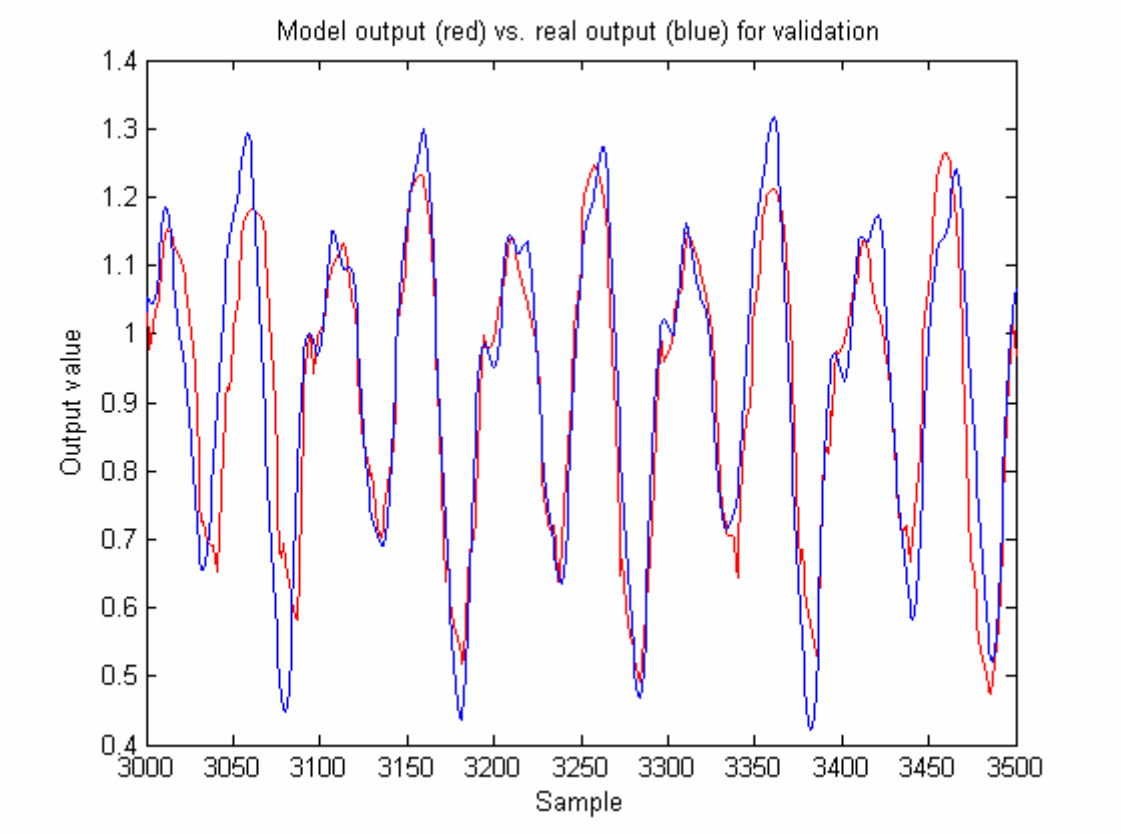

Figure 9: Model output for the testing data.

Scenario E

We decide to make some improvements and for that we pick up the thresholds idea of the subtractive clustering algorithm, also already used by Angelov, scenario B.

Condition 1: the potential of the new data point is higher than the potential of the existing centres, i.e. $P_k(z_k) > P_k(z_i^*)$, or between two thresholds $\underline{\varepsilon}$ and $\overline{\varepsilon}$

Condition 2: the new data point is close to an old centre, given by $\frac{d_{\min}}{radii} < 0.5$

Condition 3: same as Condition 1

In a more formal way we have:

If $\left\{P_k(z_k) > P_k(z_i^*) \quad or \quad \underline{\varepsilon} < P_k(z_k) < \overline{\varepsilon}\right\}$ and $\frac{d_{\min}}{radii} < 0.5$ **Then** $z_j^* = z_k$

Else If $\left\{P_k(z_k) > P_k(z_i^*) \quad or \quad \underline{\varepsilon} < P_k(z_k) < \overline{\varepsilon}\right\}$ **Then** $R = R + 1; z_R^* = z_k$

The values for the lower and upper threshold are $\underline{\varepsilon} = 0.50 * P_k^{Ref}$ and $\overline{\varepsilon} = 0.75 * P_k^{Ref}$.

Table 3 summarizes some of the information obtained for different values of radii.

Radii	**New rules**	**Modified rules**	**RMSE Training**	**RMSE Validation**	**NDEI Training**	**NDEI Validation**
0.1	80	266	0.07316	0.06762	0.32311	0.29851
0.2	28	320	0.07876	0.07673	0.34787	0.33873
0.3	19	313	0.08150	0.07948	0.35997	0.35086
0.4	15	305	0.08272	0.08024	0.36536	0.35422
0.5	9	322	0.08813	0.08684	0.38927	0.38336
0.6	9	294	0.09097	0.09090	0.40178	0.40126
0.7	8	286	0.09551	0.09330	0.42183	0.41186

Table 3: Results from scenario E.

For the particular case of radii = 0.4 we present the evolution of the potential, the rule base evolution and the model output for the testing data set.

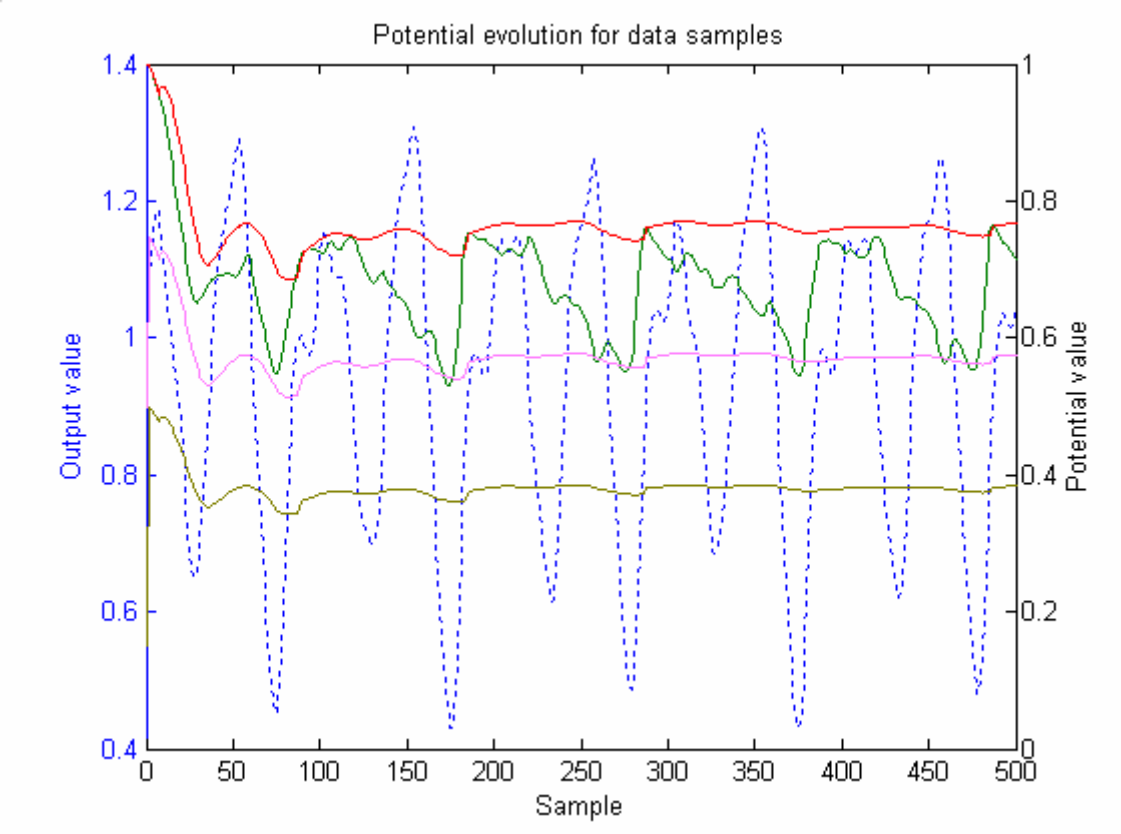

Figure 10: Evolution of the potential for the first 500 samples.

From the previous figure we conclude that the lower threshold has no influence in the process of creation and modification of rules.

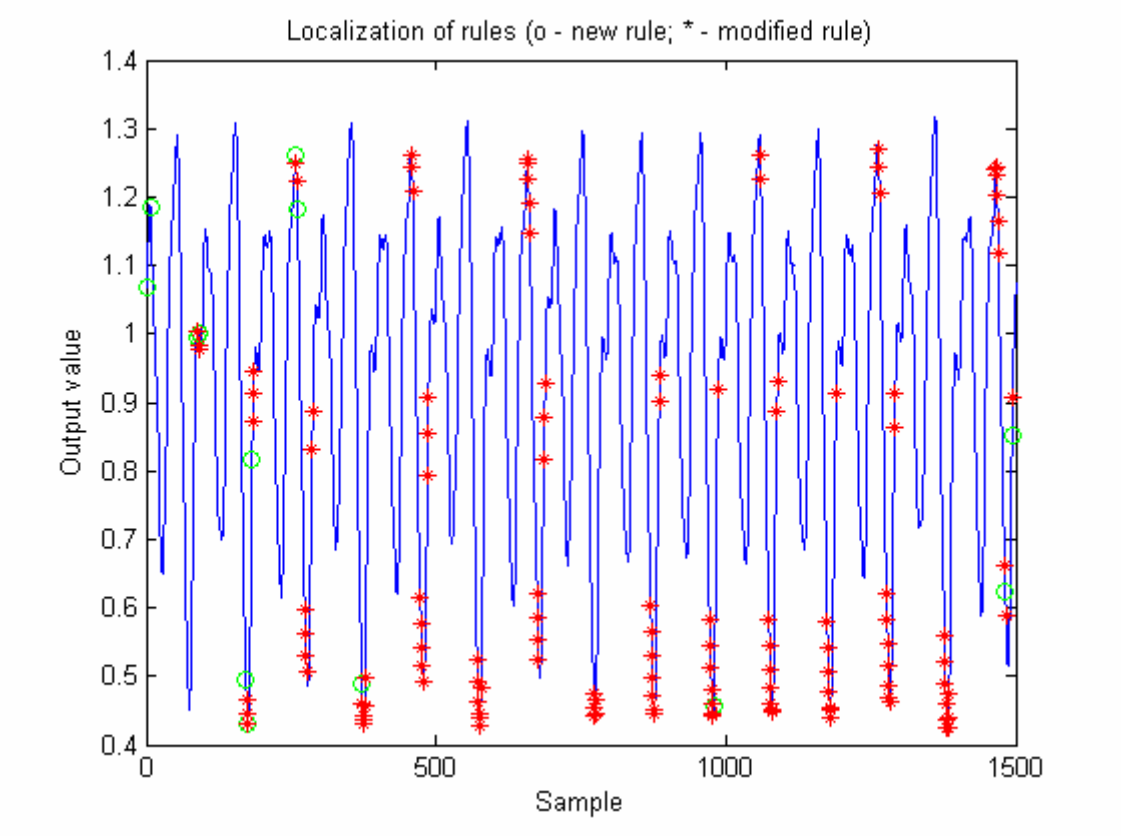

Figure 11: Rule base evolution.

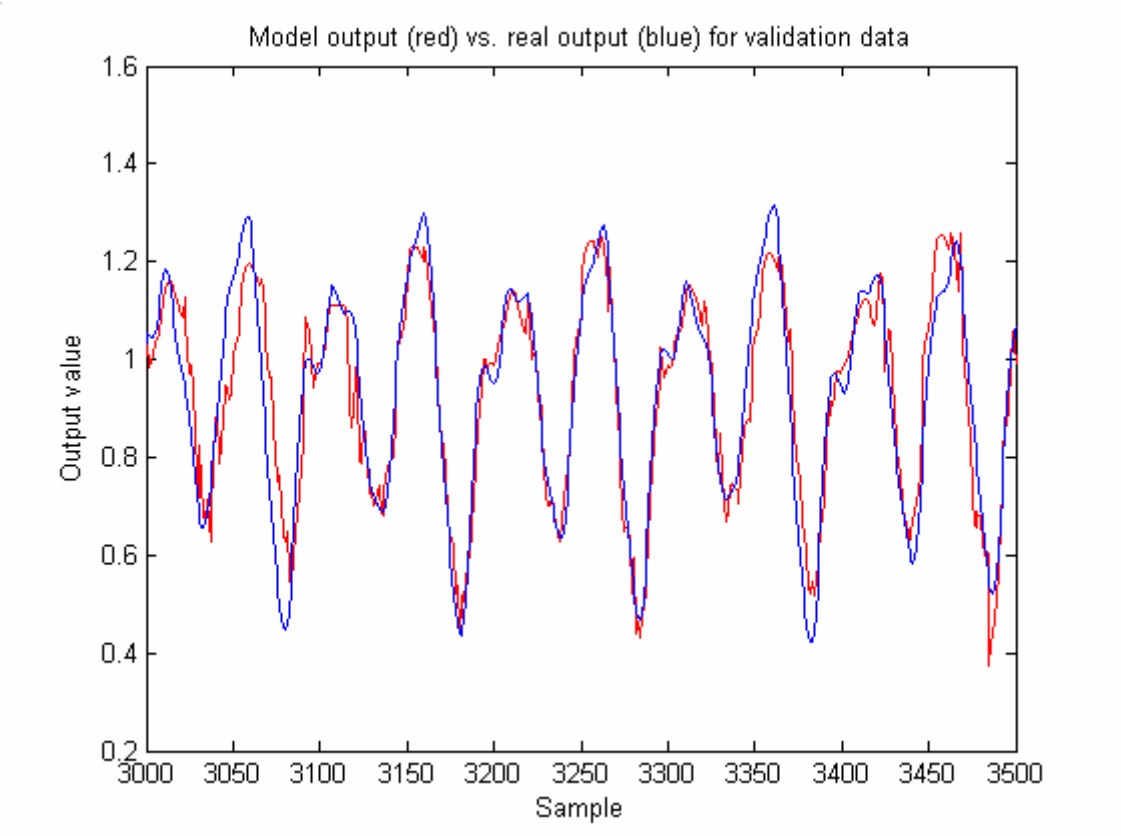

Figure 12: Model output for the testing data.

At a first glance it seems that this scenario allows better results than the previous one, which is true, but that is achieved with an increase in the number of rules, so it is natural that better values for RMSE and NDEI has been obtained.
This scenario also brings up something interesting. If we take a closer look at Figure 10 we can see that data points with minimum potential (local minima) match some of the local minima of the series output. This interesting fact can be useful in order to improve the models behaviour in these regions.

Scenario F

As we saw in scenario E we can use not only the maxima of the potential but also the minima in the process of creation and modification of rules.

Condition 1: the potential of the new data point is higher than the potential of the existing centres, i.e. $P_k(z_k) > P_k(z_i^*)$ or lower than the potential of the existing centres, i.e. $P_k(z_k) < P_k^{\inf}(z_i^*)$

Condition 2: the new data point is close to an old centre, given by $\frac{d_{\min}}{radii} < 0.5$

Condition 3: same as Condition 1

In a more formal way we have:

If $\{P_k(z_k) > P_k(z_i^*) \quad or \quad P_k(z_k) < P_k^{\inf}(z_i^*)\}$ and $\frac{d_{\min}}{radii} < 0.5$ **Then** $z_j^* = z_k$
Else If $\{P_k(z_k) > P_k(z_i^*) \quad or \quad P_k(z_k) < P_k^{\inf}(z_i^*)\}$ **Then** $R = R + 1; z_R^* = z_k$

Table 4 summarizes some of the information obtained for different values of radii.

Radii	New rules	Modified rules	RMSE Training	RMSE Validation	NDEI Training	NDEI Validation
0.1	61	7	0.05745	0.05762	0.25373	0.25437
0.2	32	37	0.06219	0.06121	0.27467	0.27021
0.3	23	51	0.06779	0.06657	0.29942	0.29388
0.4	17	52	0.07263	0.07210	0.32079	0.31827
0.5	13	67	0.08291	0.08173	0.36621	0.36081
0.6	9	87	0.08979	0.09084	0.39657	0.40099
0.7	8	95	0.08657	0.08467	0.38237	0.37376

Table 4: Results from scenario F.

For the particular case of radii = 0.4 we present the evolution of the potential, the rule base evolution and the model output for the testing data set.

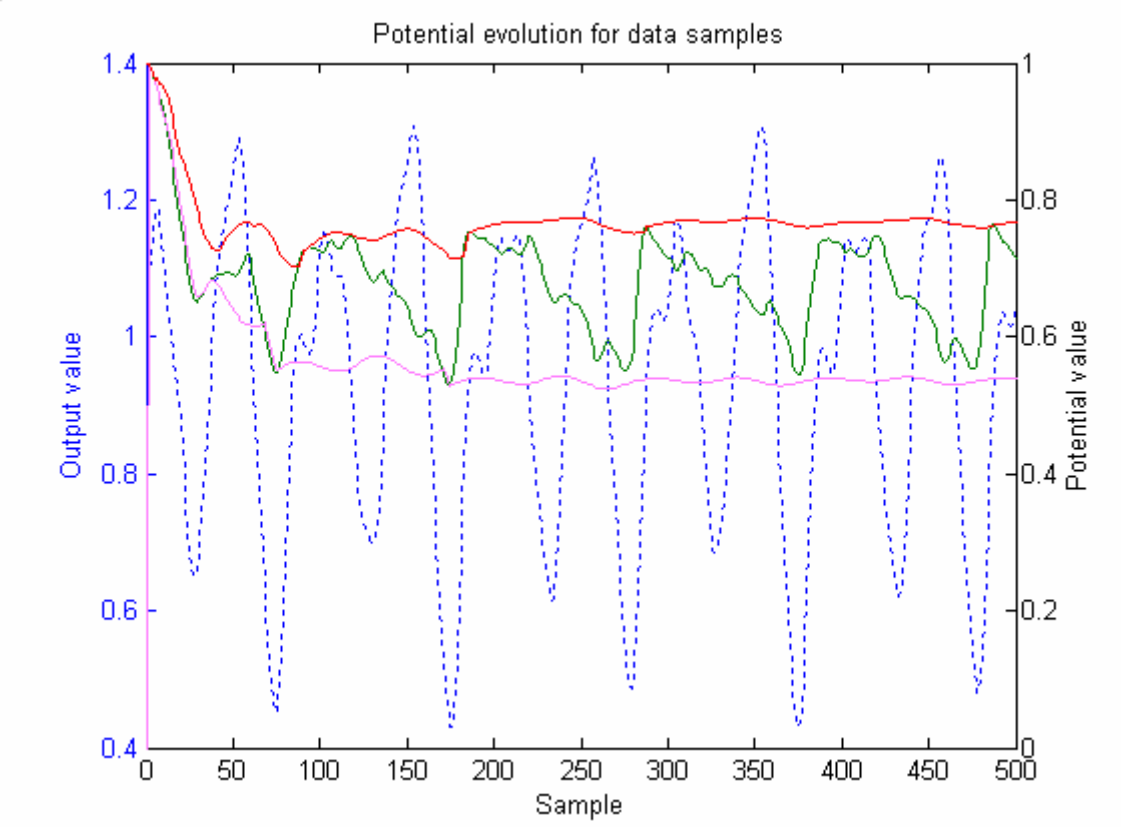

Figure 13: Evolution of the potential for the first 500 samples.

Figures 13 and 14 shows that there are some data points for which the potential is lower than the potential of the existing centres. More precisely the number of data points that verify the condition $P_k(z_k) < P_k^{\inf}(z_i^*)$ is 45 and the number of data points that verify the condition $P_k(z_k) > P_k(z_i^*)$ is 23. From the 17 new rules 3 were created with $P_k(z_k) > P_k(z_i^*)$ and 13 were created with $P_k(z_k) < P_k^{\inf}(z_i^*)$, the other one corresponds to the initial data point. In this scenario the rules creation process is more extended in time.

In Figure 15 is presented the model output for the testing data set and we can see that the model output has improved considerably in the regions of lower extrema. In the regions of upper extrema the behaviour remains the same.

A question that immediately arises and that must be answered is why the minimum potential of the existing centres should be considered. The minimization of the information potential makes sense since we are searching for some information details to better predict the series output. When the information potential is maximized the models generated predict the essential of the information, but the details are not present.

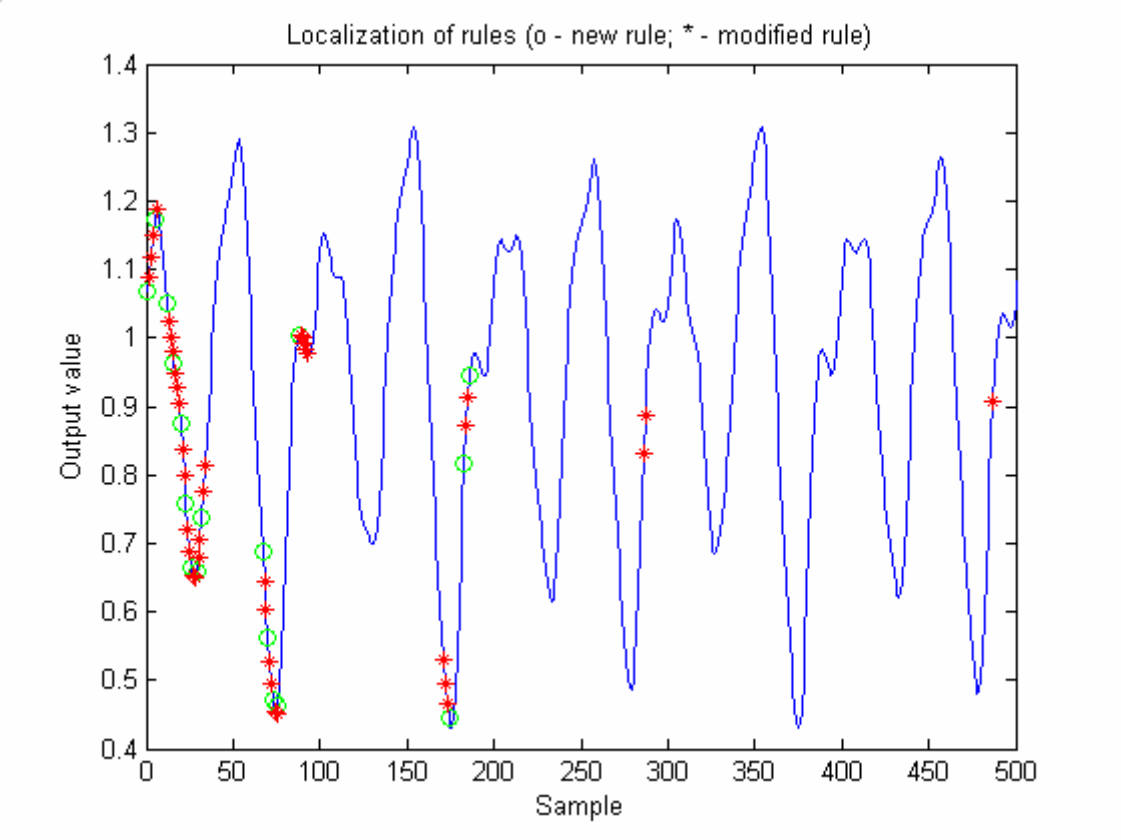

Figure 14: Rule base evolution.

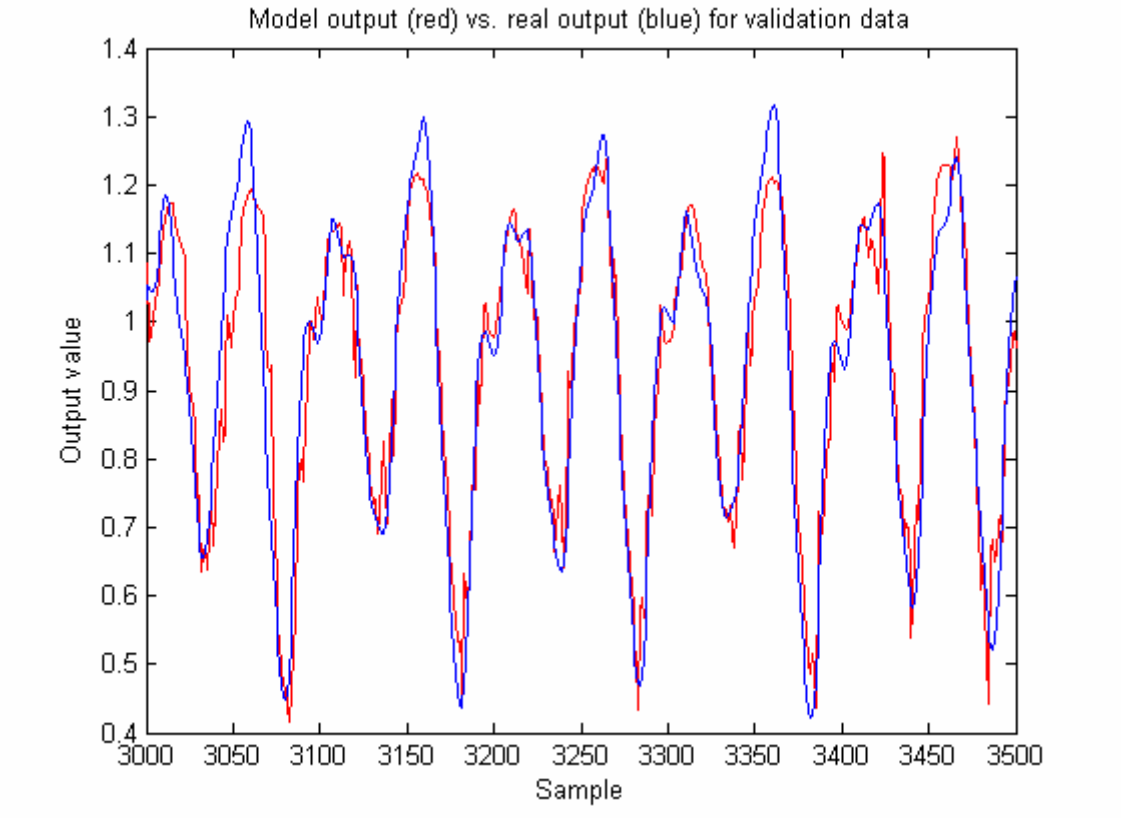

Figure 15: Model output for the validation data.

The minimization of the information potential allows recovering some of the missing details. The same for instance happens with PCA (Principal Component Analysis) technique where the main components describe the major of the information of the signal but the small information details (regions of extrema, noise, etc.) are precisely described by the components with less descriptive power.
If we make an analogy between information potential and entropy it is well known that when the entropy is maximized that does not mean it will be obtained the model with the smaller error, what is obtained is a model that maximizes the quantity of information it can describe.
Another important issue related with the introduction of the new condition in the process of creation and modification of rules is the model interpretability enhancement. This seems contradictory since the number of rules has increased, but if we analyse the location of the fuzzy sets we conclude that almost all the universe of discourse is covered, which did not happen before.

Scenario G

In scenario F the performance of the models has improve considerably, but we are interested in investigate if it is possible to achieve similar performances with a small number of rules.

Condition 1: the potential of the new data point is higher than the potential of the existing centres, i.e. $P_k(z_k) > P_k(z_i^*)$ or lower than the potential of the existing centres, i.e. $P_k(z_k) < P_k^{\inf}(z_i^*)$

Condition 2: the new data point is close to an old centre, given by $\frac{d_{\min}}{radii} < 0.5$ if $P_k(z_k) > P_k(z_i^*)$ or $\frac{d_{\min}}{radii} < 0.85$ if $P_k(z_k) < P_k^{\inf}(z_i^*)$

Condition 3:

In a more formal way we have:

If $\left\{ P_k(z_k) > P_k(z_i^*) \text{ and } \frac{d_{\min}}{radii} < 0.5 \right\}$ or $\left\{ P_k(z_k) < P_k^{\inf}(z_i^*) \text{ and } \frac{d_{\min}}{radii} < 0.85 \right\}$ **Then** $z_j^* = z_k$
Else If $P_k(z_k) > P_k(z_i^*)$ or $P_k(z_k) < P_k^{\inf}(z_i^*)$ **Then** $R = R + 1; z_R^* = z_k$

We conduct a study for different values of radii and Table 5 summarizes some of the information obtained.

Radii	New rules	Modified rules	RMSE Training	RMSE Validation	NDEI Training	NDEI Validation
0.1	42	23	0.06177	0.06155	0.27284	0.27173
0.2	23	42	0.06937	0.06782	0.30640	0.29938
0.3	14	67	0.08240	0.08064	0.36394	0.35597
0.4	11	70	0.08649	0.08481	0.38199	0.37439
0.5	8	65	0.09133	0.09398	0.40338	0.41486
0.6	7	78	0.09568	0.09678	0.42259	0.42724
0.7	5	153	0.09816	0.09595	0.43353	0.42355

Table 5: Results from scenario G.

For the particular case of radii = 0.4 we present the evolution of the potential, the rule base evolution and the model output for the testing data set.

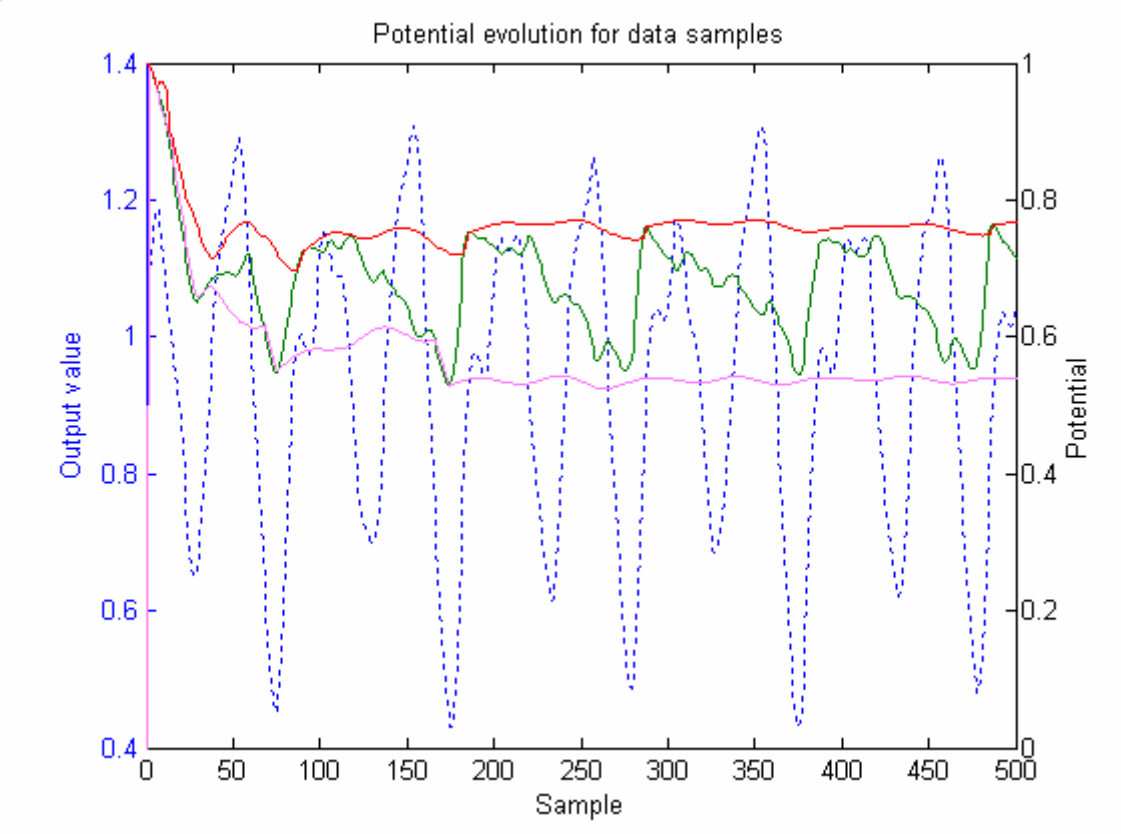

Figure 16: Evolution of the potential for the first 500 samples.

The consequence of the modification of Condition 2 is a reduction in the number of created rules, which slightly affects the model performance. On the other hand, the complexity of the models is reduced and the interpretability improves considerably.

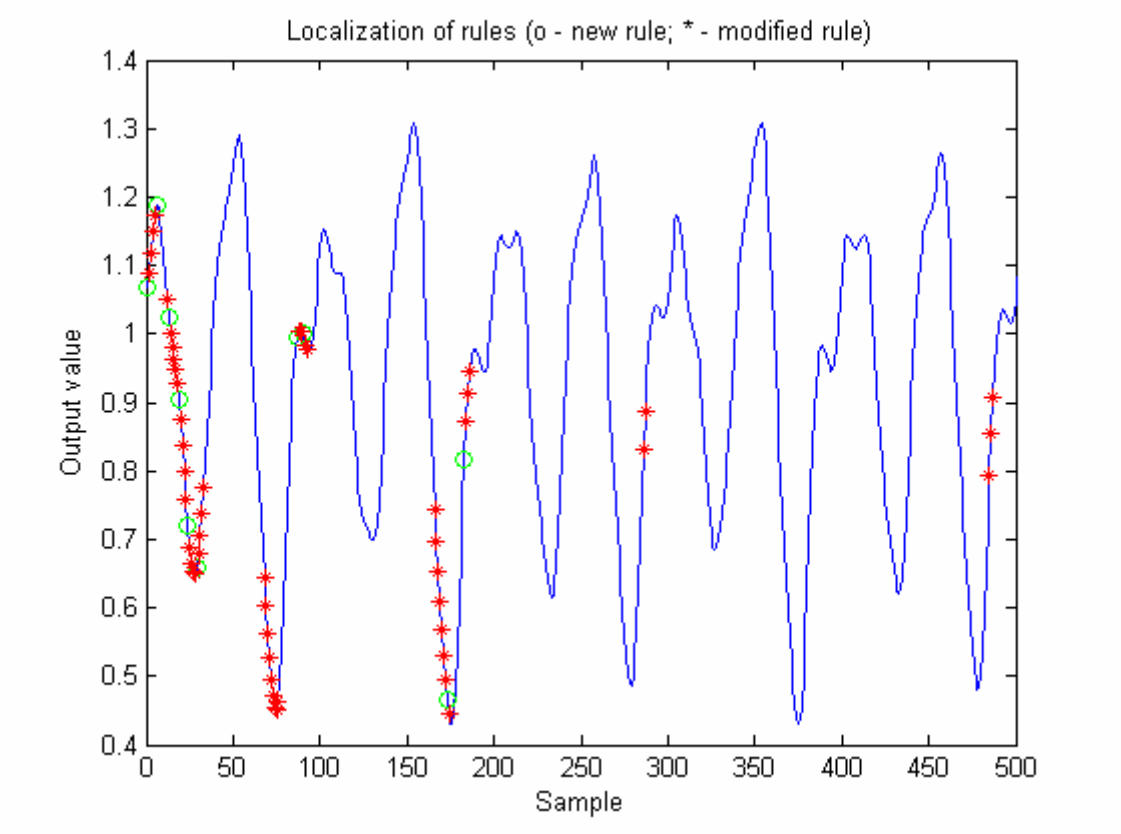

Figure 17: Rule base evolution.

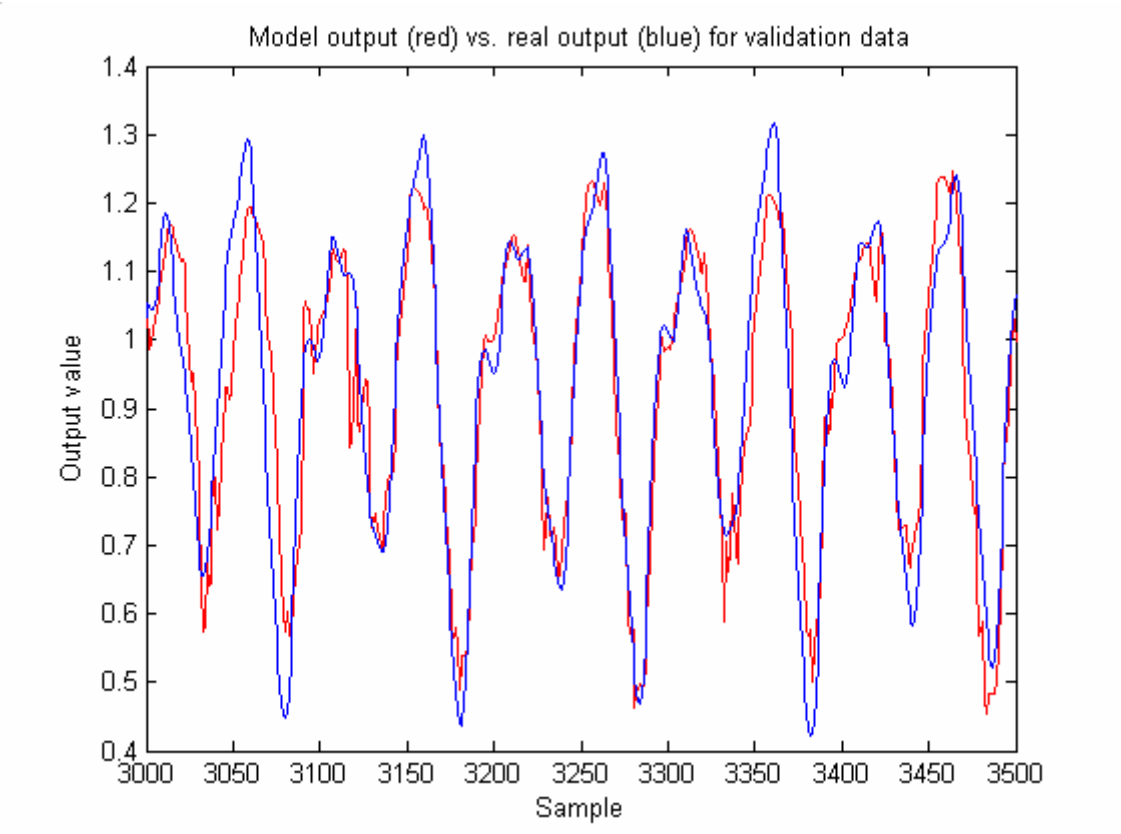

Figure 18: Model output for the testing data.

After all the experiments for the scenarios described in the report in Table 6 are presented the results for a particular value of the constant radii, radii = 0.4.

Scenario	Radii	New rules	Modified rules	RMSE Training	RMSE Validation	NDEI Training	NDEI Validation
A	--	--	--	--	--	--	--
B	--	--	--	--	--	--	--
C	0.4	19	0	0.08547	0.08670	0.37749	0.38272
D	0.4	6	27	0.09622	0.09392	0.42497	0.41460
E	0.4	15	305	0.08272	0.08024	0.36536	0.35422
F	0.4	17	52	0.07263	0.07210	0.32079	0.31827
G	0.4	11	70	0.08649	0.08481	0.38199	0.37439

Table 6: Results of the different scenarios for radii = 0.4.

From the table above we can see that for the same value of radii very different values for the number of rules created and modified are obtained. The performance of the models, measured by the RMSE and NDEI, also varies significantly.

In order to better understand the different scenarios in Table 7 are presented the results for a similar number of created rules.

Scenario	Radii	New rules	Modified rules	RMSE Training	RMSE Validation	NDEI Training	NDEI Validation
A	--	--	--	--	--	--	--
B	--	--	--	--	--	--	--
C	--	--	--	--	--	--	--
D	0.20	9	12	0.09610	0.09301	0.42444	0.41059
E	0.50	9	322	0.08813	0.08684	0.38927	0.38336
F	0.60	9	87	0.08979	0.09084	0.39657	0.40099
G	0.45	9	113	0.08933	0.08931	0.39456	0.39426

Table 7: Results for a similar number of new rules.

Table 7 lists the results when the number of rules created is 9. With the exception of scenario D the value for the constant radii are around 0.5. The performance in this case is better for scenario E.
However, it must be stated that in scenario E the definition of the relevant threshold, the upper one, was made after some experimentation, which means that if another data set is considered probably it will not work.

4.3 Comparative analysis

For the purpose of a comparative analysis we consider some existing online learning models applied on the same task. The models considered are RAN (Platt, 1991), ESOM (Deng and Kasabov, 2000), EFuNN (Kasabov,1998), DENFIS (Kasabov and Song, 2002) and ETS (Angelov and Filev, 2003).

Table 7 lists the prediction results (NDEI on testing data after online learning) and the number of rules (nodes, units) evolved in each model.

Methods	Rules (nodes, units)	NDEI
RAN	113 units	0.373
ESOM	114 units	0.320
EFuNN	193 rule nodes	0.401
DENFIS	58 fuzzy rules	0.276
ETS	113 fuzzy rules	0.095
ETS*	13 fuzzy rules	0.360

Table 8: Online learning models.

From the table above we conclude that it is possible to build a model with reasonable accuracy with a small number of rules, i.e. more transparent models. In this study a model with only 13 rules (result from scenario F) has a performance similar to models with a much higher number of units or rule nodes.
It should be noted that models with much lower NDEI have been reported, but the number of rules (nodes or units) is in the range of thousands, which completely undermines their transparency.
As we previously stated our concern is not only to obtain accurate models but also interpretable ones. The interpretability of the models generated with the approach described in this study can be significantly improved since some of the fuzzy sets are quite similar. Improvements on the interpretability of the models will reduce its complexity but the price to pay is a smaller accuracy.

5. Conclusions

The approach to on-line identification of ETS models proposed by Angelov is based on recursive, non-iterative building of the rule base by hybrid learning. The rule-based model evolves by replacement or upgrade of rules and parameter estimation. It is computationally effective, as it does not require retraining of the whole model.
The adaptive nature of this model, in addition to the highly transparent and compact form of fuzzy rules, makes them a useful tool for on-line modelling. The main advantages of the approach are (Angelov and Filev, 2003):

- It can develop/evolve an existing model when the data pattern changes, while inheriting the rule base;
- It can start to learn a process from a single data sample or evolve an existing model and improve the performance of predictions on-line;
- It is non-iterative and recursive and hence computationally very effective.

The approach however still needs some improvements. The conditions to modify and upgrade the fuzzy rules must be studied more deeply since they influence the number of created and modified rules and it is not easy to adjust the definitions for a specific problem.
Another vital issue is the on-line clustering algorithm, particularly the function for recursive calculation of the potential of each new data sample. Angelov used different functions (Cauchy type function of first order, exponential with the summation in the exponent) for recursive calculation of the potential but all present limitations. New functions or estimators from information theory need to be tested to achieve a better placement for focal points covering not only the regions of higher density of points but also other regions of interest.
The transparency of the models must also be improved since the fuzzy sets most of the times are very similar, particularly when the number of rules is high. This objective however is not easy to perform on-line.

References

P. Angelov and D. Filev (2003). *An Approach to On-Line Identification of Takagi-Sugeno Fuzzy Models* (to appear).

P. Angelov and R. Buswell (2002). *Identification of Evolving Rule-Based Models*. IEEE Transactions on Fuzzy Systems, vol. 10, no. 5, pp. 667-677.

P. Angelov and D. Filev (2002). *Flexible Models with Evolving Structure*. IEEE Symposium on Intelligent Systems.

P. Angelov (2002). *An approach for Fuzzy Rule-Base Adaptation using On-Line Clustering*.

P. Angelov (2002). *Evolving Rule-Based Models: A Tool for Design of Flexible Adaptive Systems*, Physica-Verlag.

S. Chiu (1994). *Fuzzy Model Identification Based on Clustering Estimation*. Journal of Intelligent and Fuzzy Systems, vol. 2, pp. 267-278.

D. Deng and N. Kasabov (2000). *Evolving Self-Organizing Maps for Online Learning, Data Analysis and Modeling*. Proceedings of IJCNN'2000, vol. VI, pp. 3-8.

N. Kasabov and Q. Song (2002). *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, April.

N. Kasabov (1998). *Evolving Fuzzy Neural Networks – Algorithms, Applications and Biological Motivation*. Methodologies for the Conception, Design and Application of Soft Computing, T. Yamakawa and G. Matsumoto Eds, Singapore: World Scientific, pp. 271-274.

J. Platt (1991). *A Resource Allocation Network for Function Interpolation*. Neural Computation, vol. 3, pp. 213-225.