

Temporal Extrapolation within a Static Clustering¹

Tim K. Cocx

Walter A. Kusters

Jeroen F.J. Laros

Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands

1 Introduction and background

Predicting the behaviour of individuals is a core business of policy makers. This paper discusses a new way of predicting the “movement in time” of items through pre-defined classes by analysing their changing placement within a static, preconstructed 2-dimensional clustering. It employs the visualization realized in previous steps within item analysis, rather than performing complex calculations on each attribute of each item. For this purpose we adopt a range of well-known mathematical extrapolation methods that we adapt to fit our need for 2-dimensional extrapolation.

It is common practice to visualize analysis results as a clustering within the 2-dimensional plane, approximating the correct, multi-dimensional distances. As long as the error made within the visualization is as small as possible there is no preferred method of obtaining such a clustering within our prediction system. However, the algorithm must allow for incremental addition of single elements from the sequence to be extrapolated, which is not supported by all of the known methods.

All interpolation schemes are suitable starting points for the process of extrapolation. In most cases, it is sufficient to simply continue the fabricated interpolation function after the last existing data point. In the case of a spline interpolation, however, a choice can be made to continue the polynomial constructed for the last interval (which can lead to strange artifacts), or extrapolate with a straight line, constructed with the last known derivative of that polynomial (Figure 1 and 2). In our approach both x and y are coordinates and therefore inherently independent variables. They depend on the current visualization alone. Within our model, they do however depend on the time variable t . Because our methods aim to extrapolate x, y out of one other variable t , we need a form of 2-dimensional extrapolation.

2 Approach

The data used as reference within our approach is represented by a square $q \times q$ distance matrix describing the proximity between all q items. These items are clustered and visualized in a 2-dimensional plane with dots representing our reference items. This step in the approach is done only once so the focus should be on the quality of the clustering instead of the computational complexity. From this point on this clustering is considered to be fixed or static.

Analysis of the behaviour of new items should start with the calculation of the attributes for each time-unit. These units are supposed to be cumulative, meaning that they contain all the item’s *baggage*, i.e., its whole history, up to the specified moment. Using the same distance measure that was used to create the initial distance matrix, the *distance vector per time-unit* can now be calculated. This should be done for all t time-units, resulting in t vectors of size q . These vectors can now be visualized as before.

After selecting the best extrapolation scheme for our type of data our method creates a function that extrapolates item behaviour. One advantage of this approach is that the extrapolation or prediction is immediately visualized to the end-user rather than presenting him or her with a large amount of numerical data.

¹Published in: Foundations of Intelligent Systems, Proceedings of ISMIS 2008, LNAI 4994, pp.189–195, Springer 2008.

This research is part of the DALE (Data Assistance for Law Enforcement) project as financed in the ToKeN program from the Netherlands Organization for Scientific Research (NWO) under grant number 634.000.430.

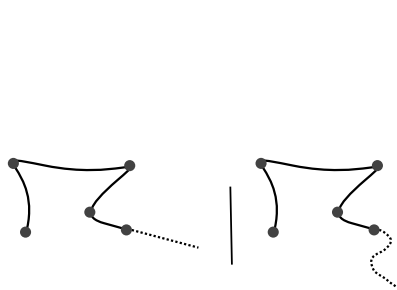


Figure 1: Straight line extrapolation (left) and polynomial continuation (right)

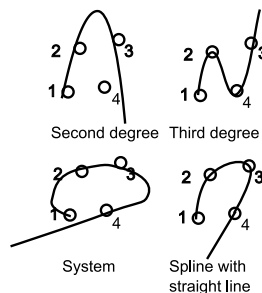


Figure 2: Different extrapolation methods yield very different results

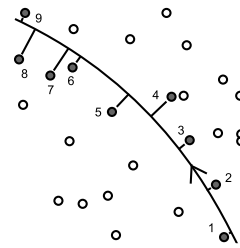


Figure 3: Selecting points with the shortest distance to the extrapolation line

If the user is familiar with the data under consideration, he/she can analyse the prediction in an eye blink. Augmenting the system with a click and point interface would enable the end-user to use the prediction as a starting point for further research. In most cases it is desirable to predict which class the item under consideration might belong to in the future. In that case it is important to retrieve further information from some of the reference items and assign future attribute values and a future class to the item.

A first step would be to select r reference items closest to the extrapolation line. This can easily be done by evaluating the geometric distance of all reference points to the line and selecting those with the smallest distance, see Figure 3. We order these points by their respective distance to the last known data point of the extrapolated item: the confidence of the prediction declines with this distance and calculate the expected future attributes of the item under consideration based upon this weighted average. The extrapolated item can now be visualized into the clustering according to its future attributes and be classified accordingly.

3 Experiments

The detection, analysis, progression and prediction of criminal careers is an important part of automated law enforcement analysis [1]. Our approach of temporal extrapolation was tested on the national criminal record database of The Netherlands. This database contains approximately one million offenders and their respective crimes (approximately 50 types). For each item (i.e., person) in the test set we only consider the first $t = 4$ time periods. The accuracy is described by the mean similarity between the calculated and the expected values of the attributes.

Although the runtime needed for visual extrapolation is much less than that of regular methods, the accuracy is comparable. For this database the best result is the spline extrapolation with a straight line, having a very short runtime while reaching an accuracy of 88.7%.

4 Conclusion and Future Directions

In this paper we demonstrated the applicability of temporal extrapolation by using the prefabricated visualization of a clustering of reference items. We demonstrated a number of extrapolation techniques and employed them to predict the future development of item behaviour. Our methods were tested within the arena of criminal career analysis, predicting the future of unfolding criminal careers. We showed that our novel approach largely outperforms standard prediction methods in the sense of computational complexity, with a loss in accuracy smaller than 1 percentage point. The visual nature of our method enables the analyst of the data to immediately continue his/her research since the prediction results are easily displayed within a simple graphical interface.

References

[1] J.S. de Bruin, T.K. Cocx, W.A. Kusters, J.F.J. Laros, and J.N. Kok. Data mining approaches to criminal career analysis. In *Proceedings of the Sixth IEEE International Conference on Data Mining (ICDM 2006)*, pages 171–177, 2006.