Survey: Visual Analysis Approaches to Time Series Prediction

Fabian Otto[*]

Technical University Darmstadt

Abstract

1 Introduction

Predictive analytics is concerned with the prediction of future probabilities and trends based on observed events. It encompasses a multi-perspective approach that includes integrated reasoning, pattern recognition and predictive modeling associated with domain knowledge. However, the primary use of such systems tends to be reactive, meaning that analytic systems are typically.

The increasing availability of digital data provides both opportunities and challenges. The potential of utilizing these data for increasing effectiveness and efficiency of operations and decision making is vast. Harnessing this data with effective tools can transform decision making from reactive to proactive and predictive. However, the volume, variety, and velocity of these data can actually decrease the effectiveness of analysts and decision makers by creating cognitive overload and paralysis by analysis, especially in fast-paced decision making environments.

In spatiotemporal data, analysts are searching for regions of space and time with unusually high incidences of events. In the cases that hotspots are found, analysts would like to predict how these regions may grow in order to plan decision support and preventative measures. Furthermore, analysts would also like to predict where future hotspots may occur.

Four stages:

1. data cleaning,
2. feature selection,
3. modeling,
4. and validation

2 Temporal Time Series

content

2.1 Simple Time Series (maybe just use Sect. 2.2, check what makes more sense)

2.2 Pattern and Anomaly Detection

The easiest kind of prediction can be found with simple time series, they are only based on a single variable and include one single history.

TS is also multivariate.

An early work with simple time series can be found from Buono et al. [3]. They built upon the TimeSearcher system proposed by Hochheiser and Shneiderman [5], which focuses explicitly on high usability even for users without specialized skills such as statistics. Based on their idea, Buono et al. continued with using timeboxes, i.e. rectangular regions that filter the data and reduce the scope of each search. Their updated version, extends the original timeboxes with another variant for pattern search in the remaining data. Further, they also allow users to work with long time series of multiple heterogeneous variables. For data exploration the user is initially presented with a multivariable time series viewer, which allows to visually analyze multiple variables in parallel in different levels of detail. In order to detect pattern, the systems allows the user to highlight a pattern within the time series to search for. Therefor, the search space can be limited to only interesting areas. The initial matches of the given pattern can afterwards be refined to support the goal better. This process of returning many results and consequently enable the user to reduce the outcome to his/her liking, can also be found in spatiotemporal approaches (see Sect. 3), esp. in the template driven approach of [11]. Moreover, TimeSearcher also provides different data transformations to match patterns more easily. One drawback of the system is that it cannot deal with missing data points and pattern matching still remains very depended on the user chosen parameters.

This second version of TimeSearcher was updated once more by Buono et al. [4]. It focuses mainly on a similarity-based data-driven forecast, which uses line charts for visualization. To avoid the overlapping issue for an increasing number of items, the system offers a summarized view based on a river plot, which also displays a confidence bands. This data-driven approach extrapolates the time series based on a similarity search of past events to predict future events. As in [3] the system is meant for users without statistical knowledge. For that reason, a simultaneous preview interface allows to compare multiple parameter choices in parallel. Given those ideas, some drawbacks of have to be considered as well. Data-driven approaches normally require larger datasets compared to a model-driven approach. Furthermore, the system's prediction output still relies heavily on the choices of an untrained user.

Another extension of an existing system is from Bögel et al. [1], who built a prediction functionality on top of their earlier work [1]. Unlike [3–5] the system is model-driven, and therefore it requires less data points. Originally, the system was designed after the Box Jenkins method [2] and meant for model selection, however, during their evaluation, they found that an actual prediction functionality would also improve the model selection process. While being able to adjust different model parameters (e.g. for Autoregressive, Moving Average or Autoregressive integrated moving average models), the prediction visualizations are changing in parallel. As visualizations, the system provides, similar to [4], line charts with corresponding confidence bands. Additionally, they specifically visualize the difference of true and predicted values for each data point as well as the direction (positive or negative). This gives the analyst a quick overview if the model is constantly over- or underestimating the time series and how long and often this occurs.

Another system can be found from Steed et al. [12]. Their approach is focused on understanding patterns in log and imagery data collected by 3D printers.

2.3 Time Series Prediction with Peak Preservation

content

2.4 Multivariate Time Series

Visual Analytics approaches for simple time series prediction (Sect. 2.1) offer a great variety for single time series. However, in a practical environment it is often necessary to evaluate multiple

[*] e-mail: fabian.otto@stud.tu-darmstadt.de

time series in parallel or deal with multidimensional input variables. In order to solve these problems, visual analytics approaches for simple time series are not sufficient anymore.

An older Visual Analytics approaches for multivariate time series prediction is from Ichikawa et al. [6]. Their goal was to predict multiple daytime stock prices and simultaneously visualize a set of prediction from different simulation systems within a single system. Therefor, their system supports multiple prediction for a single stock as well as predictions of multiple stocks. One major finding was, visualizing multivariate predictions in a 3-dimensional space creates large amounts of occlusion, thus it is not suitable. Instead the system utilizes line charts with cluttering control and color charts with level-of-detail control. The line charts support multivariate scaling, i.e. the time axis of the plot can be scaled differently to focus one certain areas, e.g. with less overlap. Further, the lines are colored differently, depending on the amount of overlap/uniform predictions. The color chart is composed of horizontal color-bands for each prediction, whereas a vertical color band can be seen as a period in time. In order to improve the visualization, the different color-bands are cluster based on the user-specified level-of detail. This results in a smaller amount of horizontal color-bands where the individual properties are diminished. In the workspace the above plots can be displayed with different axis. This enables the user to compare a set of predictions for different parameter ranges (e.g. sales organizations) as well as different stocks simultaneously. Consequently, the user can get a overview or trend, but due to clustering and simply the amount of predictions displayed, he can hardly extract specific information.

3 Spatiotemporal Time Series

The previous section (Sect. 2) presented an overview of systems for temporal time series prediction as well as pattern and anomaly detection. However, in other application areas such as crime prevention as well as emergency and epidemic intelligence, not only the temporal data is of value, but also locations. Typically, these locations fit into a hierarchical categorization structure, which can be filtered. Further, the data categories are processed either as aggregated time series over a spatial location (e.g. county, zip code, collection station) or represent a spatial snapshots of a small time aggregate (e.g., day, week).

One system which focuses on spatiotemporal prediction is the model-driven approach from Maciejewski et al. [7]. It is based on their previous work [9, 10] and centers around on categorical geospatiotemporal event data, where events consist of locations in time and/or space, where each event fits into a hierarchical categorization structure. Specifically, they used a data set for detecting adverse health events using pre-diagnosis information from emergency departments. The system itself provides a line chart with certainty bands and colored geospatial window, which shows the percentage of events in a certain area, e.g. patients at an emergency department, which where classified with respiratory syndromes. Further, the user is able to apply filtering on a fine and coarse-grained level. The systems also differentiates between the time series and the geospatial prediction. The time series prediction is achieved by cumulative summation or a Seasonal Trend decomposition based on locally weighted regression. For multivariate data, each event category is modeled as a separate time series signal. Equivalent to the time series prediction, the granularity/level of aggregation (e.g. state, county, etc.) for geospatial predictions can be adjusted by the user. Further, the system provides a Kernel Density Estimation approach which allows to display the spatiotemporal distribution on a more fine-grained level. In order to detect anomalies, e.g. outbreaks, (also see page 1) the system calculates the difference between the predicted and the actual values and highlights areas above a user specified threshold. One system which focuses on spatiotemporal prediction is from Maciejewski et al.

Another system also from Maciejewski [8] follows a similar approach. The goal is to handle data, which contains multiple variables, high signal to noise ratio and a degree of uncertainty. Equivalent to [7] it also provides a linked environment of geospatial data and time series graphs and allows users to filter data. Additionally, the system also uses cumulated summation and Kernel Density Estimation. The major difference is, this work focuses on finding and understanding the pattern, rather than only predicting them. Therefor, the system establishes temporal contour maps, which are overlaid contour maps over period of time. This allows the user to view shifting hotspots across time and analyze the movements of trends and patterns over this period. Further, the system allows to search for correlations between multiple variables via overlaying contour maps, heat maps and/or including height. However, one issue with this visualization is, it only works with three different variables and cannot help to find larger correlations. Moreover, aggregating too many data points, may yield to largely exaggerated hotspots or a uniform surface, through too many hotspots.

Recent work from Malik [11] focuses on a data-driven visual analytics approach that provides domain experts (i.e. non statistic experts) a proactive and predictive environment, which allows them to utilize their domain expertise. Similar to [7] the system applies Seasonal Trend decomposition based on locally weighted regression and Kernel Density Estimation for its predictions. One issue they identified was that domain experts need additional guidance in order to improve their analysis. Hence, they provide geospatial and temporal scale templates, this presents the users with a starting point and avoids searching over the complete parameter space. For the geospatial template, the system separates the space in subregions and filters for regions, which show a high predicted activity and provide sufficient data. Further, the system allows the user to interactively change the initial template to include e.g. police beats and avoid zero counts with no predictive statistical value. In order to compensate for small amounts of data points in some regions, the system can make use of demographically similar neighborhoods within a certain radius by averaging their prediction. Temporal templates follow the motivation of peak preservation (see page 1) as trends can occur on a hourly basis as well. Therefore, the system provides a clock view to highlight high activity on a hourly granularity. Moreover, it enables the user to filter on daily and monthly basis to detect such patterns.

4 Conclusion

content

References

[1] Visual analytics for model selection in time series analysis. *IEEE Transactions on Visualization and Computer Graphics, Special Issue "VIS 2013"*, 19, 2013.

[2] G. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976.

[3] P. Buono, A. Aris, C. Plaisant, A. Khella, and B. Shneiderman. Interactive pattern search in time series. In *Visualization and Data Analysis 2005*, vol. 5669, pp. 175–187. International Society for Optics and Photonics, 2005.

[4] P. Buono, C. Plaisant, A. Simeone, A. Aris, G. Shmueli, and W. Jank. Similarity-based forecasting with simultaneous previews: A river plot interface for time series forecasting. In *Information Visualization, 2007. IV'07. 11th International Conference*, pp. 191–196. IEEE, 2007.

[5] H. Hochheiser and B. Shneiderman. Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, Mar. 2004. doi: 10.1145/993176.993177

[6] Y. Ichikawa, T. Tsunawaki, I. Fujishiro, and H. Yoon. A visualization environment for multiple daytime stock price predictions. In *Proceedings of the 2nd VIIP International Conferences on Visualization, Imaging and Image Processing, Malaga, Spain*, 2002.

[7] R. Maciejewski, R. Hafen, S. Rudolph, S. G. Larew, M. A. Mitchell, W. S. Cleveland, and D. S. Ebert. Forecasting hotspots: A predictive analytics approach. *IEEE Transactions on Visualization and Computer Graphics*, 17(4):440–453, April 2011. doi: 10.1109/TVCG.2010.82

[8] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert. A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, 16(2):205–220, 2010.

[9] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, M. Wade, and D. S. Ebert. Understanding syndromic hotspots - a visual analytics approach. In *2008 IEEE Symposium on Visual Analytics Science and Technology*, pp. 35–42, Oct 2008. doi: 10.1109/VAST.2008.4677354

[10] R. Maciejewski, B. Tyner, Y. Jang, C. Zheng, R. V. Nehme, D. S. Ebert, W. S. Cleveland, M. Ouzzani, S. J. Grannis, and L. T. Glickman. Lahva: Linked animal-human health visual analytics. In *2007 IEEE Symposium on Visual Analytics Science and Technology*, pp. 27–34, Oct 2007.

[11] A. Malik, R. Maciejewski, S. Towers, S. McCullough, and D. S. Ebert. Proactive spatiotemporal resource allocation and predictive visual analytics for community policing and law enforcement. *IEEE transactions on visualization and computer graphics*, 20(12):1863–1872, 2014.

[12] C. A. Steed, W. Halsey, R. Dehoff, S. L. Yoder, V. Paquit, and S. Powers. Falcon: Visual analysis of large, irregularly sampled, and multivariate time series data in additive manufacturing. *Computers & Graphics*, 63:50–64, 2017.