Data Set A is a Pattern Matching Problem

Jens Kohlmorgen and Klaus-Robert Müller
GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany
E-mail: jek@first.gmd.de, klaus@first.gmd.de

Key words: time series prediction, benchmarking, Santa Fe Competition, pattern matching

Abstract. Several data sets have been proposed for benchmarking in time series prediction. A popular one is Data Set A from the Santa Fe Competition. This data set was the subject of analysis in many papers. In this note, it is shown that predicting the continuation of Data Set A is nothing else than a pattern matching problem. Looking at studies of this data set, it is remarkable that most of the very good forecasts of Data Set A used upsampled training data. We explain why upsampling is crucial for this data set. Finally, it is demonstrated that simple pattern matching performs as good as sophisticated prediction methods on Data Set A.

1. Learning from One Example

Data Set A from the Santa Fe Competition [9] consists of sampled values from the emission intensity of a NH_3-FIR (far-infrared) laser (see Hübner et al. [1]). Given a time series of 1000 data points, the objective of the competition was to predict the following 100 points of the data and to estimate the error bars of the prediction. In Fig. 1, the time series and its true continuation, starting at $t = 1000$, is shown ($t = 0, ..., 1200$). In order to predict 100 or more data points after $t = 1000$, the collapse of the intensity at $t = 1060$ has to be predicted properly. A similar collapse can be found in the training data at $t = 605$. In fact, the sequence 545-617 in the training data is *almost identical* to the sequence 1000-1072, i.e. the first 73 points of the continuation (see Fig. 2). Moreover, there is no other collapse in the training data that is similar to the one in the continuation. Thus, there is only *one* example in the training set that represents the collapse to be predicted. Therefore, a learning method that should succeed in predicting the collapse in the continuation has to learn the sequence 545-617 *by heart*, and has to reproduce it exactly at $t = 1000$.

2. How to predict Data Set A

This question can be reformulated: how to find out, that the sequence 545-617 is a good forecast at $t = 1000$ without knowing the true continuation? If we just consider the training data (Fig. 1, $t = 0, ..., 999$), we can infer that

three continuations are reasonable. These three cases are characterized by an increasing amplitude, which fits to the data sequence at the end of the training set (see also Lendaris & Fraser [3] and Kostelich & Lathrop [2] for a discussion). In Fig. 3, these three sequences are aligned with the true data sequence at the end of the training set ($t = 960, ..., 999$). It turns out that all of them fit very well at least the last 30 points (4 periods) of the training set. Therefore, they all make perfect sense as continuations. However, only one of the sequences includes the collapse of the true continuation, which occurs at $t = 1060$ (Fig. 4). Interestingly, just this sequence does not coincide with the data points before $t = 970$, whereas the others still do.

How then is it possible for a prediction model to choose the right continuation out of those three, very similar cases? Surprisingly, the best matching sequence for e.g. the last 25 data points of the training set, with respect to the euclidian distance, is *not* the one whose continuation fits best to the true continuation. More surprisingly, this also holds for any other number of data points taken from the end of the training set: there is always another pattern sequence in the training set that fits better to the end of the training data than a sequence whose continuation is 545-617. Thus, at first glance it seems that there is no reason why a predictor should choose the 'right' sequence 545-617 as forecast.

The solution to this dilemma is as follows: to resolve the fine differences in waveform and phase between all reasonably fitting training sequences (and there are even more than three), these sequences have to be sampled at a higher rate. Thus, if the training data set is upsampled by a factor of 10 (by filling in points using an interpolation technique, as suggested by Sauer [5]), and the data sequences to be compared include the upsampled data points, then the desired training sequence with continuation 545-617 is actually the best match for the last 13 or more (up to 24) data points of the original training set. 13-24 points include 1-3 periods of the waveform, which represents a reasonable range of sequence lengths resp. embedding dimensions for Data Set A. However, larger sequence sizes yield again other best matches.

3. Comparing Pattern Matching with Other Methods

Several elaborate predictions [4, 5, 6, 7, 8] have been carried out for Data Set A so far. The best forecast within the Santa Fe Competition was achieved by Wan [7], who assembled a network of Finite Impulse Response (FIR) linear filters. The network basically reproduced the training set sequence 545-619 as forecast for the first 75 time steps of the continuation of Data Set A at $t = 1000$. After 75 steps (after the predicted collapse), the prediction 'deteriorated' and 'the remaining 25 points were selected by adjoining a

similar [i.e. plausible] sequence taken from the training set' (cf. [7]). The normalized mean squared error (NMSE) for this 100-point prediction was NMSE(100) = 0.0273. Using the above pattern matching approach on the training data, and consequently taking the training set sequence 545-644 as forecast, yields a NMSE(100) of 0.0265.

As far as we know, the best result so far was obtained by Weigend & Nix [8] after the competition. This result was achieved by taking the *mean* of the post-collapse oscillations between $t = 619$ and 699 as (constant) forecast for all data points after $t = 1072$. Instead of predicting the (wrong) phase given in the training data (Fig. 2), this prediction strategy obviously improves the test set error between $t = 1073$ and 1099. Weigend & Nix reported a NMSE(100) of 0.016. If we replace the last 27 data points in the sequence 545-644 by their mean and use this as prediction, we get a NMSE(100) of 0.0137.

In both cases, the more sophisticated prediction method yields a result that is very similar to the simple pattern matching approach. This underlines the necessity to reproduce the training data starting at $t = 545$.

4. Conclusion

As shown above, a successful forecast of Data Set A requires an accurate distinction between several, very similar choices. All of them make perfect sense as continuations. However, because of the chaotic nature of the signal, all these continuations diverge after approx. 40 time steps, i.e. $t = 1040$ (Fig. 4). Only one sequence includes the collapse of the true continuation at $t = 1060$. Therefore, a prediction model has to learn this particular training sequence by heart, and has to reproduce it at $t = 1000$. To resolve the fine differences between all plausible continuations and to find the right solution, some tricky effort, as e.g. upsampling, is necessary.[1]

Learning the training data by heart is not desirable in most prediction/regression tasks, because then the prediction model is likely to overfit, in particular it fits the noise in the data. In fact, a regression model that generalizes well is not expected to reproduce the training data exactly. Yet, this is required for Data Set A. This requirement, and the good performance of simple pattern matching, leads us to the conclusion that Data Set A is rather a benchmark for exact reproduction of training data than a benchmark to test the generalization ability of a prediction method (see also the discussion in [2, 6]). However, the participants of the Santa Fe

[1] Another 'trick' is to use only the first 900 data points for training/matching, as e.g. in [7], where the last 100 points were used for validation. This restriction yields the right match at least for some particular embedding dimensions.

Competition did not know the true continuation in advance, and therefore could hardly come to this conclusion.

References

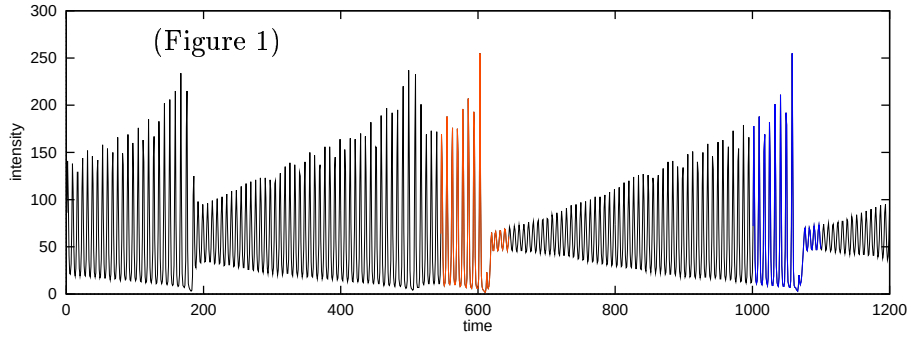
1. Hübner, U., et al., Lorenz-Like Chaos in NH_3-FIR Lasers (Data Set A), in [9], pp. 73–104, 1994.
2. Kostelich, E.J., Lathrop, D.P., Time Series Prediction by Using the Method of Analogues, in [9], pp. 283–295, 1994.
3. Lendaris, G.G., Fraser, A.M., Visual Fitting and Extrapolation, in [9], pp. 319–322, 1994.
4. Molina, C., Niranjan, M., Pruning with Replacement on Limited Resource Allocating Networks by F-Projections, *Neural Computation* 8, pp. 855–868, 1996.
5. Sauer, T., Time Series Prediction by Using Delay Coordinate Embedding, in [9], pp. 175–193, 1994.
6. Smith, L.A., Does a Meeting in Santa Fe Imply Chaos?, in [9], pp. 323–343, 1994.
7. Wan, E.A., Time Series Prediction by Using a Connectionist Network with Internal Delay Lines, in [9], pp. 195–217, 1994.
8. Weigend, A.S., Nix, D.A., Predictions with Confidence Intervals (Local Error Bars), in: Proceedings of the International Conference on Neural Information Processing (ICONIP'94), Seoul, Korea, pp. 847–852, 1994.
9. Weigend, A.S., Gershenfeld, N.A. (eds.), *Time Series Prediction: Forecasting the Future and Understanding the Past*, Addison-Wesley, 1994.

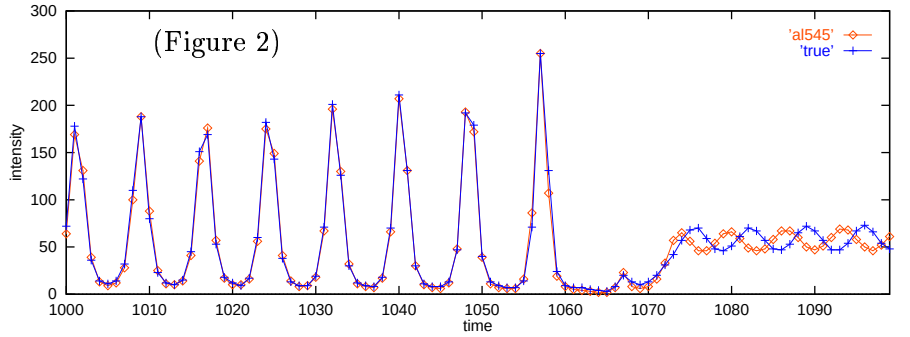
Figure 1. 1200 points of Data Set A from the Santa Fe Competition. The first 1000 data points are the training set ($t = 0, ..., 999$). Only these points were given to the participants of the competition. The continuation, starting at $t = 1000$, is to be predicted. It was only available after the competition. It turns out that the sequence 545-644 of the training set (red) is very similar to the sequence 1000-1099 of the continuation (blue).

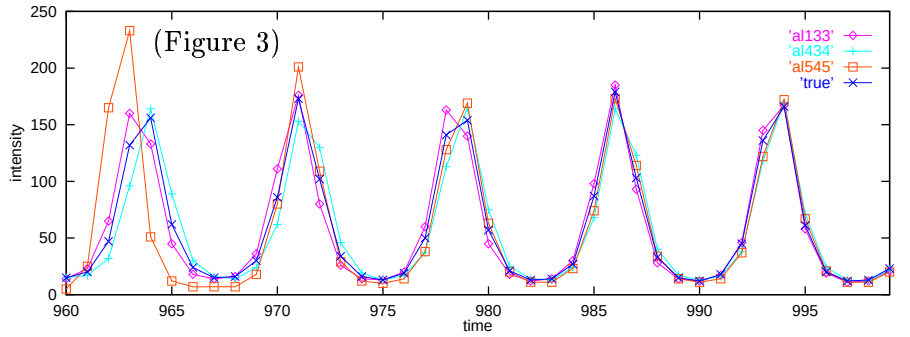
Figure 2. The sequence 545-644 of the training set (al545) aligned with the sequence 1000-1099 of the continuation (true). Clearly, they are almost identical for the first 73 points.

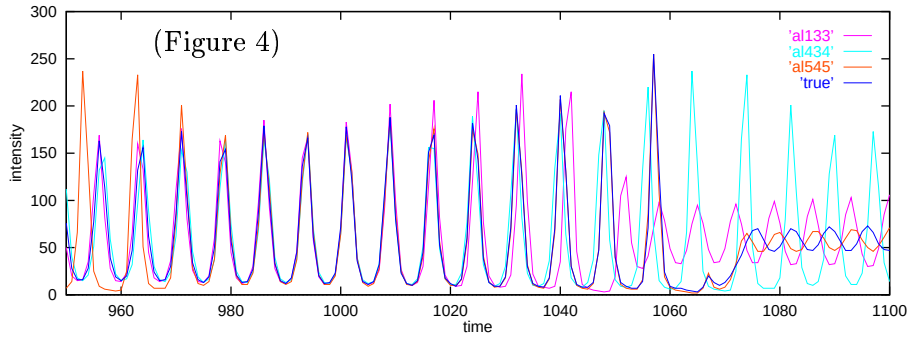
Figure 3. Three sequences from the training set that fit very well with the data points (true) at the end of the training set ($t = 960, ..., 999$). Note, that the sequence (al545), whose continuation is an excellent prediction for Data Set A, does not coincide with the data points before $t = 970$.

Figure 4. The sequences from Figure 3 including their continuations ($t = 950, ..., 1100$). Only one sequence (al545) includes the collapse of the true continuation, which occurs after $t = 1060$.


(Figure 1)
intensity
time
0
50
100
150
200
250
300
0
200
400
600
800
1000
1200


(Figure 2)
'al545'
'true'
intensity
time
0
50
100
150
200
250
300
1000
1010
1020
1030
1040
1050
1060
1070
1080
1090


(Figure 3)
'al133'
'al434'
'al545'
'true'
intensity
time
0
50
100
150
200
250
960
965
970
975
980
985
990
995


(Figure 4)
'al133'
'al434'
'al545'
'true'
intensity
time
0
50
100
150
200
250
300
960
980
1000
1020
1040
1060
1080
1100