Short-Time Prediction Based on Recognition of Fuzzy Time Series Patterns

Gernot Herbst and Steffen F. Bocklisch

Chemnitz University of Technology, D-09107 Chemnitz, Germany
 gernot.herbst@etit.tu-chemnitz.de

Abstract. This article proposes knowledge-based short-time prediction methods for multivariate streaming time series, relying on the early recognition of local patterns. A parametric, well-interpretable model for such patterns is presented, along with an online, classification-based recognition procedure. Subsequently, two options are discussed to predict time series employing the fuzzified pattern knowledge, accompanied by an example. Special emphasis is placed on comprehensible models and methods, as well as an easy interface to data mining algorithms.

1 Introduction

Predicting “nasty” time series (instationary, multivariate etc.—stemming from nonlinear dynamic processes) based on a *global* model, such as difference equations, resembles a David-sized answer to a Goliath-like problem, with questionable success in the general case, though. In many real-world time series, however, we may observe patterns that recur not in identical, but similar form according to a regular or irregular scheme, often due to the diversity of loose or strict rhythms and periodicities inherent in the gamut of many natural or social processes. A modest but sound answer to time series prediction might therefore lie in gaining data-based *local* knowledge of a process and employing this for predictions later on. In the following, an approach to the latter shall be outlined which explicitly allows for the fundamental uncertainty attached to this task.

2 Fuzzy Time Series Patterns

2.1 A Multivariate Parametric Fuzzy Set

The elementary approach for modelling time series patterns followed in this paper is a sample-point-wise model using fuzzy sets for each sample. In order to cope with multivariate time series, a suitable multivariate fuzzy set from [1] will be introduced beforehand. A rather unique feature of this set is that it can be derived in parametric form from an intersection of univariate parametric fuzzy sets of the following type:

$$\mu(x) = \frac{1}{1 + \left(\frac{1}{b_{l/r}} - 1 \right) \cdot \left| \frac{x-r}{c_{l/r}} \right|^{d_{l/r}}} \quad \text{with} \quad \begin{cases} b_l, c_l, d_l & \text{for } x < r \\ b_r, c_r, d_r & \text{for } x \geq r \end{cases} \quad (1)$$

The effect of r and the side-specific parameters b and c can be understood from Fig. 1a. While $c_{l/r} > 0$ quantify the uncertainty (as in crisp sets), $b_{l/r} \in (0, 1]$ and $d_{l/r} \geq 2$ account for the fuzziness of this uncertain information, where increasing values of $d_{l/r}$ lead to sharper descents of the membership value to zero and $d_{l/r} \rightarrow \infty$ result in rectangular (crisp) sets. For the multivariate extension, N fuzzy sets of this type are being combined using a compensatory HAMACHER intersection (2), resulting in an N -dimensional parametric membership function $\mu : \mathbb{R}^N \mapsto (0, 1]$, cf. (3). Exemplary sets for $N = 2$ are shown in Fig. 1b.

$$\cap_{\text{Ham}}^N \mu_i = \left(\frac{1}{N} \sum_{i=1}^N \frac{1}{\mu_i} \right)^{-1} \quad (2)$$

$$\begin{aligned} \mu(\mathbf{x}) = & \left(1 + \frac{1}{2N} \sum_{i=1}^N [1 - \text{sgn}(x_i - r_i)] \cdot \left(\frac{1}{b_{li}} - 1 \right) \cdot \left| \frac{x_i - r_i}{c_{li}} \right|^{d_{li}} \right. \\ & \left. + \frac{1}{2N} \sum_{i=1}^N [1 + \text{sgn}(x_i - r_i)] \cdot \left(\frac{1}{b_{ri}} - 1 \right) \cdot \left| \frac{x_i - r_i}{c_{ri}} \right|^{d_{ri}} \right)^{-1} \end{aligned} \quad (3)$$

Fig. 1. The multivariate parametric membership function (3).

2.2 Modelling and Classification of Time Series Patterns

Equation (3) can now be employed to model equidistantly sampled multivariate time series patterns. For a pattern of length L (sample points), L membership functions are being used and result in a progression of fuzzy sets that can be interpreted and displayed as a fuzzy corridor for instances of the respective pattern, cf. Fig. 2. We will denote the fuzzy set for the t -th sample point of a pattern by $\mu_{p,t}(\mathbf{x})$. With (3) being an unimodal function, the mean course of a pattern is therefore being captured along with the uncertainty in its realisations.

The parameters of the fuzzy sets $\mu_{p,t}$ may either be formulated by experts, or—as described in [2]—determined automatically from sets of pattern instances. The latter case is especially suited for joint use with time series motif mining algorithms, e. g. [6],

Fig. 2. Fuzzy time series pattern along with a noisy candidate sequence.

with the sole assumption that instances of one pattern are equal in length and similar in the fuzzy sense of our model.

To classify whole time series $\mathbf{x}(1), \dots, \mathbf{x}(L)$ of the same length as the fuzzy model given by $\mu_{P,1}(\mathbf{x}), \dots, \mu_{P,L}(\mathbf{x})$, it is only necessary to combine the individual (elementary) classification results for all samples $\mathbf{x}(t)$ in their respective classes $\mu_{P,t}(\mathbf{x})$:

$$\mu = \mu_{P,1}(\mathbf{x}(1)) \cap \dots \cap \mu_{P,L}(\mathbf{x}(L)) \quad (4)$$

If (2) is being used for the operator $\cap$ again, we essentially obtain an $(N \cdot L)$ -dimensional fuzzy classifier as a natural extension of the multivariate set (3), i. e. with the same parametric structure and mode of operation. Figure 2 gives a visual example of a classifier for a univariate time series sequence.

One important advantage of this approach has to be seen in its ability to classify subsequences of a pattern, i. e. instances of a length less than L , by intersecting only classification results for sample points that are available. We will rely on this property in Sect. 3.2 to classify incompleting pattern instances.

3 Online Recognition of Patterns

3.1 Problem Statement

In a context of streaming time series, with one new datum $\mathbf{x}(t)$ being available at each point in time t , we want to be able to recognise previously known local patterns (each of them modelled by the fuzzy means introduced in Sect. 2.2) in an online manner. If we are able to detect a pattern before it is completed (say, at a stage $\tau < L$), a short-time prediction can be derived based upon the knowledge still “left” for the pattern’s $(L - \tau)$ remaining sample points.

If we consequently follow a fuzzy approach, however, a recognition system will (and should) recognise every known pattern in every possible stage of development, all at the same time. Although this wealth of information will obviously have to be narrowed down to usable “key” information later on, we will show how to work with and modify the recognition results before. On the other hand, this calls for a recognition approach with computational requirements which are not prohibitive, while still delivering every possible recognition result.

3.2 Motivation for Recursive Equations

At any point in time t , a pattern instance in a streaming time series $\mathbf{x}(t)$ may be present and will, due to the fuzzy approach to pattern recognition, be detected in every possible stage $\tau = 1, \dots, L$. We will denote the classification result for an incomplected pattern at stage τ by $\mu(t, \tau)$, which means that L such results form the complete recognition result for a pattern at one point in time t . For every possible value of τ , $\mu(t, \tau)$ can be computed as given by (5).

$$\mu(t, \tau) = \mu_{p,1}(\mathbf{x}(t - \tau + 1)) \cap \dots \cap \mu_{p,\tau}(\mathbf{x}(t)), \quad \tau = 1, \dots, L \quad (5)$$

Unfortunately, this also means that $\sum_{i=1}^L i = \frac{1}{2}L \cdot (L + 1)$ elementary classification results $\mu_{p,1 \dots L}$ would have to be computed at every time step. If we compare previous recognition results $\mu(t - 1, \tau)$ against the current results $\mu(t, \tau)$, however, it can be noticed that some of them share almost all elementary classification results, as depicted in Fig. 3a. More precisely, $\mu(t, \tau)$ could be recursively derived from $\mu(t - 1, \tau - 1)$ by incorporating $\mu_{p,\tau}(\mathbf{x}(t))$, as shown in (6). As a consequence, only L elementary classification results $\mu_{p,1 \dots L}$ would have to be computed at each point in time to update the recognition results. This equals the computational cost of detecting only completed patterns by using (4), i. e. there is no computational overhead regarding elementary classification steps to obtain recognition results for incomplected patterns at every possible stage.

$$\mu(t, \tau) = \underbrace{[\mu_{p,1}(\mathbf{x}(t - \tau + 1)) \cap \dots \cap \mu_{p,\tau-1}(\mathbf{x}(t - 1))]}_{\mu(t-1, \tau-1)} \cap \mu_{p,\tau}(\mathbf{x}(t)) \quad (6)$$

Fig. 3. Recursive classification (recognition) of time series patterns.

3.3 Update Equations for Recursive Pattern Recognition

In order to derive a recursive equation for $\mu(t, \tau)$ like (6), which must deliver equivalent results compared to the non-recursive equation (5), an intersection operator $\cap$ is needed which preserves the weight of the left- and right-hand side truth values, as $\mu(t - 1, \tau - 1)$ is already the result of the intersection of $(\tau - 1)$ truth values. In [3] we showed how

to extend the compensatory HAMACHER intersection (2) to a weighted conjunction (7) preserving given weights as needed here. This allows to reformulate (6) to (8).

$$\mu_a \overset{N_a}{\cap} \overset{N_b}{\mu_b} = \frac{1}{\frac{1}{N_a + N_b} \left(\frac{N_a}{\mu_a} + \frac{N_b}{\mu_b} \right)} \quad (7)$$

$$\mu(t, \tau) = \mu(t-1, \tau-1) \overset{(\tau-1)}{\cap} \mu_{p,\tau}(\mathbf{x}(t)) \quad (8)$$

If we rearrange the pattern recognition results $\mu(t, \tau) \forall \tau$ in a vector $\boldsymbol{\mu}_\tau(t)$, the elementary classification results of $\mathbf{x}(t)$ to all classes $\mu_{p,1..L}$ in $\boldsymbol{\mu}_p(\mathbf{x}(t))$, and define a vector of weights $\mathbf{n}_\tau$,

$$\boldsymbol{\mu}_\tau(t) = \begin{pmatrix} \mu_{\tau,1}(t) \\ \vdots \\ \mu_{\tau,L}(t) \end{pmatrix} = \begin{pmatrix} \mu(t,1) \\ \vdots \\ \mu(t,L) \end{pmatrix}, \quad \boldsymbol{\mu}_p(\mathbf{x}(t)) = \begin{pmatrix} \mu_{p,1}(\mathbf{x}(t)) \\ \vdots \\ \mu_{p,L}(\mathbf{x}(t)) \end{pmatrix}, \quad \mathbf{n}_\tau = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ L-1 \end{pmatrix},$$

we can—as done in [3]—obtain a vectorial update equation for the recursive classifier depicted in Fig. 3b:

$$\boldsymbol{\mu}_\tau(t) = \left(\begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \mathbf{0}^T & 0 \\ \mathbf{I} & \mathbf{0} \end{pmatrix} \cdot \boldsymbol{\mu}_\tau(t-1) \right) \mathbf{n}_\tau \cap \boldsymbol{\mu}_p(\mathbf{x}(t)) \quad (9)$$

3.4 Post-Processing of Recognition Results

For prediction purposes, early recognition results $\mu(t, \tau)$ with $\tau \ll L$ promise the largest prediction horizon, but are quite unreliable, as based on only very few data points. On the other hand, almost completed patterns ($\tau \approx L$) may be detected reliably, but are rather pointless by leaving nothing to predict.

It appears difficult to define strict boundaries for both reliable and usable values of τ , which is why we propose to formulate a soft compromise by means of a fuzzy set $\mu_w: \{1, \dots, L\} \mapsto [0, 1]$, called the *fuzzy window of interest* w. r. t. τ . This window may then be applied to $\mu(t, \tau)$, leading to so-called *windowed recognition results* in (10).¹ Semantically, (10) corresponds to a coincidence of interest in a certain stage τ and the actual recognition of a pattern in this stage. An advantage of the fuzzy recognition results $\mu(t, \tau)$ is that we can perform this windowing procedure in a completely fuzzy manner before any decision step.

$$\tilde{\mu}(t, \tau) = \mu(t, \tau) \cap \mu_w(\tau) \quad \forall \tau \quad (10)$$

When these results shall serve as a rationale for concrete actions such as predictions, however, crisp values for the similarity and stage of the detected pattern will often be necessary. For the time being, we will employ a first-of-maxima (FOM) approach to the defuzzification of $\tilde{\mu}(t, \tau)$ to obtain crisp results $\tilde{\mu}^*$ and τ^* at every point in time.

¹ Non-compensatory intersections (such as all T-norm operators) should be employed for $\cap$ in (10). In this paper, the HAMACHER product will be used.

4 Pattern-Based Short-Time Prediction

As soon as one or more evolving patterns are being detected in a streaming time series (resulting in $\tilde{\mu}^*$ and τ^* as described in the previous section), a short-time prediction of the time series can be provided for a horizon of $(L - \tau^*)$ samples for every detected pattern. While the fuzzy recognition system will be able to detect any pattern at any time, dissimilar pattern instances (with small values of $\tilde{\mu}^*$) will obviously not form a good foundation for predictions and should therefore be neglected, i. e. by requiring $\tilde{\mu}^*$ to pass a certain threshold. Subsequently, two prediction methods will be presented which work with *one* detected pattern; the combination of several of these predictions will be discussed afterwards.

4.1 Prediction Methods

(I) *Knowledge-based prediction based on a pattern's mean course.* Given the fuzzy time series model of Sect. 2.2 and especially its visual representation in Fig. 2, it almost suggests itself to base a prediction of a detected, but not yet completed pattern (stage τ^*) on its mean course. In the underlying fuzzy sets from Sect. 2.1, this corresponds to the modal parameters $\mathbf{r}$ in (3) for each sample point. For a prediction horizon p , $1 \leq p \leq L - \tau^*$, we can simply use these values for a prediction of the time series $\mathbf{x}$ beyond time t :

$$\hat{\mathbf{x}}(t + p) = \mathbf{r}(\tau^* + p) \quad (11)$$

(II) *Extrapolating prediction based on fuzzy implication.* While this method from [7] was first and foremost designed for univariate ($N = 1$) global time series and their fuzzy models, it may be used for multivariate ($N > 1$) local time series patterns as well, if applied for every component x_i , $i = 1, \dots, N$ of the time series according to the following procedure.

In the fuzzy description of the τ^* -th sample point, the position of $x(t)$ in comparison to $r(\tau^*)$ is determined as in (12). For the predicted value of the next time step $\hat{x}(t + 1)$ it is firstly assumed that the qualitative position $\delta(\tau^* + 1)$ in relation to $r(\tau^* + 1)$ will remain the same. Secondly it is implied that if the overall recognition result for the known sample points is $\mu(t, \tau^*)$, the predicted value $\hat{x}(t + 1)$ would produce the same value for its future classification result in $\mu_{p, \tau^*+1}(\hat{x}(t + 1))$, cf. (13), similar to the propagation of a truth value from the antecedent part of a fuzzy rule to its consequence.

$$\delta(\tau^*) = \begin{cases} -1, & x(t) < r(\tau^*) \\ +1, & x(t) \geq r(\tau^*) \end{cases} \quad (12)$$

$$\mu_{p, \tau^*+1}(\hat{x}(t + 1)) \stackrel{!}{=} \mu(t, \tau^*) \quad (13)$$

With these premises and $\delta(\tau^*)$ at hand, it is possible to obtain $\hat{x}(t + 1)$ by rearranging the membership function μ_{p, τ^*+1} , given in (1), to (14).

$$\hat{x}(t+1) = r(\tau^* + 1) + \begin{cases} - \left(\left(\frac{1}{\mu(t, \tau^*)} - 1 \right) \cdot \left(\frac{b_l(\tau^* + 1)}{1 - b_l(\tau^* + 1)} \right) \right)^{\frac{1}{d_l(\tau^* + 1)}} \cdot c_l(\tau^* + 1), & \delta(\tau^*) = -1 \\ + \left(\left(\frac{1}{\mu(t, \tau^*)} - 1 \right) \cdot \left(\frac{b_r(\tau^* + 1)}{1 - b_r(\tau^* + 1)} \right) \right)^{\frac{1}{d_r(\tau^* + 1)}} \cdot c_r(\tau^* + 1), & \delta(\tau^*) = +1 \end{cases} \quad (14)$$

For a multi-step prediction, (14) may either be recursively repeated (with $\hat{x}(t + p - 1)$ being reclassified for the prediction of $\hat{x}(t + p)$), or the implication of constant relative position $\delta(\tau^* + p)$ and classification results $\mu_{p, \tau^* + p}$ may be extended to farther sample points $p > 1$. If $\hat{x}(t + p)$ is based on only one detected pattern (cf. Sect. 4.2), both approaches lead to identical results.

4.2 To Combine or Not to Combine

If several (K) patterns are being detected and used for (different) predictions, the latter have to be combined to obtain a compromise that reflects the reliability of the individual recognition results. The center-of-gravity method is one approach to this, resembling the defuzzification step for several active fuzzy rules:

$$\hat{x}(t+1) = \frac{\sum_{k=1}^K \tilde{\mu}^{*k} \cdot \hat{x}^k(\tau^{*k} + 1)}{\sum_{k=1}^K \tilde{\mu}^{*k}} \quad (15)$$

One has, however, to decide application-specifically if such a combination does make sense from a semantic point of view: May different patterns be “active” at the same time? This is especially questionable if the pattern knowledge was gained through motif mining algorithms like [6], which almost always assume only one active pattern at a time. Common sense would probably opt for basing a prediction only on the latest information available. Figure 4 sketches a case where it indeed seems advisable to discard older results and solely use the best and most recently recognised pattern for a local prediction. In other cases, however, (15) may just as well lead to (quantitatively) better results, despite the fact that such a combination might (semantically) not be well justified.

Fig. 4. Pattern-based prediction with varying horizons. From time t_2 , a prediction only based on x^A would yield better results than a combination of both patterns.

4.3 Selection of a Suitable Method

While the prediction method (I) in Sect. 4.1 may be described as cautious or conservative, only reproducing the knowledge about a pattern’s mean course, method (II) is, in principle, able to extrapolate to the entire universe of discourse. When should we select which of these methods?

This question ultimately leads to another, more philosophical question regarding the fuzzy sets describing a pattern: What is the source of the uncertainty and fuzziness represented by the parameters b , c and d in (1)? In time series datasets, we encounter different phenomena that may help answering our initial question. To mention two examples: In Fig. 5a, instances of the pattern and the mean course run mostly in parallel—the instances “breathe”. Entirely different is especially the second half of the pattern in Fig. 5b, where the instances exhibit a large amount of high-frequency noise added to the mean course. In the latter case, the cautious prediction method (I) would—on average—deliver better results, whereas method (II) is suited to extrapolate “breathing” instances of the pattern in Fig. 5a.

Fig. 5. Different reasons for uncertainty in time series patterns (data from [4]).

5 Example and Conclusion

In the random time series displayed in Fig. 6a, five randomly chosen instances from each of the two patterns in Fig. 5 were embedded in alternating order, with additional slight noise added. To complicate the recognition process, both pattern ensembles were made more similar by normalisation and resampling to the same mean, variance and length (100 samples). The random parts between the instances were filtered and scaled to closely match the patterns’ properties and shapes. For the recognition and prediction, fuzzy models as described in Sect. 2.2 were learned from the datasets, and a fuzzy window of interest formulated such that pattern instances should be at least half-way completed.²

² For $\mu_w(\tau)$, the membership function (1) was employed and parameterised with $r = 60$, $c_l = 10$, $c_r = 40$, $b_l = 0.5$, $b_r = 0.7$, $d_l = 4$, $d_r = 2$ to obtain high values μ_w for $\tau > 50$ and a rather steep decrease of interest for lower values of τ .

(a) Random time series with embedded pattern instances (marked black)

(b) Results $\tilde{\mu}^*(t)$ of the fuzzy decision (black: 'Coffee' pattern, grey: 'FaceAll')

(c) Embedded (grey) and crisp decision on detected patterns (black)

Fig. 6. Example for the recognition of patterns embedded in a random time series.

The results $\tilde{\mu}^*(t)$ of the decision step based on windowed recognition results $\tilde{\mu}(t, \tau)$ for both patterns are presented in Fig. 6b. Coming to a crisp decision on one active pattern and applying a threshold of $\tilde{\mu}^*(t) \geq 0.5$, we can see from Fig. 6c that the embedded instances are recognised reliably throughout the time series, each—as desired—in its second half of development. Using a more sophisticated decision procedure, which is beyond the scope of this paper, additional short-lived recognition results (the spikes in Fig. 6c) could be filtered out.

To compare the prediction methods qualitatively, two sections of the time series along with different short-time predictions are displayed in Fig. 7 in higher detail. In Fig. 7a, the predictions based on method (II) yield better results, while the simple method (I) outmatches (II) in Fig. 7b. One special property of the prediction approach presented in this paper, which is also visible in Fig. 7, is that a new prediction with a different horizon may be available at each point in time. Due to this fact and that predictions may—depending on the size of the knowledge base—not be available at any time, a quantitative comparison to existing prediction approaches calls for new, suitable

performance measures. The comparability with (say: model-based) methods designed for fixed horizons, however, will always be impaired by the missing flexibility on the one (model-based) side, and the unguaranteed availability on the other (pattern-based).

Fig. 7. Local predictions of $x(t)$ (light grey) using method (I) (black) and (II) (grey).

In contrast to earlier works in similar fields [5,8], the pattern model of this article can directly employ results of motif-oriented data mining algorithms. Besides the possibility of soft windowing, subsequent work should explore further advantageous uses of the fuzzy recognition results, especially a more sophisticated decision strategy.

References

1. Bocklisch, S.F.: Prozeßanalyse mit unscharfen Verfahren. Technik, Berlin (1987)
2. Hempel, A.J., Bocklisch, S.F.: Fuzzy pattern modelling of data inherent structures based on aggregation of data with heterogeneous fuzziness. In: Rey, G.R., Muneta, L.M. (eds.) Modelling Simulation and Optimization, chap. 28, pp. 637–655. INTECH (2010)
3. Herbst, G., Bocklisch, S.F.: Online recognition of fuzzy time series patterns. In: 2009 International Fuzzy Systems Association World Congress and 2009 European Society for Fuzzy Logic and Technology Conference (IFSA-EUSFLAT 2009). pp. 974–979 (2009)
4. Keogh, E., Xi, X., Wei, L., Ratanamahatana, C.A.: The UCR time series classification/clustering homepage. http://www.cs.ucr.edu/~eamonn/time_series_data/
5. Šket Motnikar, B., Pisanski, T., Čepar, D.: Time-series forecasting by pattern imitation. OR Spectrum 18(1), 43–49 (1996)
6. Mueen, A., Keogh, E., Zhu, Q., Cash, S., Westover, B.: Exact discovery of time series motifs. In: Proceedings of the SIAM International Conference on Data Mining (SDM 2009). pp. 473–484. American Statistical Association (ASA) (2009)
7. Päßler, M., Bocklisch, S.F.: Fuzzy time series analysis. In: Hampel, R., Wagenknecht, M., Chaker, N. (eds.) Fuzzy Control: Theory and Practice, pp. 331–345. Physica, Heidelberg (2000)
8. Singh, S.: Multiple forecasting using local approximation. Pattern Recognition 34, 443–455 (2001)