Learning Useful Abstractions from the Web

Abhishek Bafna(abafna@umich.edu) Jenna Wiens(wiensj@umich.edu)

Background/Objective: The successful application of machine learning to electronic medical records typically turns on the construction of an appropriate feature vector. Defining abstractions to create high-level, lower-dimensional feature vectors can help in identifying clinically meaningful similarities among patients especially when the number of training examples is limited. In this work, we compare conventional unsupervised dimensionality reduction techniques (e.g., principal component analysis, PCA) to novel approaches which leverage a large corpus of freely available expert knowledge in unstructured form from the Web. As a case study, we consider the task of learning useful abstractions from a list of d medications.

Methodology: We begin with a bag-of-words model, in which a patient is represented by a d dimensional binary feature vector, where each feature corresponds to one of d different medications. Next, we build a corpus of documents, one document corresponding to each medication, retrieved from the Web. We then apply latent Dirichlet allocation (LDA) to this corpus. LDA represents each medication in a k dimensional feature space where k is the number of topics (and $k < d$). Using this representation and a distance metric we transform each patient into lower dimensional space and compute a patient similarity/kernel matrix. We consider two distance metrics when building the patient similarity matrix 1) earth mover's distance (EMD) and 2) Euclidean distance. In contrast to the Euclidean distance, EMD allows for incorporation of the underlying distances between topics. Finally, we also apply standard dimensionality reduction techniques such as PCA.

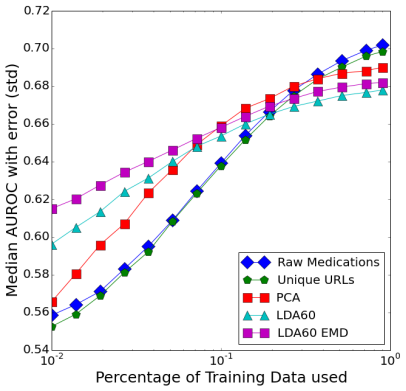

Figure 1: When the amount of training data is limited the models based on abstractions learned from the Web help yield better classification performance.

Evaluation Results: We consider over 25,000 patient admissions from the MIMICII database [1]. In this database each medication is represented by a free text string. After minimal preprocessing, we identified 2,285 "unique" medications. Applying the methods described above we represented each patient admission based on the medications received in three different feature spaces: $\mathcal{X}_{original}$, $\mathcal{X}_{PCA}$, and $\mathcal{X}_{LDA}$. To evaluate the utility of the learned abstractions, we considered three different clinical classification tasks. We split the data into a training and test set limiting the portion of training data used from 1% to 100% and learn five different support vector machine (SVM) classifiers:

- Raw Medications - in which each patient is represented in the original feature space
- Unique URLs - uses a transformation based on the mapping from medications to unique web pages
- PCA - employs the principal component analysis transformation of the data
- LDA60 - uses the transformation based on the results of the LDA analysis and a Gaussian kernel
- LDA60 EMD - uses the LDA feature vectors and Gaussian kernel based on the EMD

to predict different adverse outcomes: mortality within 30 days of discharge, readmission to the hospital after discharge, and readmission to the ICU during the same hospital visit. Figure 1 shows the results of the classification performance on the test set for the third task. On all three classification tasks, when the number of training examples was limited the topic-based classifiers outperformed a baseline classifier built using the set of original medication names as features.

Conclusions: Based on our results, we conclude that LDA can be used for harnessing the large quantity of unstructured medical knowledge on the Web to produce meaningful and useful abstractions in an unsupervised manner.

References:

[1] Saeed, Mohammed, et al. "Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database." Critical care medicine 39.5 (2011): 952.