Learning to Identify Locally Actionable Health Anomalies

Kuang Chen and **Emma Brunskill**
Computer Science Division
University of California, Berkeley
{kuangc|emma}@cs.berkeley.edu

Jonathan Dick
Department of Medicine
University of Chicago
jjd@uchicago.edu

Prabhjot Dhadialla
Earth Institute
Columbia University
pdhadialla@ei.columbia.edu

Abstract

Local information access (LIA) programs tap into existing public health data flows, and present data in simple and useful ways to ground staff. LIAs hold great potential for improving rural health systems in developing regions; benefits include more evidence-based decision making and optimizations at a local scale, as well as improved service delivery and data quality. Our fledgling LIA program in rural Uganda currently provides clinicians with a small set of static data visualizations for discussion. To increase the program's effectiveness, we want to automatically identify relevant data visualizations. We propose an adaptive tool that learns from local clinicians' decision-making processes to predict and generate visualizations that show actionable anomalies.

Introduction

In recent years, progress toward effective public health reporting in developing regions has been driven by a call for better monitoring and evaluation (M&E). According to a World Bank report, "prioritizing for M&E has become a mantra that is widely accepted by governments and donors alike." (Mackay 2009). However, M&E is often implemented specifically to meet top-down requirements. Thus, ground staff often perceive little benefit from data collection, and do the minimum amount of work necessary to meet reporting requirements. Worse, clinicians miss opportunities for local interventions, including those which get lost at higher levels of data aggregation, and those which might only be actionable based on local insight. We posit that local use of data will lead to better data quality, and contribute to better health care and decision-making.

In rural Uganda, we have started a LIA program that features simple visualizations of data from existing sources to support clinicians in evidence-based management of their health program. A core challenge to encouraging and sustaining local data analysis is to provide data visualizations that indicate locally actionable health anomalies.

A rural health system can include a variety of data: from community health worker (CHW) surveys to clinical medical records. In our implementation, we worked with over 60 variables from two data sources. The number of potential visualizations is a combinatorial function of these variables, and it is unrealistic to expect a clinician to look through an exhaustive set. Instead we aim to learn a classifier that can automatically identify actionable health trend outliers, and present visualizations of these trends to clinicians.

To build a classifier, we require labels to identify the relevant visualizations which are anomalous and actionable. Collecting a sufficient set of labels in advance, outside of the normal work flow, would be time consuming and the labels collected would potentially be less reliable. Instead we propose an active learning approach to identify the boundary between relevant and irrelevant visualizations. Select visualizations will be presented to clinicians as candidate discussion items for biweekly CHW-clinician meetings. In the process of creating the agenda and summarizing meeting minutes, we can capture labels for a subset of visualizations, which can be used to train the classifier.

Background

With the Millennium Villages Project[1] in Uganda, we implemented a simple data visualization tool using a Microsoft Excel workbook. The first tab is for the entry of weekly community health worker data. The second tab features dynamic import of clinic visit records from the villages' electronic medical records (EMR) via a database connection. Having clinic and CHW information in Excel enable additional tabs to feature summary tables and visualizations based on PivotTables and PivotCharts. The tool provides visualizations such as: "Patients under 5 years old with symptom X by village." In the initial weeks of use, a clinician found in a visualization that the number of clinic-to-CHW requests were abnormally low at some clinics. At the next meeting, he held an evidence-based discussion of the issue. This was an exemplary demonstration of the potential impact of appropriate data analysis, and it motivated our interest in trying to automatically identify clinically relevant and actionable visualizations.

Problem Definition

In the MVP Uganda health system, clinicians receive weekly updates to their health data records. From these records, a finite, though combinatorially large, set of visualizations can be constructed. We would like to extract features about visualizations that generalize over the changing data. First, we

[1] http://www.millenniumvillages.org

define a variable-length parameter vector p of the database table entries relevant to the variables plotted in a particular visualization. Next we compute features over this parameter vector p, such as the intercept and slope of a line fit through the data. Our objective is to learn a mapping from a fixed-length feature vector to a label that indicates whether the associated data visualization is anomalous and actionable.

To gather labels, each week the system will select a small set of visualizations as potential meeting agenda items. At the meetings, clinicians will be asked to vote upon which suggested visualizations would make likely agenda items, and which are not relevant: this provides a source of positive and negative labels to the system. After each meeting, we will extract the visualizations discussed at the meeting based on the weekly minutes: initially we intend to rely on the clinician in charge of note-taking, but later this could be automated. These will provide an additional source of positive labels. These labels will be incorporated into the classifier for re-training, along with any new clinical data that arrives over the following week. The proposed process is displayed in Figure 1. As the dataset will consist of a small set of labeled data, and a large amount of unlabeled data, with the ability to gain more labels, this problem is an instance of semi-supervised active learning. The mentioned methods to obtain labels may be too slow, so we are also exploring bootstrapping the system by getting local experts to label a large set of visualizations, and we hope to identify other points in the standard health worker process flow where labels could be added.

Initially we intend to follow Tong and Koller's (2001) approach for active learning using Support Vector Machines (SVMs). The authors use a heuristic that approximates which unlabeled visualizations would, if labeled, most reduce the size of the *version space*. The version space is the set of decision boundaries that separate the labeled data (as represented in the feature space defined by the particular kernel used in the SVM). However at each step, we also wish to identify potentially relevant visualizations. These will be identified by using transductive SVM (Joachims 1999) which uses both the labeled and unlabeled data to build a classifier, and then use the labels from this classifier to predict relevant visualizations.

There are a number of interesting technical features of this problem, including the high dimensional feature space relative to the amount of data and incorporating prior information over labels. A particularly important aspect of this problem is the tension between which visualizations the system should present to clinicians, assuming there is a fixed reasonable number. Clinicians are primarily interested in new visualizations that the system predicts to be relevant. Thus the system should present the visualizations with positive labels. However, for improving the classifier, e.g. actively learning the decision boundary, the system should query clinicians about visualization labels with the most uncertainty. This problem is related to the classic exploration/exploitation tradeoff. To help address this, we could define a utility function over the benefit of identifying or cost of failing to identify, an actionable, clinically relevant data visualization. Maximizing the expected infinite-horizon sum of utilities would provide an objective measure for trading between the benefit of displaying more data visualizations that the system is confident are relevant, with the benefit of presenting data visualization where the label is highly uncertain, but may be very useful for learning the classification function. Recent work by Kappor and Horvitz (2009) that reasons about better labeling past data versus classifying a new sample may also be applicable to selecting which visualization labels to query.

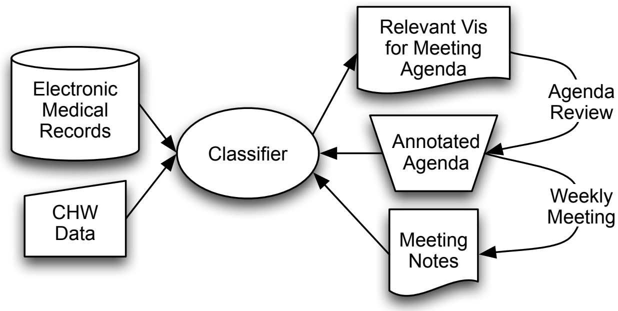

Figure 1: Proposed data flow.

Proposed Pilot Evaluation

Our LIA program featuring the static Excel tool is currently in production use. We are in the process of extending the tool, and plan to conduct a pilot evaluation, following existing workflows in the MVP health program. To bootstrap initial performance we will obtain labels by querying domain experts about which of a a small fixed subset of visualizations would be interesting, and potentially actionable.

During the pilot we expect to gain valuable qualitative feedback about the program design: the user interface is also likely to be critical for ensuring that system is easy to use, and the suggested anomalies are duly noted by the clinicians. After each meeting, we will quantitatively evaluate the classifier by the percentage of discussed items that were classified as relevant (recall), and the percentage of classified relevant items that were discussed (precision).

We hope this tool may increase the relevancy of local data and aid local clinicians in performing evidence-based interventions and management that improve community health.

Acknowledgements

We would like to thank Ben Bellows, Jahanzeb Sherwani, Joe Hellerstein, Julia Frydman, Melissa Ho, Roni Rosenfeld, Tapan Parikh and Yaw Anokwa.

References

Joachims, T. 1999. Transductive inference for text classification using support vector machines. In *ICML*.

Kapoor, A., and Horvitz, E. 2009. Breaking boundaries: Active information acquisition across learning and diagnosis. In *NIPS*.

Mackay, K. 2009. *How to Build M&E Systems to Support Better Government*. Washington, DC: World Bank.

Tong, S., and Koller, D. 2001. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research* 45–66.