Using Anchors to Estimate Clinical State without Labeled Data*

Yoni Halpern[1], Youngduck Choi[1], Steven Horng MD MMSc[2], David Sontag PhD[1]
[1]New York University, New York, NY
[2]Beth Israel Deaconess Medical Center, Boston, MA

1 Introduction

We present a novel framework for learning to estimate and predict clinical state variables. Using this, we seek to create a new middleware application layer consisting of hundreds of clinical state variables that summarize a patient's past and current state. These variables can used for electronic phenotyping and to trigger clinical decision support. In its simplest form, one could envision this as a collection of binary classifiers. However, most machine learning based approaches require experts to label positive and negative examples. This can be time consuming and becomes prohibitive as we scale to a large number of clinical variables. Additionally, rules learned on a data set in one institution may not transfer to other institution if the underlying data representation changes.

We suggest a different approach; instead of manually labeling patient examples, an expert identifies certain highly informative variables which we term "anchor variables". We show that if the anchor variables satisfy two properties, they can be used instead of labels to learn decision rules from large amounts of otherwise unlabeled data. The ability to build anchors upon standardized ontologies and the framework's ability to learn from unlabeled data promote generalizability across institutions.

We developed a user interface to enable experts to choose anchor variables in an informed manner, and applied the framework to electronic medical record-based phenotyping to enable real-time decision support in the emergency department. We validate the learned models using a prospectively gathered set of gold-standard responses from emergency physicians for nine clinically relevant variables.

2 Anchor variable framework

Rather than take the standard machine learning approach of labeling patients, we introduce a learning procedure that uses *positive anchor variables*, defined in Definition 1.

Definition 1. *An observation X_j is an* anchor *for a latent variable Y_i if X_j is conditionally independent of all other observations, X_k $\forall k \neq j$, conditioned on the value of Y_i. It is a positive anchor if it has the additional property that $P(Y_i = 1|X_j = 1) = 1$.*

In other words, anchor observations provide a direct, albeit noisy, view of the underlying latent variable we wish to predict. The key characteristic of an anchor is the conditional independence property which states once the value of the latent variable is known, no other observables provide additional information about the anchor variable. The positive condition means that observing the anchor to be positive unambiguously reveals the state of the latent variable to be positive, while observing it to be negative does not necessarily reveal the state of the latent variable. This setting has previously been studied under the name *positive-only labels* and a procedure for learning with positive-only labels is described in [1]. Briefly, the procedure consists of learning classifiers to predict the presence of the *anchor* rather than the presence of the underlying latent variable and performing a correction to account for the difference in predicting the anchor and the latent variable.

Anchors like this exist in medicine. For example, a positive rapid antigen test for Group A streptococci is very specific for strep throat and can serve as a positive anchor. The absence of a positive test result can be uninformative either due to the low sensitivity of the test, or due to the possibility that the test was never performed because the diagnosis was obvious and the patient was treated without testing [2]. In real use cases, no anchors will ever perfectly meet the criteria in Definition 1, and missing data will not be completely at random. Nonetheless, the definition gives theoretical principles by which to choose good anchors. As we demonstrate with experimental results in the next section, approximate anchors can perform well on real data.

*This abstract is a summary of our AMIA 2014 Annual Symposium paper, which can be found at `http://clinicalml.org`.

3 Data and Methods

For training and evaluation we use a collection of 273,174 emergency department patient records collected from a 55,000 patient/year Level 1 trauma center and tertiary academic teaching hospital between 2008 and 2013. Each record represents a single patient visit. All consecutive ED patient visits were included in the data set and no visits were excluded. This study was approved by our institutional review board.

To evaluate the utility of the latent variable predictors learned, we collected gold standard labels from the primary clinical provider caring for a patient at the time of disposition from the emergency department, (admission, discharge, or transfer). As part of the routine clinical workflow of disposition, clinicians were asked a series of 2-3 questions chosen randomly from a rotating pool of questions (see Table 1). Overall, from 1,082 (`isAnticoagulated`) to 62,589 (`hasInfection`) responses were collected for each question.

Medications are represented by generic sequence number (GSN) and diagnosis codes by ICD9 codes. Age was discretized by decade with a binary indicator for each decade. Patients are represented as a binary feature vector representing the presence or absence of each distinct diagnosis code, current medication, dispensed medication (as recorded by Pyxis machine), word, discretized age value, and sex. The free text was derived from the chief complaint, triage assessment and physician's comments. Observations that occur in fewer than 50 patients in the entire dataset were discarded, leaving a final binary feature vector of size 20,334.

We designed a user interface for eliciting anchor variables from experts (described in the extended version of this paper), which they used to specify anchors for the 9 clinical state variables. For example, for `hasSepticShock` two anchors were specified: the diagnosis code 785.42 (septic shock) and the Pyxis medication group for cardiac sympathomimetics. In the evaluation, diagnosis codes are removed in the test patients in order to simulate a real-time decision environment.

4 Results

We compare classifiers learned with anchors against a simple rules baseline which predicts 0 if the anchor is absent and 1 if the anchor is present, as well as a logistic regression model using a portion of the collected labels for training purposes. Table 1 shows the performance of the classifiers using the area under the ROC cure (AUC). Across the nine clinical state variables considered, our anchor-based unsupervised learning algorithm obtains prediction accuracy comparable to and in many cases better than a supervised prediction algorithm.

Variables	**Rules**	**Supervised**						**Anchors**
		100	**200**	**500**	**1K**	**2K**	**3K (min, max)**	
didFall	0.725 ± 0.008	0.814	0.852	0.900	0.914	0.920	**0.924** (0.917,0.934)	0.883 ± 0.006
hasCardiacEtiology	0.611 ± 0.006	0.772	0.827	0.824	0.875	0.900	**0.906** (0.891, 0.920)	0.881 ± 0.006
hasInfection	0.723 ± 0.002	0.728	0.767	0.804	0.830	0.861	0.883 (0.881, 0.886)	**0.903± 0.001**
fromNursingHome	0.620 ± 0.005	0.725	0.792	0.822	0.869	0.894	0.891 (0.873, 0.906)	**0.918 ± 0.004**
hasCancer	0.822 ± 0.018	0.635	0.673	0.693	0.810	0.882	0.902 (0.880, 0.930)	**0.945 ± 0.01**
hasPneumonia	-	0.856	0.907	0.933	0.947	0.956	0.963 (0.954, 0.972)	**0.971 ± 0.003**
isAnticoagulated	0.849 ± 0.03	-	-	-	-	-	-	**0.930 ± 0.02**
isImmunosuppressed	0.650 ± 0.01	0.584	0.659	0.740	0.814	0.842	**0.862** (0.840, 0.877)	0.840 ± 0.009
hasSepticShock	0.738 ± 0.02	-	0.760	0.773	0.863	0.920	0.952 (0.928, 0.967)	**0.967 ± 0.008**

Table 1: Comparing AUC in the real-time setting. The supervised method is trained using logistic regression with a small number of gold standard labels. When the anchors are composed entirely of diagnosis codes, the rules approach cannot be meaningfully evaluated on the test set (in the real-time setting, diagnosis codes are not available at test time). When we had insufficient data to train, the supervised approach could not be evaluated. Best methods in each row are bolded. The anchor approach uses 200K *unlabeled* examples in training. Standard errors of the AUC for Rules and Anchors are computed using 1000 bootstrap samples of the test set. Min and max values for the 3K supervised baseline are from the 4-fold cross validation.

References

[1] Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. In: KDD; 2008. p. 213–220.

[2] Tanz RR, Gerber MA, Kabat W, Rippe J, Seshadri R, Shulman ST. Performance of a rapid antigen-detection test and throat culture in community pediatric offices: implications for management of pharyngitis. Pediatrics. 2009;123(2):437–44.