Bulk Learning on EHR Data

Po-Hsiang Chiu, PhD[a], George Hripcsak, MD, MS[a]
[a]Department of Biomedical Informatics, Columbia University, New York, NY

Introduction

A central computational task in data-driven disease phenotyping is to use the variations of clinical concepts inferred from the electronic health records (EHR) to stratify phenotypic cohorts, each of which represents a clinical condition or a potential disease subtype that goes beyond historical disease definitions. Patients are essentially represented by clinical traits such as medication history and laboratory measurements. Phenotyping using statistical learning methods[1] may seem to reach a higher level of automation when compared to the conventional rule-based approach; however, predictive analytics is not without its own challenges in the creation of training data since both the tasks of feature engineering and gold standard labeling require significant human intervention. To reduce the overhead of data annotation and achieve a higher level of automation, we adopt a hierarchical learning approach based on the ensemble learning paradigm[2], using infectious diseases as the domain of study, to construct abstract features that represent the shared clinical traits among the target infections. Statistical models such as logistic regression can then be built from within the abstract feature space of low dimensionality, thereby reducing the demand for labeled data.

Method

Bulk learning, as the name suggests, is a simultaneous training procedure for multiple clinical conditions. The key idea is to use diagnostic codes (e.g. ICD-9s) readily available in the codified EHR as *surrogate labels*, out of which we can train a hierarchical model with levels of feature abstraction that captures the common denominator of the target conditions, referred to as the *bulk learning set*. Clinical evidence suggests that different infectious diseases share degrees of similarity in antibiotic prescriptions, various laboratory tests, among other clinical factors, all of which can potentially serve as high-level *phenotypic measures* for separating cohorts with different infections. Through the learned *abstract features* representing importance weightings of phenotypic measures of interest, the labeling of disease cases for the entire bulk learning set can be realized by training statistical models from within the abstract feature space. Since abstract feature space has a significantly lower dimension compared to the raw feature space spanned by patient attributes extracted from the EHR data, we postulate that a relatively small annotated dataset is only required for stratifying disease cohorts, which would otherwise have a substantially higher demand of labeled data considering the size of the raw feature set that can be used to model these phenotypic measures. An example three-tier hierarchical learning architecture is depicted in Figure 1 with two levels of feature abstraction: i) Base or level 0 consists of a set of classifiers, each of which is modeled to express a high-level phenotypic measure relevant to the target infections. Each measure is represented by the feature set extracted from Medical Entity Dictionary (MED) developed at NewYork-Presbyterian Hospital[3], a semantic network for medical concepts. ii) The level-1 tier is a meta-classifier that takes probabilistic scores from base classifiers and a set of indicators combined as the abstract feature set, where each indicator represents the absence or presence of data for the corresponding phenotypic measure (e.g. not all patients had intravenous chemistry tests during their clinical visits); up until this level, ICD-9 codes are used as surrogate labels. iii) The level-2 tier inherits the probabilistic output from level 1 coupled with the ICD-9 codes together as the abstract feature set while using annotated labels for model training.

Results and Discussion

In the experiment, we used the clinical data repository from Columbia University Medical Center and examined 100 clinically different infectious conditions. For simplicity, L2-regularized logistic regression is applied in both the base and meta-classifiers. In predicting the ICD-9 as surrogate labels, each base classifier has different predictive strength due to their varying degrees of clinical relevance. The phenotypic measure involving microbiology tests in general has the best predictive strength. The level-1 classifier consistently outperforms individual base classifiers, which can be explained by the diversity in predictions at the base level, a crucial condition for ensemble learning to work well. In the model evaluation with the gold standard, we annotated 83 clinical cases. Under this small annotation set, the ICD-9-modulated level-2 model reaches an AUC of 0.78. However, the model tends to result in false positives due to the inconsistent labeling between the ICD-9 and the gold standard; overcoming this performance shortage is an ongoing effort. Additionally, we experimented on alternative feature abstractions such as the ICD-9-modulated level-1 model, which combines the level-1 features with the ICD9; the performance tradeoff with its level-2 counterpart depends on the size of labeled data. Figure 2 illustrates the relative performance of the global level-1 model in reconstructing the ICD-9 labeling (through predictions), where the model was trained on the aggregated level-1 data across the bulk learning set (hence global to all diseases), followed by predicting individual diseases as the evaluation; the horizontal axis denotes the codified diseases. The empirical result indicates that the global level-1 model approximates the ICD-9 labeling well, which is promising for using such hierarchical learning method for large-scale, high-throughput phenotyping in the near future.

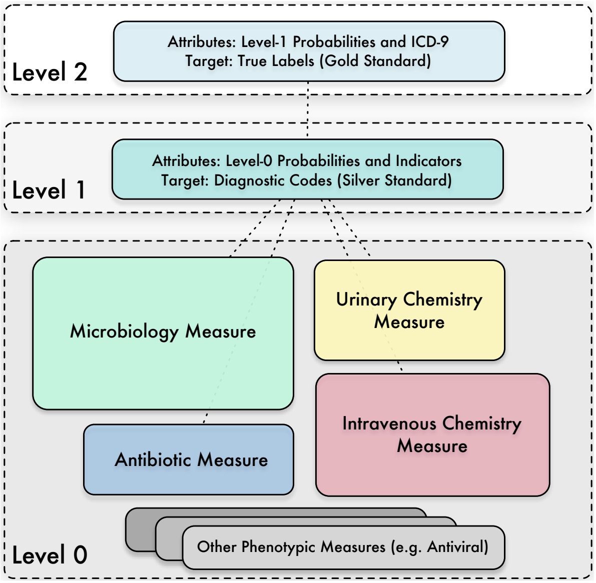
Level 2
Attributes: Level-1 Probabilities and ICD-9
Target: True Labels (Gold Standard)
Level 1
Attributes: Level-0 Probabilities and Indicators
Target: Diagnostic Codes (Silver Standard)
Microbiology Measure
Urinary Chemistry Measure
Antibiotic Measure
Intravenous Chemistry Measure
Other Phenotypic Measures (e.g. Antiviral)
Level 0

Figure 1. Bulk training with the hierarchical learning architecture of 4 phenotypic base models.

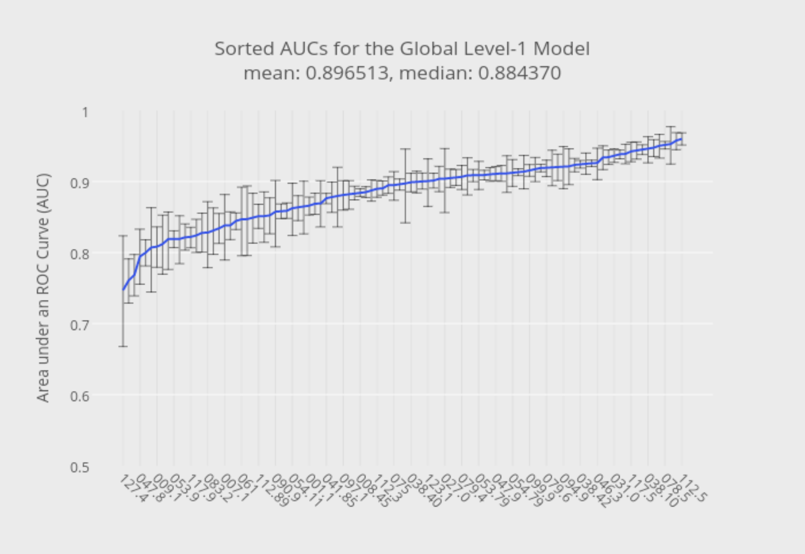

Figure 2. Sorted performance in AUCs for the global level-1 model with a grand mean at 0.897.

Acknowledgment

This work was funded in part by National Library of Medicine grant LM006910.

References

1. Matheny ME, Miller RA, Ikizler TA, Waitman LR, et al. (2010). Development of inpatient risk stratification models of acute kidney injury for use in electronic health records. Med Decis Making 30: 639–650.
2. Whalen S, Pandey GK. A comparative analysis of ensemble classifiers: case studies in genomics. InData Mining (ICDM), 2013 IEEE 13th International Conference on 2013 Dec 7 (pp. 807-816). IEEE.
3. David Baorto, James Cimino, et al. Available: http://med.dmi.columbia.edu. Access date: March 10, 2016.