On Data–Driven Fuzzy Partition in the Fuzzy–Probabilistic Inference System Framework

CAO Nhung and HOLČAPEK Michal and VALÁŠEK Radek

Institute for Research and Applications of Fuzzy Modeling, University of Ostrava
30. dubna 22, 701 03 Ostrava, Czech Republic
E-mail: {nhung.cao, michal.holcapek, radek.valasek}@osu.cz

1 Preliminaries and motivation

This paper is focused on *fuzzy–probabilistic IF–THEN rules based systems*, where the antecedents encode fuzzy information and consequents represent probability distributions of the output variable. This system, which combines both types of uncertainty in one framework, can be applied effectively in time series analysis and forecasting. Let $\mathbb{U}$ be a universe, let $\Delta = \{A_1, \ldots, A_m\}$ be a fuzzy covering of $\mathbb{U}$, i.e., for every $x \in \mathbb{U}$, at least one $A_k(x) > 0$, and Y denote a random variable defined in a probability space $(\Omega, \mathcal{A}, P)$. Then, the rules take the form (see, [1]):

$$R_k: \text{ IF } X \text{ is } A_k \text{ THEN } Y \sim Q_k(p), \quad k = 1, \ldots, m, \tag{1}$$

where $Q_k(p)$ denotes the quantile function of the conditional distribution of Y given A_k. It is worth noting that, in practice, uniform or generalized fuzzy partitions are typically constructed by shifting equidistant fuzzy sets along the domain axis. The quantile functions in the antecendents of rules in (1) are estimated from data as follows. Assume a dataset $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{U} \times \mathbb{R}$ and let Δ denote a fuzzy covering of $\mathbb{U}$. For any $p \in [0, 1]$, the quantile function $Q_k(p)$ is defined as the p–weighted quantile of the pairs $(y_1, A_k(x_1)), \ldots, (y_n, A_k(x_n))$. In particular, the p–weighted quantile is the value $z_p \in \mathbb{R}$ that minimizes the functional

$$\Phi_p(z) = \sum_{i=1}^{n} \rho_p(y_i - z)\, A_k(x_i), \quad \rho_p(u) = \begin{cases} p|u|, & u > 0, \\ (1-p)|u|, & u \leq 0. \end{cases} \tag{2}$$

The inference mechanism, which estimates the quantile function $Q_x(p)$ for any $x \in \mathbb{U}$, is defined as the weighted average of quantiles $Q_k(p)$:

$$Q_x(p) = \frac{\sum_{k=1}^m A_k(x) Q_k(p)}{\sum_{k=1}^m A_k(x)}. \tag{3}$$

Note that this is not the only way to define the inference mechanism for (1); several alternatives can be found in [2].

The effectiveness of fuzzy–probabilistic inference systems (FPIS) has been demonstrated in various applications, for example, in time series analysis [2]. However, several aspects of such systems remain open and challenging, particularly the construction of an appropriate fuzzy

Acknowledgement The study is supported by the project "Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583" which is co-financed by the European Union.

https://www.doi.org/10.15452/978-80-7599-515-5.2026.05

partition. In practice, uniform fuzzy partitions are commonly used because of their simplicity, yet they fail to capture the complex structures hidden in the data. This raises a key question: can we construct a data-driven fuzzy partition that better reflects local behavior under a well-defined criterion? This short paper introduces three methods based on different criteria for designing non-uniform, data-dependent fuzzy partitions, presented in algorithmic form. A detailed theoretical analysis of these methods is left for future work.

2 Data-driven fuzzy partitions

Criterion 1: Limited range size for data values. Assume, for simplicity, that the dataset is given by $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{U} \times \mathbb{R}$, where $x_i = i$. Let $\varepsilon > 0$ be a predefined threshold. For a finite vector $\bar{y} = (y_1, \ldots, y_k)$, recall that its *range* is defined as

$$R(\bar{y}) = \max\{y_1, \ldots, y_k\} - \min\{y_1, \ldots, y_k\}.$$

We construct consecutive subvectors $\bar{y}_1, \ldots, \bar{y}_n$, where each $\bar{y}_k = (y_k, \ldots, y_{k+m_k})$ is determined such that

$$R(\bar{y}_k) \le \varepsilon \quad \text{and} \quad R(\bar{y}_k^+) > \varepsilon, \tag{4}$$

where $\bar{y}_k^+ = (y_k, \ldots, y_{k+m_k+1})$. Each vector $\bar{y}_k$ is associated with an interval $I_k = [k, k+m_k]$, which serves as the *support* of the fuzzy set A_k in the resulting fuzzy partition. The node of A_k is placed at the midpoint of I_k. Since not all constructed intervals are useful for the final partition, we propose three *selection strategies* to determine the next interval $I_{k_{\ell+1}}$, given the previously selected ℓ intervals $I_1, \ldots, I_{k_\ell}$. The next index $k_{\ell+1}$ is chosen based on $\bar{y}_{k_\ell}$ as follows:

(i) **Extrema-based:** choose the smaller of the indices corresponding to the minimum and maximum elements of $\bar{y}_{k_\ell}$;

(ii) **Median-based:** choose the index corresponding to the (classical) median of $\bar{y}_{k_\ell}$, ensuring that consecutive supports do not overlap in median values;

(iii) **Weighted median-based:** choose the index corresponding to the *weighted median* of $\bar{y}_{k_\ell}$, where the weights are induced by the membership degrees $A_{k_\ell}(i)$ for $i = k_\ell, \ldots, k_\ell + m_{k_\ell}$.

Note that the threshold ε must be appropriately adjusted to obtain a reasonable number of fuzzy sets in the resulting fuzzy partition. Obviously, a smaller value of ε results in a higher number of fuzzy sets, and vice versa.

Example 1 *Consider synthetic data $y_i = g(i)$ with*

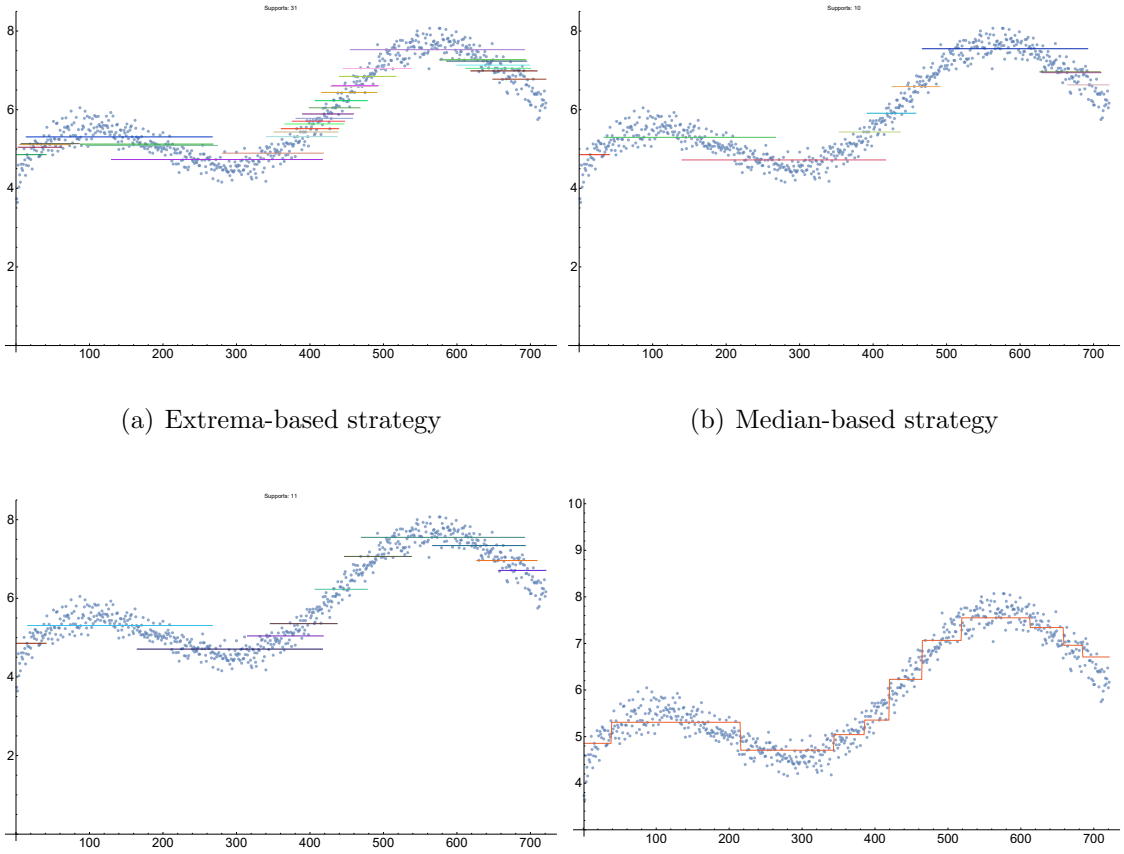
$$g(x) = \ln(x+20) + \cos(x/90) + \tfrac{1}{3}\sin\left(\frac{x\log(x+50)}{200}\right) + \upsilon(x),\ \upsilon(x) \sim \mathcal{N}(0, 0.2^2).$$

Let $\varepsilon = 1.7$. Figure 1 illustrates the supports obtained from the three starting-index methods and the corresponding moving median computed via L_1-based minimization [2]. The moving median exhibits favorable behavior, aligning well with the observed data pattern.

Criterion 2: Limited variance of data values. This criterion is a modification of the previous one, in which the range is replaced by the variance. Specifically, condition (4) is replaced by

$$\mathrm{Var}(\bar{y}_k) \le \varepsilon \quad \text{and} \quad \mathrm{Var}(\bar{y}_k^+) > \varepsilon, \tag{5}$$

where the variance of the vector components is computed in the standard way. All remaining steps, including the selection strategies, are identical to those described earlier. Figure 2 illustrates the result for Example 1 with $\varepsilon = 0.3$ and the weighted-median-based strategy.



(a) Extrema-based strategy

(b) Median-based strategy

(c) Weighted median-based strategy

(d) Moving median based on the third strategy

Figure 1: Resulting fuzzy set supports based on Criterion 1 and the moving median.

Criterion 3: Relative error of variances of local linear regression residuals. The third strategy evaluates variance stability by analyzing the residuals of linear regressions across intervals and their subintervals. The domain is partitioned by recursive bisection whenever significant relative changes in variance are detected.

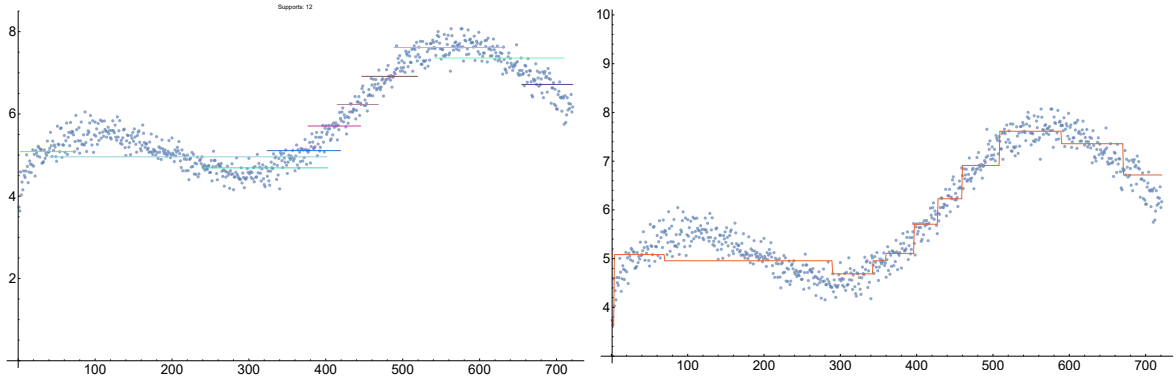
Let I denote an interval (e.g., $I = [1, n]$) and let σ_I denote the standard deviation of the residuals for the linear regression model over the interval I. We define the *relative error of residual variances* over the interval I (with $\sigma_I > 0$) and its subinterval J as

$$R_{I,J} = \frac{\sigma_I - \sigma_J}{\sigma_I}. \tag{6}$$

Denote by $\bar{x}_I$ the mean, by $V_I = \sigma_I / \bar{x}_I$ the coefficient of variation, and by n_I the number of integers in I. Let $R_{\min}$, $V_{\min}$, and $n_{\min}$ be the minimum thresholds for the relative error, the coefficient of variation, and the interval size, respectively.

The partition procedure recursively divides each interval into two half-subintervals according to the following rule. Assume that an interval I has been divided by the partition procedure into two half-subintervals. Let $I/2$ denote the left (or right) subinterval whose right (or left) endpoint coincides with the midpoint of I, where the endpoints are appropriately rounded to consecutive integers. The subinterval $I/2$ is further divided into its two half-subintervals if $n_{I/2} > n_{\min}$ and either of the following conditions holds:

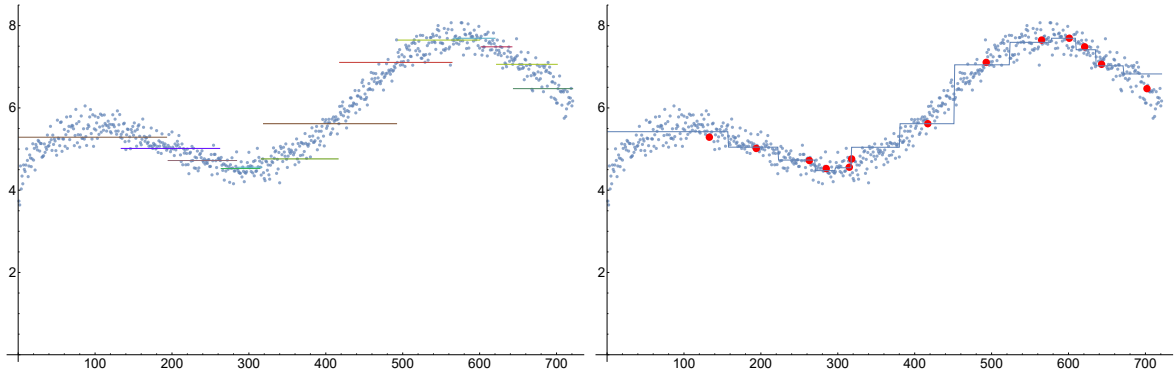
$$R_{I,I/2} > R_{\min} \quad \text{or} \quad V_{I/2} > V_{\min}. \tag{7}$$



(a) Weighted median-based strategy (b) Moving median based on the third strategy

Figure 2: Resulting fuzzy set supports based on Criterion 2 and the moving median.

When $I/2$ is further subdivided and the bisection procedure is invoked with new parameters, the linear regression is recomputed on the same dataset. Hence, σ_J can be passed as a parameter to replace σ_I in the next step. After finitely many steps, we obtain a disjoint family of subintervals $\{I_j\}$ of $I = [1, n]$. To construct overlapping supports for the fuzzy partition, each I_j is extended by redefining its endpoints as the indices of the medians of values in the adjacent intervals I_{j-1} and I_{j+1} (cf. Median-based selection strategy). Figure 3 illustrates this procedure on data from Example 1, with red bold dots indicating medians within intervals.



(a) Resulting supports (b) Moving median based on the received supports

Figure 3: Resulting fuzzy set supports based on Criterion 3 and the moving median.

References

[1] Madrid, N.: *Significance measures for rules in probabilistic-fuzzy inference systems based on fuzzy transforms.* Fuzzy Sets and Systems. **467** (2023) 108575.

[2] Cao, N., Holčapek, M., Valášek, R., Madrid, N., Neděla, D.: *An Investigation of Alternative Methods for the Inference of Probabilistic-Fuzzy Systems.* In: International Conference on Modeling Decisions for Artificial Intelligence. Cham: Springer Nature Switzerland (2025) 104–116.