Interactive Temporal Feature Construction: A User-Driven Approach to Predictive Model Development

David Gotz[*]
University of North Carolina at Chapel Hill

Rashnil Chaturvedi[†]
University of North Carolina at Chapel Hill

Abstract

As organizations gather ever larger and more detailed datasets, predictive modeling is becoming a widely used technology in support of data-driven decision making. In a diverse set of disciplines, ranging from advertising to medicine, temporal event data (such as click streams and electronic health records) are increasingly being used as the basis for training these predictive models. In these cases, temporal relationships between events (e.g., one event occurring before another vs. the same events in opposite order) can be highly predictive. However, existing methods for feature construction make it difficult to incorporate this sort of information, and often require domain experts to manually specify patterns of interest. This poster introduces Interactive Temporal Feature Construction (ITFC), a visual analytics technique designed to enable more effective, data-driven temporal feature construction. The primary contributions for this work include a new interactive workflow for model refinement, a set of algorithms and visual representations designed to support that workflow, and a use case which demonstrates how ITFC can result in more accurate predictive models when applied to complex cohorts of electronic health data.

1 Introduction

As digital technologies proliferate, organizations are instrumenting a wide range of systems to gather ever larger and more complex collections of data. These rich, real-world datasets are then being analyzed using advanced machine learning techniques to extract insights that support data-driven decision making. For example, predictive modeling is widely used in the medical domain to determine patients at high risk. Accurate predictive models can be enormously valuable for both automated and human decision-making tasks.

Determining the most effective feature vector representation for a complex dataset can be challenging for high-dimensional data. In these cases, where datasets can contain thousands or even hundreds-of-thousands of variables, significant effort must be made to identify which variables (or combinations of variables) would be most useful to include as features during model training [2]. The automated creation of feature vectors in the training set is accomplished in part through the use of algorithms, such as feature agglomeration, which search the training set for combinations of variables that are highly predictive. These *constructed features*, which are built from a combination of native variables in a dataset, can then be incorporated into the model-building process with the hope of enabling more accurate predictions.

While this automated approach to feature construction can work well in many cases, it can be difficult to use with temporal event datasets where the time of occurrence for individual data points is of importance. The space of possible constructed variables grows combinatorially. Model developers are therefore often left to manually construct features based on domain expert guidance, or to incorporate additional constraints on the construction process (e.g., [3]).

This poster introduces *Interactive Temporal Feature Construction* (ITFC), a visual analytics technique designed to overcome these challenges using pattern-based feature construction. As the use case presented demonstrates, ITFC has been applied to real-world medical datasets and has allowed users to successfully execute an iterative model building process that resulted in quantifiable improvements to model accuracy. The key research contributions for ITFC include:

- An iterative, user-in-the-loop model development workflow which includes: (1) evaluation of model performance, (2) visualization of event patterns in erroneously predicted data records, (3) construction of new features from patterns identified in the visualization, (4) training of new models that incorporate the newly constructed features, and (5) a visual comparison of model performance to understand benefits of the newly constructed features.
- Results and discussion from a use case where the ITFC approach has proven effective for the improvement of predictive models trained on real-world medical data.

2 Motivation: Predicting Health Care Outcomes

The concept of time is central to nearly all aspects of health care. Patients have symptoms which evolve over time. Clinicians make diagnoses of specific conditions and/or perform diagnostic procedures. A timeline of events can contain many thousands of data points per person, and in modern health systems is captured within an Electronic Health Record (EHR).

One of the most compelling applications in population health analytics is predictive modeling [1], where the goal is to use the large collection of EHR data within an institution to build models that predict the onset of disease, the risk of hospitalization, treatment efficacy, and other outcomes of interest. Given the high costs (both economic and health) in medical care and the large numbers of people who receive treatment, even small improvements in the ability to anticipate outcomes can be hugely beneficial across a population. For this reason, health analysts can invest significant time to iteratively build, evaluate, and refine new versions of a predictive model (through parameter tuning and additional efforts at feature construction). This is an arduous task, especially given the nearly unlimited combinations of temporal relationships. This iterative model building process is what the methods introduced in this poster are designed to support.

3 ITFC Workflow

The ITFC workflow includes two high-level phases. At the start, an *initialization phase* results in the construction of an initial predictive model. Then a *refinement phase* supports iterative feature construction to drive model improvements. During the initialization phase, a data set is selected and temporal event sequence data is aligned around a sentinel event (e.g., aligning medical data around the first date of diagnosis of a disease for each patient). The dataset is then split into three subsets: a training set, a testing set, and a refining set. Using the training set, an initial baseline model is then configured and trained.

Once a baseline model has been trained, the iterative refinement phase of ITFC begins. The refinement cycle consists of three steps performed repeatedly until the model has achieved an acceptable level of predictive performance. This cycle includes (1) model evaluation on the testing set, (2) interactive visualization-driven feature construction using the refining set, and (3) refined model construction.

[*] e-mail: gotz@unc.edu

[†] e-mail:rashnil@live.unc.edu

IEEE VIS 2017
October 1–6, Phoenix, Arizona, USA

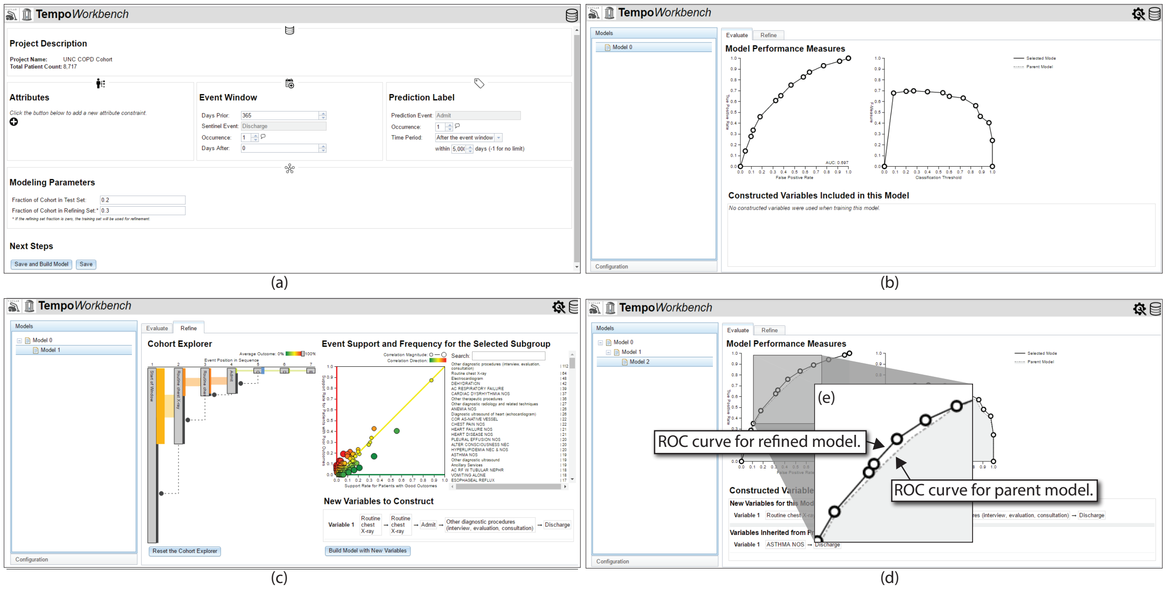
TempoWorkbench
Project Description
Attributes
Event Window
Prediction Label
Modeling Parameters
Next Steps
Save and Build Model
Save
(a)
Model Performance Measures
Constructed Variables Included in this Model
(b)
Cohort Explorer
Event Support and Frequency for the Selected Subgroup
New Variables to Construct
Reset the Cohort Explorer
Build Model with New Variables
(c)
(e)
ROC curve for refined model.
ROC curve for parent model.
(d)

Figure 1: Screen captures of the prototype ITFC system showing (a) inclusion criteria and model configuration, (b) initial model evaluation, (c) feature construction, and (d) refined model evaluation showing (e) ROC curves to illustrate improvements in performance.

4 COPD Use Case

To demonstrate the ITFC approach to predictive model development, a prototype was developed and applied to the task of hospital readmission risk prediction for patients suffering from Chronic Obstructive Pulmonary Disease (COPD). For this use case, data for a cohort of 8,717 COPD patients was gathered from the UNC Health System's Clinical Data Warehouse [4].[1]. The dataset contained diagnoses, procedures, and encounters for each patient, for a total of approximately 1.25 million point events.

Focusing on readmission risk after discharge, we configured the initial model with the following settings: "hospital discharge" as the sentinel event, and an event window focused on one year of events prior to the time of discharge. The prediction target was specified as a "hospital admission" occurring *after* the time of discharge. Finally, we distributed 50% of the matching patients to the testing set, 20% to the training set, and 30% to the refining set. Figure 1(a) shows a screenshot of the prototype during this configuration step.

After configuration, the initial model was trained and the performance measurements shown in Figure 1(b) were calculated and visualized. The AUC [2] for the initial model was 0.697. Switching to the refine tab, we began the model refinement phase of the ITFC. In the first refinement iteration, we used the cohort visualization tool to define a pattern related to asthma prior to discharge from a hospital ("Asthma NOS → Discharge"). We then re-trained the model, examined the new model's evaluation results, then re-visited the refinement tab to continue the iterative process. Figure 1(c) shows the refinement tab at this stage of the process. In this case, we defined yet another new variable related the event pattern being visualized in the figure: "x-ray → x-ray → hospital admission → routine procedures → discharge".

Using this additional variable to train a 3rd model, we see an improved result with an AUC of 0.715. This represents an prediction performance improvement of nearly 2.6%. The evaluation panel showing this final result is shown in Figure 1(d). The model tree on the left displays the sequence of three models described in this use case (i.e., Models 0, 1, and 2). The inherited pattern variables shown in Figure 1(d) represent the variables constructed during the refinement of Model 0 to create Model 1. Finally, the performance measure plots in Figure 1(d) allow for the comparison of performance between the current model (Model 2) and its parent (Model 1). This comparison shows that the addition of the new for Model 2 has produced a bump in the middle of the ROC curve (when false positive rates are between 0.3 and 0.6). This shows that the new model is indeed helpful, allowing us predict more accurately some of the harder to classify patients.

5 Conclusion

This poster has introduced a new visual analytics approach, Interactive Temporal Feature Construction (ITFC), which supports an iterative temporal predictive model development and refinement process in which users are placed "in-the-loop" for the critical feature construction process. A key research contribution for this work is the introduction of a novel visual analytics-based interactive workflow for predictive modeling. A prototype system implementing the ITFC workflow has been developed and applied to real-world datasets from the medical domain. A use case has been presented to showcase the ability of ITFC to help users to construct informative pattern-based features that result in quantifiable model improvements.

These initial results are promising, but several challenges remain to be addressed in future research. Future plans include more comprehensive user evaluations, including longer-term case studies. Another area of interest for future work is the integration of other types of constructed features, such as hierarchical aggregation of event types, that could further improve prediction performance.

Acknowledgments

This work was funded in part by NSF Award DMS 1557593.

References

[1] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7):1123–1131, 2014.

[2] I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.*, 3:1157–1182, Mar. 2003.

[3] K. Malhotra, S. B. Navathe, D. H. Chau, C. Hadjipanayis, and J. Sun. Constraint based temporal event sequence mining for Glioblastoma survival prediction. *Journal of Biomedical Informatics*, 61:267–275, June 2016.

[4] NC TraCS. Clinical Data Warehouse for Health. https://tracs.unc.edu/index.php/services/biomedical-informatics/cdw-h. Accessed: 2017-03-24.

[1] COPD was represented using the family of 490.*-496.* ICD-9 codes.

[2] The area-under-the-curve (AUC) is a widely used measure of predictive accuracy which ranges from 0.5 (random chance) to 1.0 (perfect prediction).