Visualizing Accuracy to Improve Predictive Model Performance

David Gotz and Jimeng Sun

Fig. 1. This screenshot shows a modified version of a temporal event sequence visualization called DecisionFlow [2]. It has been connected with a predictive modeling platform [4] and updated to visually integrate model accuracy data. It supports the identification of event types that are highly correlated with incorrect predictions, and subgroups of the data that are poorly predicted. This provides model developers with a powerful set of tools for understanding erroneous predictions, identifying problems with predictive models, and eventually developing more targeted and precise models that improve overall accuracy.

Abstract— Visualization methods have traditionally focused on visualizing retrospective data, often with the goal of helping users identify data attributes with strong associations to specific outcomes of interest. This can be very helpful during various stages of predictive model development including feature selection, feature construction, and model configuration. While less studied, visualization can also be a powerful tool in the steps that come after a model has been trained: validation and refinement. This paper describes our preliminary work exploring the use of interactive visualization for two specific validation and refinement tasks. In particular, we focus on (1) the visual identification of features associated with incorrect predictions, and (2) visual cohort segmentation to support the development of more targeted predictive models for poorly predicted sub-populations.

Index Terms—Information Visualization, Visual Analytics, Medical Informatics, Predictive Modeling, Temporal Event Data

1 Introduction

Visualization methods have long been used to support exploratory data analysis tasks. A wide variety of techniques have been designed to provide overviews of a dataset, to support interactive filtering of datasets to identify subsets of interest, and to enable the inspection of details for individual (or sets of) data elements [5]. The many interactive visualization methods developed over the years provide a rich set of exploratory tools for the predictive model developer. By visually exploring their data, model developers can leverage both computational power and their visual perception system to improve data quality, detect anomalies, identify patterns, and explore high-dimensional data. These activities can support various stages of the predictive model development process including feature selection, feature construction, and model configuration/selection (e.g., [1]).

Though less frequently explored, interactive visualization techniques can also be used to help developers after model has been trained. Work in this area in has largely focused on validation (e.g., [3]), a critical step in understanding how effective a given model is at predicting specific phenomena. During validation, the developer often seeks more than just a measure of accuracy. She/he also needs insights

- *David Gotz is with the University of North Carolina at Chapel Hill. E-mail: gotz@unc.edu.*
- *Jimeng Sun is with the Georgia Institute of Technology. E-mail: jsun@cc.gatech.edu.*

Manuscript submitted 31 March 2014 for review as part of IEEE VAST 2014.

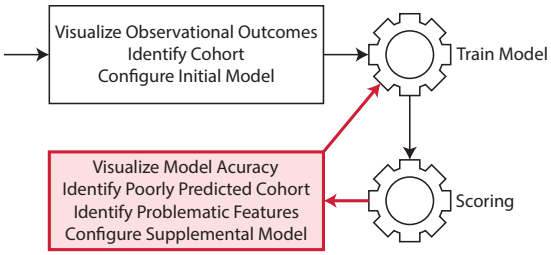

Fig. 2. In our iterative workflow, visualization techniques are first used to explore outcomes, identify cohorts of examples, and configure initial parameters for model training. The model is then used to score additional examples. The accuracy of the predictive model is then visualized to identify problematic features and examples. These in turn support additional rounds of more focused model training.

that can point to ways in which a model can be refined and optimized for a given prediction task.

In the work described in this paper, we are focusing on the development of visualization-based methods for two specific types of model validation and refinement tasks. The prototype implementation of our methods has been developed in the context of temporal event data. However, our approach can be generalized to a broad set of predictive modeling problems in which models are developed using a training set of *samples*, each of which consists of a set of *features*.

First, we are exploring visual methods that enable the identification of *problematic samples*. The goal here is to identify "hard to predict" subsets of the population: subgroups of the input samples on which the model exhibits poor prediction performance.

Second, we are exploring the use of interactive visualization to help identify *problematic features* during validation. The goal of this aim is to help developers determine, for a given predictive model, which features are most strongly correlated with poor prediction performance.

When used in combination, these techniques can be used by developers and data scientists to both construct new features and to identify subgroups that might benefit from more targeted, population-specific predictive models because they differ so significantly from the general training population. In this paper, we outline our basic approach and describe our preliminary prototype implementation.

2 Our Approach and Preliminary Results

The focus of our work is on visualizing the results of predictive models to identify (1) *problematic samples*, the subsets of the scored population that exhibited especially low levels of prediction accuracy, and (2) *problematic features*, the individual features that most strongly associated with incorrect predictions. These insights can be very valuable to predictive model development. However, these capabilities form just one part of a larger iterative workflow that we aim to support for predictive model development and validation. This workflow is illustrated in Figure 2.

The workflow starts with a model developer using exploratory visualization tools to better understand their data, both in terms of input features and the outcome variable that is to be predicted. Once those tools help the developer identify a representative cohort of examples and an initial model configuration, a new predictive model can be trained. The new model is then used to score a test population for which ground truth is known. In a traditional model development process, techniques such as cross-validation are used to measure accuracy and generalizability. The resulting statistics can be compared across alternative models.

However, our workflow aims to use visualization to provide a more detailed view of model accuracy. In particular, we use exploratory visualization tools similar to those used in the initial cohort identification phase to examine how specific features correlate with incorrect predictions. We also allow users to identify problematic subsets of the population. These are defined groups for whom the predictive model yielded highly inaccurate results in comparison to the larger scored population. Armed with these new insights, users can select problematic subgroups as training sets for new models which can be configured in a highly focused and precise way for the targeted subgroup.

This new accuracy visualization portion of our proposed workflow is illustrated in red in Figure 2. As the diagram shows, this process can be repeated iteratively, resulting eventually in a suite of predictive models that together are intended to provide a more accurate prediction rate for the overall population.

We have developed a prototype implementation that supports this workflow by connecting a modified version of the DecisionFlow visual analytics system [2] to the scalable PARAMO predictive modeling platform [4]. The DecisionFlow interface was modified to (a) visualize prediction accuracy (using models computed from PARAMO) rather than outcomes, and (b) pass model configuration information to PARAMO to initiate the training of new models directly from the visual interface.

2.1 Problematic Samples

DecisionFlow visually represents temporal event sequence data using a horizontal timeline that is subdivided into individual pathway segments that represent subgroups of sequences that share common event milestones. In its original form, DecisionFlow color-coded the segments based on the average of the actual outcomes observed for the corresponding subgroup of event sequences.

To help users visually identify problematic subgroups for which prediction accuracy was especially poor, we extended the DecisionFlow system to accept prediction accuracy data from PARAMO. We then revised the visual encoding used in the visualization to map color (on a red-to-yellow-to-green color scale) to average prediction accuracy.

For example, Figure 3 shows a screenshot from our modified version of DecisionFlow being used to visualize results from a predictive model designed to predict Heart Valve Replacement for a set of patients with cardiovascular diseases. DecisionFlow was used to segment the predicted population into three distinct subgroups based on which lab tests were performed on the patients. The visualization shows significant variation in model performance based between these groups as represented by the distinct background colors (orange, yellow, and green). While the model performed quite strongly for patients represented by the green segment, those represented by orange were harder to predict.

The screenshot shown in the figure illustrates three subgroups. However, DecisionFlow allows the predictive modeler to define ad hoc subgroups as they explore the accuracy data provided by PARAMO. The milestone events used to define the subgroups can be dynamically added and/or removed via direct manipulation to make subsetting fast and intuitive. More information about the exploratory interactions supported by DecisionFlow can be found in the original paper describing its methods [2]. As described in Section 2.2, the definition of new subgroups is driven by the portion of the visualization showing correlation between features and prediction accuracy

DecisionFlow was further modified to allow users to quickly train new models that target a specific problematic population of samples. For example, the user inspecting the model in Figure 3 has clicked on the orange subgroup highlighted with thick black edge. This selects the corresponding patients which can be sent to PARAMO to train a new model. The initiation of a new training run is triggered by a menu command anchored to the 'Actions' menu. This feature completes the round trip between DecisionFlow and PARAMO required to support the iterative process outlined in Figure 2.

2.2 Problematic Features

In addition to modifying DecisionFlow's timeline, we altered both the calculations and color coding used in the event statistics panel that appears at the bottom of the user interface. The original design highlights features correlated with observed outcomes. We instead use the view to convey the prevalence of specific feature types in the correctly and incorrectly predicted populations. Each event type is represented by a circle in a scatter plot with X, Y axes representing positive support (p)

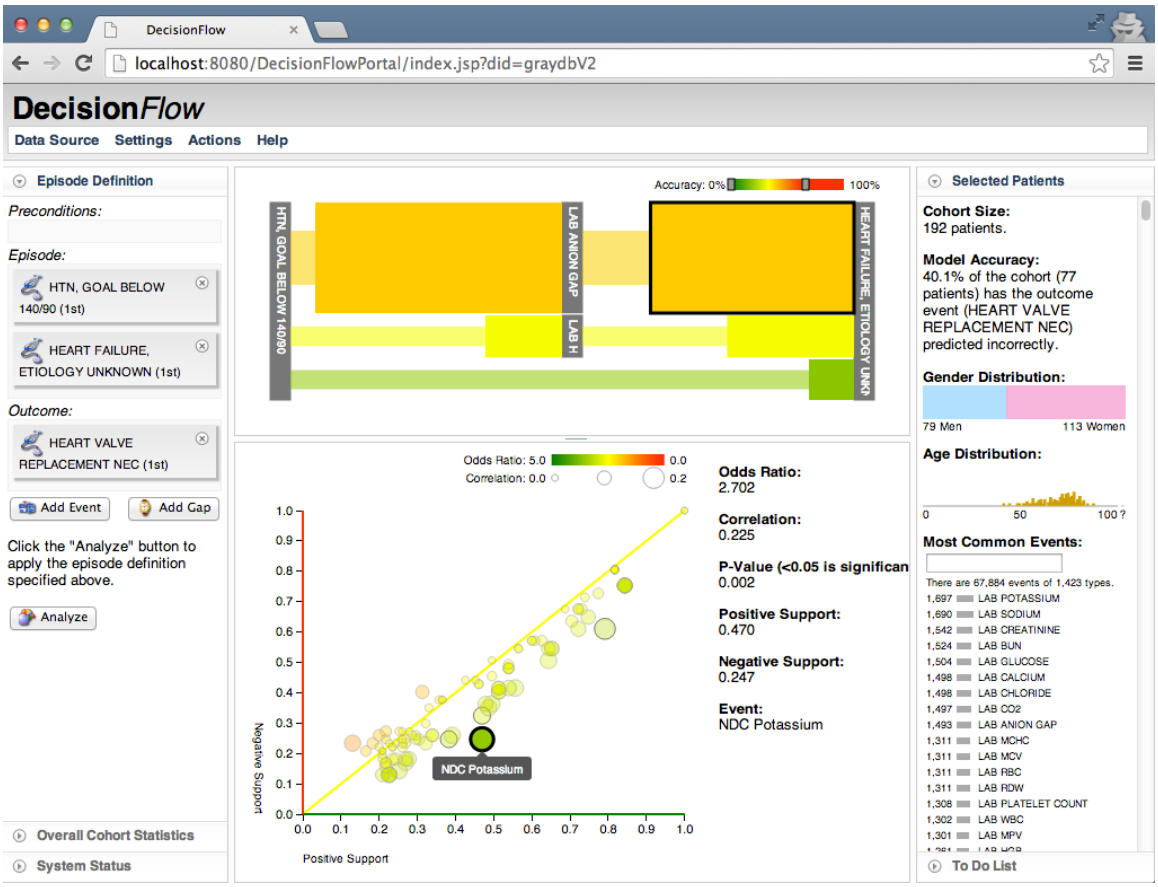

Fig. 3. In this example, a user has used a modified version of DecisionFlow to segment the example population into subgroups based on the appearance of specific features in the data. Predictions for each subgroup have differing levels of prediction accuracy as illustrated by the color-coded bands in the top portion of the visualization. The scatterplot view shows individual features that associate strongly within the subgroup with correct (green) or incorrect (red) predictions and can be used to interactively define additional subgroups. Once a problematic cohort is isolated, users can quickly send the samples to the linked PARAMO platform to begin training a new predictive model that targets the narrower, more focused set of examples.

and negative support (n), respectively. The support measures p and n for event type t are defined as follows:

$$p(t) = \frac{\text{count of correctly predicted patients with event } t}{\text{count of all correctly predicted patients}} \quad (1)$$

$$n(t) = \frac{\text{count of wrongly predicted patients with event } t}{\text{count of all wrongly predicted patients}} \quad (2)$$

The color-coding and size of each circle show odds ratio and correlation statistics, both computed from the model accuracy data.

For example, the scatter plot in the lower portion of Figure 1 shows that the "Lab Anion Gap" test was one of a several features in the data that correlated significantly with incorrect predictions. A model developer wishing to learn more about those patients could select "Lab Anion Gap" for *promotion*, which uses the event to define new subgroups in the timeline view. As seen in Figure 3, which shows the same model being validated after two different features have been promoted, this allows the developer to focus on specific subgroups as outlined in Section 2.1. This same figure shows the event "NDC Potassium" with support in roughly 50% positive support but only 20% negative support.

3 Conclusion

As more and more data is collected and analyzed, predictive modeling is quickly becoming an everyday tool in many domains. However, challenges remain in training accurate models for complex real-world applications. Achieving higher levels of accuracy requires methods that help developers better understand why some predictions fail while others succeed. Interactive visualization can play an important role in these tasks in the same way that it has helped a wide variety of other exploratory data analysis activities. This paper presented an overview of our first steps in this direction and provided a glimpse at our emerging platform for predictive model validation and refinement.

References

[1] M. Bgl, W. Aigner, P. Filzmoser, T. Lammarsch, S. Miksch, and A. Rind. Visual analytics for model selection in time series analysis. *IEEE transactions on visualization and computer graphics*, 19(12):2237–2246, Dec. 2013.

[2] D. Gotz and H. Stavropoulos. DecisionFlow: Visual analytics for high-dimensional temporal event sequence data. *To Appear in IEEE Transactions on Visualization and Computer Graphics*, 2014.

[3] T. Muhlbacher and H. Piringer. A partition-based framework for building and validating regression models. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1962–1971, 2013.

[4] K. Ng, A. Ghoting, S. R. Steinhubl, W. F. Stewart, B. Malin, and J. Sun. PARAMO: A PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. *Journal of Biomedical Informatics*, 48:160–170, Apr. 2014.

[5] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343, Sept. 1996.