PredVis: Interaction Techniques for Time Series Prediction

Sakshi Sanjay Pratap*
Georgia Institute of Technology

Alex Endert†
Georgia Institute of Technology

Abstract

With the increasing collection of time series data, both related to business and personal use, a substantial amount of research and development efforts are being directed to gain deeper insights from such records. Data mining techniques like similarity search and segmentation are used as tools to enhance the comprehension of this data. While innovative techniques have been examined, less work has been done in creating functional tools that use these methods for visual analysis. We have created PredVis, a mixed-initiative system that uses statistical and machine learning techniques to facilitate visual analysis of time series data for forecasting. The system provides tools to interact with model calculated results, visually query variances and integrate other dimensions of the data to improve user comprehension and decision-making. The methods developed in the system are generic and can be employed for analysis in domains including finance, health, marketing and astronomy.

Index Terms: time series; prediction; visual analytics;

1 Introduction

With the extensive recording of events, there is an increased interest in mining time series data and predicting trends and events [4]. A relatively recent development is the use of mixed-initiative interfaces, systems that integrate computational assistance for data analysis tasks, to create visual predictions. Studies have shown that interactive refinement of machine learning tasks can result in better user experiences and more effective learning systems [1]. Our work strives to develop effective techniques that enable users to use their knowledge in conjunction with the advantages of the prevalent analysis techniques and generate predictions that are more effective and interpretable.

TimeSearcher3 [3] introduced a similarity-based data-driven forecaster that provides interactive feedback related to the parameters used in the forecasting. Hao et al. [5] integrated multiple prediction and similarity-based models. They used a visual analytics approach for peak-preserving prediction of large seasonal time series, that can be steered by the users. TimeFork [2] presented a technique for visual prediction of multivariate time series that considers inter-variable relationships. While several techniques have been examined, domain independent visual analysis of time series prediction still remains a challenge with huge potential for further improvement.

PredVis provides seamless integration of existing machine learning methods and visualization techniques with human knowledge to enhance time series predictions. It is common for an analyst to have information that the model might not be able to account for. For example, in the study of product sales, a remote spike after an election could be miscalculated by the model as a data anomaly whereas an informed user will be able to correctly identify the pattern. Alternatively, the user's inaccurate assumptions can be more easily caught by computational analysis. The system is designed to utilize techniques that support effective exploration of single or multiple time series data records. Predictions are calculated on a single series once the desired record has been selected. By providing features such as series similarity search, models' historic performance evaluation and event integration, we create a fully functional prediction system.

2 System Techniques

PredVis currently uses two state of the art algorithms - Long Short Term Memory [6] (LSTM), a variant recurrent neural networks and Support Vector Regression [7] (SVR) to generate the predictions. Though the predictions are initialized using these models, the system is flexible to integrate most standard models. The user can analyze and incrementally update these predictions. Our goal is to enable users to intuitively make more complete estimates by using simple and effective interactions. Below, we describe the user interface, model, visualization and interaction strategies.

2.1 User Interface

The interface has 3 main components: uploading data, record selection and visual analysis. The user begins the analysis process by uploading simple (.csv) files, each corresponding to a time series record with columns for the time stamps and values. The system then processes and stores these records which are easily searchable based on their names or past trends. For example, if a user is interested in stocks that have decreased by at least 10% in the last year or areas whose electricity consumption has increased by 5% in the last two months, the user can simply adjust the parameters to find the matching records. In the next step, the user analyzes and interacts with the results of the model's predictions. To represent the historical data and forecasts, we use an enhanced time series graph, a commonly-used representation. To provide a more comprehensive view of the data, the system has the provision for integrating event data in addition to the time-series records. This is essentially a csv file containing timestamps and corresponding events (news articles, social media posts, etc.). The event counts are then displayed as clickable histograms on top of the original time series graph as seen in Fig. 1. Clicking any event bar provides details of all the events for that timestamp.

2.2 Prediction Modeling

To initialize the system, we use two independent models (LSTM and SVR) that calculate the predictions for a given record. These models are trained on the historical time period selected and for the desired time required. LSTM network, is a recurrent neural network that is trained using Back-propagation through time. We use a window size of 7 for prediction and 2 epoches to train our data. We develop a simple version of SVR where the next state is constructed using a state vector consisting of the previous 7 values. These parameters are configurable and adjusting them could give better predictions for certain datasets.

For calculating a single predicted value, the previous available values are fed as input to the model. For longer sequences, this input is incrementally shifted and predictions are calculated using the previous predicted data in conjunction with the actual historical data. After a point, the forecasting model would only contain predicted records as input which might result in decreased accuracy.

*e-mail: spratap7@gatech.edu
†e-mail: endert@gatech.edu

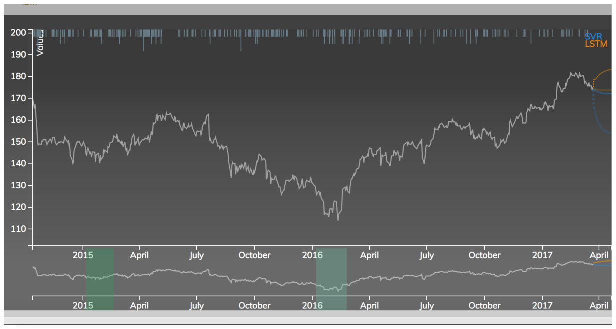


Figure 1: Using weighted regions and event integration. The figure shows the initial and weighted predictions for the two models. The intensity of a region is representative of its weight. The height of the bars above the graph represents the count of events on that day.

Analyzing results of multiple algorithms and a large amount of training data helps in providing a broader perspective and more reliable predictions.

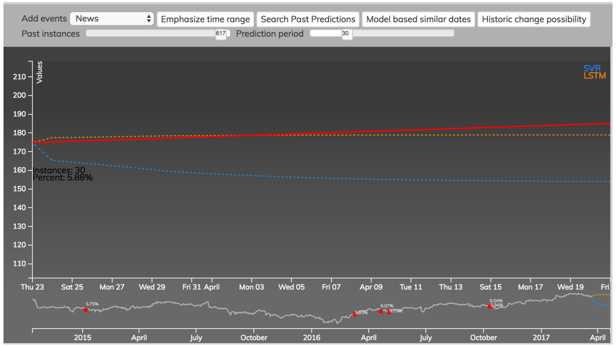
2.3 Model input and Weight Adjustment

Once a record is selected, users can interactively control the data values used to create the model as well as the time range to forecast for. The system also uses the concept of weighted prediction wherein users can select time intervals and give them weights to emphasize that area is of higher importance than the rest of the data. This is useful when users might have additional knowledge about the record or effects of past events. For example, recent data or seasonal records might be more relevant in calculating predictions in some cases. By adjusting the weights, users can steer the predictions making them more representative of the actual state. Observing the model's original prediction and the weighted prediction, users can then dynamically update the weights of any region and see how the results vary. The underlying model is an ensemble of models created using predictions from the individual regions and weighing them according to the user defined weights. Fig. 1 shows the selection and assignment of weights of 2 timeframes.

2.4 Assumption and Model Evaluation

A user might postulate that the product sales will increase by 10% in the next one year. She could evaluate this by adjusting his expected percentage changes in the UI. The system then highlights all the points in time, if any, when similar changes have occurred. This helps the user gauge if his assumed change is feasible and insight into the frequency and conditions for it to occur. Fig. 2 shows the user testing her assumption and receiving feedback from the system.

Evaluating the performance of a model is particularly important for the user. The user can see how the algorithms would have predicted at a given time instance by selecting that point in the graph and observing the results of the different models. Considering the model's dynamic nature and the time required to compute the predictions, we have constrained the user to select past instances one at a time. When calculating the predictions for historical periods, the model is trained on data until that date. Another method for the user to evaluate the model is by using Sequence Matching. The system provides the option to find the data sequences that are most similar to the 2 predicted lines generated by the models. It uses dynamic time warping to calculate these subsets from the overall timeseries. The user could then analyze the events or conditions during that period to get further insight.


Add events News
Past instances 617
Prediction period 30

Instances: 30
Percent: 5.88%

Figure 2: Assumption evaluation: Here the user is searching for instances where the price of IBM's stock has increased by 5.9% in 30 days by adjusting the red line. The system in turn highlights all such instances in the overview panel

3 Conclusion

One of the main drawbacks of many prediction systems is the lack of user feedback to the system, resulting in predictions that are less interpretable and may miss important domain expertise that can be provided by users. Our system uses advances from visualization and machine learning research to support this task. By allowing the users to steer the prediction results and analyze the effects of each of these incremental changes, they can better comprehend the underlying workings and data used for generating the model and subsequently make better predictions. Our preliminary tests using stock market data from Yahoo Finance and news articles from NYT are very encouraging. We hope comprehensive user testing will help us refine our system and make it production ready.

Techniques introduced in PredVis are just some of the methods that can be applied to effectively create intelligent interfaces. Visual Prediction Systems can employ a number of data mining techniques like dimension reduction, sub-sequence searching as well as domain specific interactions to help the user carry out more complex tasks. In our next phase, we will study more such techniques that can augment the analytic process. While there are many advantages of interactive systems, some of challenges in using human directed models like the introduction of bias, lack of effective evaluation techniques and balancing speed-accuracy trade-offs still remain.

References

[1] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014.

[2] S. K. Badam, J. Zhao, N. Elmqvist, and D. S. Ebert. Timefork: Mixed-initiative time-series prediction. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pp. 223–224. IEEE, 2014.

[3] 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.

[4] J. G. De Gooijer and R. J. Hyndman. 25 years of time series forecasting. *International journal of forecasting*, 22(3):443–473, 2006.

[5] M. C. Hao, H. Janetzko, S. Mittelstädt, W. Hill, U. Dayal, D. A. Keim, M. Marwah, and R. K. Sharma. A visual analytics approach for peak-preserving prediction of large seasonal time series. In *Computer Graphics Forum*, vol. 30, pp. 691–700. Wiley Online Library, 2011.

[6] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] T. Joachims. Making large-scale svm learning practical. Technical report, Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, 1998.