Semi-supervised Learning for Phenotyping Tasks

Dmitriy Dligach, PhD, Timothy Miller, PhD, Guergana K. Savova, PhD
Boston Children's Hospital and Harvard Medical School, Boston, MA

Abstract

Supervised learning is the dominant approach to automatic electronic health records-based phenotyping, but it is expensive due to the cost of manual chart review. Semi-supervised learning takes advantage of both scarce labeled and plentiful unlabeled data. In this work, we study a family of semi-supervised learning algorithms based on Expectation Maximization (EM) in the context of several phenotyping tasks. We first experiment with the basic EM algorithm. When the modeling assumptions are violated, basic EM leads to inaccurate parameter estimation. Augmented EM attenuates this shortcoming by introducing a weighting factor that downweights the unlabeled data. Cross-validation does not always lead to the best setting of the weighting factor and other heuristic methods may be preferred. We show that accurate phenotyping models can be trained with only a few hundred labeled (and a large number of unlabeled) examples, potentially providing substantial savings in the amount of the required manual chart review.

Introduction

Mining massive databases of electronic health records for patients who satisfy a set of predefined criteria is known in medical informatics as phenotyping. Phenotyping has numerous use cases: clinical trial recruitment, outcome prediction, survival analysis, and other kinds of retrospective studies [1]. Several translational science initiatives have made identifying patient phenotype cohorts from the Electronic Health Records (EHR) their primary focus. Among them are Electronic Medical Records and Genomics (eMERGE) [2], Pharmacogenomics Network (PGRN) [3], and Informatics for Integrating Biology and the Bedside (i2b2) [4]. Within each of these initiatives, disease-specific phenotyping algorithms are developed and run against repositories containing millions of patient records. The identified patient cohorts are subsequently linked to biobanks for genetic analysis.

Supervised machine learning is currently the predominant paradigm in phenotyping [1]. Unfortunately, supervised learning can be very costly due to the expenses associated with the manual chart review. A separate machine learning model is developed for each phenotype, typically requiring hundreds of manually labeled examples. For instance, about 600 hundred patient records had to be reviewed manually for each disease within the i2b2 initiative. At the same time, thousands more were left unlabeled. Investigating cheaper alternatives to supervised learning that take advantage of the bountiful *unlabeled* patient records thus holds great promise.

Semi-supervised learning is a class of methods that is concerned with incorporating unlabeled data within machine learning models along with the data that have labels. In contrast with the general domain, semi-supervised learning has not received sufficient attention in the clinical NLP community. In this paper, we report on our experiments with several versions of the expectation maximization (EM) algorithm [5], a simple, yet powerful approach for semi-supervised learning. We experiment with four phenotyping datasets developed within the i2b2 initiative and analyze the interaction between the key aspects of semi-supervised learning that determine its success: the amount of labeled and unlabeled data and the relative weight of the unlabeled data. For our experiments, we adapt and implement several flavors of the EM algorithm [6,7] for modeling the i2b2 phenotypes. We first experiment with the basic EM algorithm. While the outcome of this experiment is promising, it becomes clear that unlabeled data often overwhelms the model, resulting in suboptimal parameter estimates. This shortcoming is ameliorated by introducing the weighting factor, allowing the unlabeled data to help characterize the parameter space without overwhelming the relatively small number of high quality gold labels. We investigate several alternatives for setting the value of the weighting factor including cross-validation and a simple heuristic. We discover that cross-validation is not always the optimal approach with the competing methods often leading to better models.

Our results indicate that overall unlabeled data can be highly beneficial to the models. We show that high-quality models can be trained with only a few hundred labeled examples when combined with unlabeled instances, potentially providing large savings in the amount of the required manual chart review. We also discuss several

practical lessons that can be drawn from our experiments. Our work is a bid to bring semi-supervised learning to the attention of the community by detailing a simple yet versatile baseline method.

The only study we are aware of that touches the subject of semi-supervised learning in the context of clinical NLP is the work by Garla et al. [8], who look at an application of Laplacian SVMs for detecting the presence of malignant liver lesions. While this work presents an encouraging result, it has several shortcomings that make it difficult to predict how well their findings would generalize to other tasks. First, the liver lesion model is based on complex rule-based features, whereas our model utilizes a simple bag of Unified Medical Language System (UMLS) [9] features that are successful in phenotyping [1,10–12]. Second, Garla et al. do not examine the effects of varying the amount of labeled data, which in our work leads to valuable insights about the practical aspects of semi-supervised learning. Finally, Garla et al. study a single dataset, employing a model with a large number of tunable parameters. In contrast, our work involves multiple datasets, while utilizing a simple and scalable model with a single tunable parameter (the weighting factor), the effects of which we carefully study. While our experiments with multiple datasets paint a more complex picture concerning the effects of unlabeled data, our conclusions are likely more generalizable and complete.

This paper is organized as follows. We first discuss our models and the EM algorithm. We then outline an experiment providing the intuition why EM works. Next, we examine the effects of varying the amount of labeled and unlabeled data and different methods for setting the weighting factor. Finally, we discuss our findings and draw practical lessons from our results.

Methods

Data representation

We perform our study in the setting where the unit of phenotype classification is the complete patient chart. We represent each chart as a set of UMLS concept unique identifiers (CUIs) which we extract from the patient's records using Apache Clinical Text Analysis and Knowledge Extraction System (cTAKES; ctakes.apache.org) [13]. CUIs are commonly used as representations in clinical NLP [12,14,15] as they help to abstract from lexical variability of medical terminology and capture the clinically relevant terms. Each CUI can be either asserted or negated, as determined by the cTAKES negation module. Some CUI examples employed in one of our models can be seen in Table 3.

Although cTAKES is capable of extracting most CUIs that exist in the UMLS, we only include the CUIs listed in phenotype-specific dictionaries. The dictionaries are created manually by i2b2 domain experts and define the relevant terms for each phenotype. Each phenotype-specific dictionary has D entries: CUI_1 ... CUI_D. We model a patient i as a D-dimensional vector of CUIs $\vec{x}_i$ in which an element $x_{i,k}$ indicates how many times CUI_k from the dictionary was seen in the patient's chart. Wu et al. [16] refer to this as "sum aggregation" and found it to perform well across performance metrics in a task requiring aggregating representations of patient charts.

Model

In the setting where all data have labels, model parameter estimation is fairly straightforward: for example, such methods as maximum likelihood or maximum a posteriori estimation can be utilized. When some of the labels are not known, parameter estimation is harder, although methods for obtaining parameter estimates exist. One such method is the Expectation Maximization (EM) algorithm [5]. The EM algorithm is an iterative procedure that begins by estimating model parameters from labeled data only. The model is then used to assign probabilistically-weighted class labels to the unlabeled instances. The model parameters are subsequently re-estimated from all (labeled and unlabeled) data and the procedure repeats until the parameter estimates stabilize. The algorithm is outlined in Table 1.

- **Input**: labeled and unlabeled data
- Estimate model parameters from labeled data only (Equations 3 and 4)
- Loop until convergence
  - **E-step**: use current model parameters to compute the class distributions for patients with no labels (Equation 2)
  - **M-step**: re-estimate model parameters from both labeled and unlabeled examples (Equations 3 and 4)
- **Output**: model parameter estimates

Table 1. Basic EM Algorithm.

The probabilistic framework we utilize in conjunction with the EM algorithm is based on a multinomial Naïve Bayes classifier, an approach frequently used in the context of text classification [17]. Phenotyping is cast as a binary patient classification task where a patient $\vec{x}_i$ has a class label $c \in \{case, noncase\}$. The total number of patients in the training set is $T = T_{labeled} + T_{unlabeled}$, where $T_{labeled}$ and $T_{unlabeled}$ are the number of labeled and unlabeled examples respectively.

The posterior probability distribution over the labels for a patient $\vec{x}_i$ is given by Bayes rule:

$$p(c|\vec{x}_i) = \frac{p(c)p(\vec{x}_i|c)}{p(\vec{x}_i)} \quad (1)$$

Incorporating the standard Naïve Bayes assumption, this equation can be factorized to include class-specific CUI distributions $p(cui_k|c)$ and the prior class probabilities $p(c)$ as model parameters:

$$p(c|\vec{x}_i) = \frac{p(c)\prod_{k=1}^{D} p(cui_k|c)^{x_{i,k}}}{\sum_{c' \in \{case,noncase\}} p(c')\prod_{k=1}^{D} p(cui_k|c')^{x_{i,k}}} \quad (2)$$

The parameters of this model can be computed using maximum a posteriori estimation:

$$p(cui_k|c) = \frac{1 + \sum_{i=1}^{T} x_{i,k} p(c|\vec{x}_i)}{D + \sum_{s=1}^{D}\sum_{i=1}^{T} x_{i,s} p(c|\vec{x}_i)} \quad (3)$$

$$p(c) = \frac{1 + \sum_{i=1}^{T} p(c|\vec{x}_i)}{2 + T} \quad (4)$$

Notice that the generative model that gives rise to these equations assumes that the patients are generated from a mixture of two components that correspond to the two classes (*case* and *noncase*). These assumptions are likely to be violated in practice. The patient pool used for selecting the data to be labeled could potentially contain patients with other medical conditions whose class specific CUI distributions could not be captured well by model parameters $p(cui_k|c)$. It could also be that the "natural" (highest probability) clustering of the data does not fit well into the *case/noncase* class boundaries we adopted for the phenotyping task. Depending on the extent to which the modeling assumptions are violated, the unlabeled data could prove to be detrimental to the performance of the model. In order to alleviate the negative impact of the violated assumptions, Nigam et al. [7] augment their model with a parameter λ that determines the contribution of the unlabeled data. We investigate several strategies for setting this parameter and their effect on the quality of the ensuing model.

The augmented version of EM is written as follows. We first define λ_i to be the weighting factor λ for the *unlabeled* examples and 1 for the *labeled* examples. We then rewrite the model parameters as follows:

$$p(cui_k|c) = \frac{1 + \sum_{i=1}^{T} \lambda_i x_{i,k} p(c|\vec{x}_i)}{D + \sum_{s=1}^{D}\sum_{i=1}^{T} \lambda_i x_{i,s} p(c|\vec{x}_i)} \quad (5)$$

$$p(c) = \frac{1 + \sum_{i=1}^{T} \lambda_i p(c|\vec{x}_i)}{2 + T_{labeled} + \lambda\, T_{unlabeled}} \quad (6)$$

This augmented EM algorithm is identical to basic EM in Table 1 except Equations 3 and 4 are swapped for Equations 5 and 6 respectively. The weight of the unlabeled data λ can be set via n-fold cross-validation, splitting the labeled data into a training and test sets and attempting different values of lambda. In addition to cross-validation, we introduce a heuristic that sets the weighting factor using a simple equation:

$$\lambda = \frac{1}{T_{labeled}} \quad (7)$$

The rationale for this heuristic is the following: the model should be able to obtain accurate parameter estimates from the labeled data alone, when a large amount of it is available. Thus the weighting factor should decrease as the amount of labeled data increases.

Datasets

We utilize four datasets all of which were created within the i2b2 initiative [12,15,18,19]. We show various important characteristics of our datasets in Table 2.

Phenotype	Cohort size (patients)	Manually labeled Instances (patients)
Ulcerative Colitis (UC)	33,465	600
Crohn's Disease (CD)	33,465	600
Multiple Sclerosis (MS)	172,447	595
Type II Diabetes (T2D)	198,002	596

Table 2. Phenotyping datasets used in the experiments.

Domain experts defined the ICD-9 codes relevant for each phenotype. These were then used to create the initial cohort from more than 6 million patient EHRs of the Partners Healthcare System. From that initial cohort, about 600 patients were randomly chosen for manual labeling. Each patient chart was reviewed by a domain expert and labeled at the patient level.

Obviously, only a small portion of the initial cohort could be labeled manually, leaving a large number of unlabeled examples for our semi-supervised learning experiments. For each experiment, we sampled randomly 500, 1000, and 3000 patients from the portion of the cohort that was never labeled.

Experimental Setup

The main objective of our evaluation is to determine whether unlabeled data can improve the classification accuracy of supervised models. The most natural baseline in these circumstances is, therefore, a supervised learning baseline in which the models are trained using labeled data only. During evaluation, we examine the behavior of semi-supervised learning with respect to the supervised baseline. This type of evaluation is best conducted via an analysis of learning curves as this allows us to compare the performance of supervised and semi-supervised models at different sizes of the training set.

To generate smoother curves, we utilize 10-fold cross-validation. In a typical experiment, within each fold we allocate a held-out test set and a pool of examples from which the labeled data is sampled. To produce a single point on a learning curve, we average the performance on the held-out test sets across all folds. To produce a point on the baseline curve, we train a Naïve Bayes classifier using labeled data only and evaluate its performance on the held-out test set. To produce a point on the semi-supervised learning curve, we add the unlabeled data, run the EM algorithm for 25 iterations[1], and use the resulting model to classify the test set. The resulting curves can be

[1] In our preliminary experiments, EM typically converged after 15-20 iterations.

compared visually or numerically in terms of the area under the curve (AUC) or increase in accuracy. For each phenotype, we generate the supervised baseline and semi-supervised learning curves for 500, 1000, and 3000 unlabeled examples.

We first experiment with the basic EM algorithm, which could also be viewed as a version of the augmented EM in which the weight of the labeled and unlabeled examples is the same. To illustrate the inner working of the EM algorithm, we examine the feature weights across EM iterations. Next, we experiment with augmented EM, down-weighting the unlabeled examples. Finally we experiment with setting the weight parameter using our heuristic technique and 10-fold cross-validation. During cross-validation, the held-out test set is, of course, not used; instead a validation set is allocated within each fold and each of the following lambda values is evaluated: 0, 0.05, 0.25, 0.50, 0.75, 1.0.

Due to the limit on the number of figures, we are not able to provide individual learning curves for the four phenotypes for *all* experimental conditions. Instead, we summarize each condition by plotting the curves that are averaged across all four phenotypes. These plots are supplemented by individual phenotyping plots for several important experimental conditions. All experiments reported in this paper are based on our own implementation of the EM algorithm and the supervised baseline.

Results

The averaged learning curves obtained in the experiment with basic EM algorithm are shown in Figure 1 (upper left).

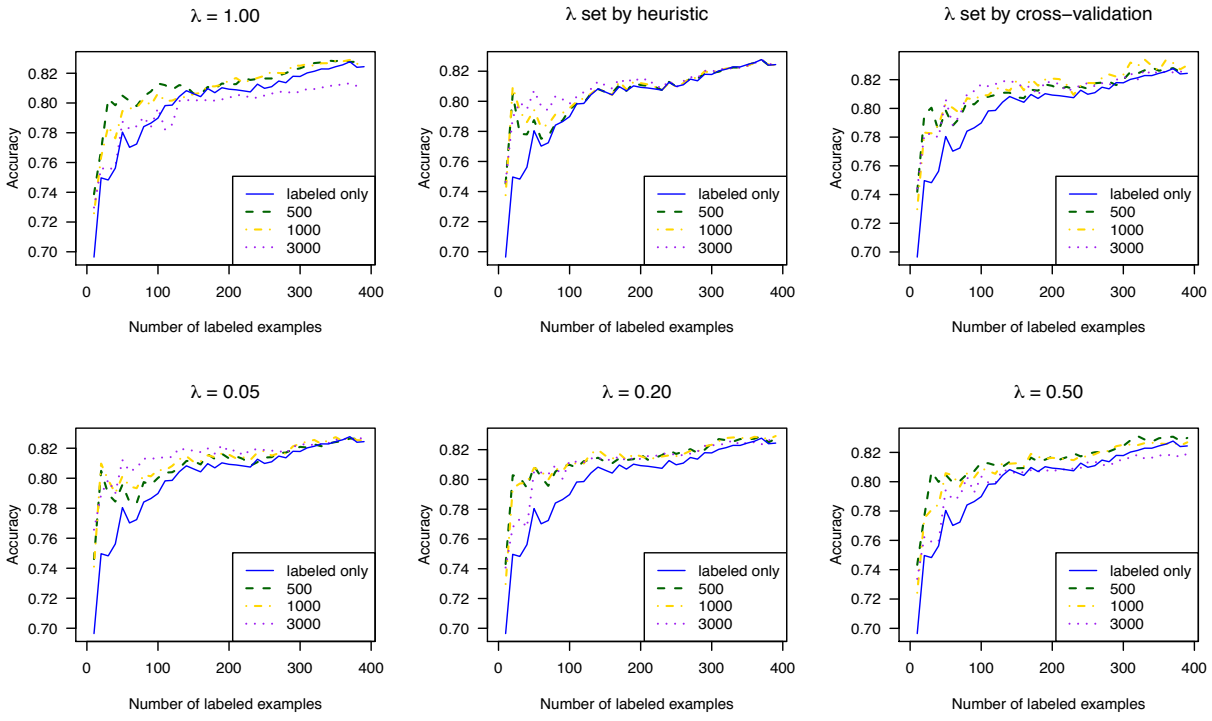

Figure 1. Average Learning Curves

The individual learning curves for each of the four phenotypes are in Figure 2. To take a peek "under the hood" of the EM algorithm, we train a Crohn's Disease model using 10 labeled examples. The top ten features selected using log-likelihood ratio are shown in the first two columns of Table 3. The top ten features selected using log-likelihood ratio after running EM for 25 iterations with 3000 unlabeled examples are in the last two columns of Table 3.

Labeled data only		Labeled and unlabeled data	
CUI	**Description**	**CUI**	**Description**
c0006826	Neoplasms malignant	-c1171255	Humira
c0032952	Prednisone	c2343521	Cimzia
c0030193	Pain	c1172734	Natalizumab
c0001418	Adenocarcinoma	c0001418	Adenocarcinoma
c0009324	Ulcerative colitis	c1171255	Humira
c0678172	Asacol	c1872109	Certolizumab pegol
-c00007097	Carcinoma	c2343521	Cimzia
-c1292819	Resection	c0162529	Ischemic colitis
c0007097	Carcinoma	-c0086492	J pouch
c0009410	Colostomy	c0678171	Pentasa

Table 3. Top ten features before and after execution of EM for Crohn's Disease. The '-' preceding the CUI indicates negation.

The results of running the augmented EM algorithm with λ set to 0.05, 0.20, and 0.50, are in the bottom row of Figure 1. We also provide the individual learning curves for λ=0.05 in Figure 3. Finally, the results of using our λ selection heuristic and 10-fold cross validation are shown in Figure 1 (top middle and top right); the individual phenotype learning curves for these conditions are also available in Figures 4 and 5 respectively.

One way to summarize these plots numerically is to examine the difference in the area under the curve (AUC) between the learning curve of a semi-supervised approach and the supervised baseline. Another is to compute the average improvement of a semi-supervised learning curve over the learning curve for the supervised baseline. We show both in Table 4 with the two metrics reported across phenotypes, number of unlabeled examples, and lambda selection criteria. The differences in AUC and the average improvement were computed across all training set sizes.

		$\lambda = 1.00$			$\lambda = 0.05$			**λ selection heuristic**			**λ cross-validation**		
		500	**1000**	**3000**	**500**	**1000**	**3000**	**500**	**1000**	**3000**	**500**	**1000**	**3000**
AUC difference	**CD**	3.50	8.61	7.99	1.27	3.58	5.30	1.00	2.26	3.25	2.23	5.95	5.35
	UC	2.22	-1.56	-5.59	2.79	4.14	5.49	1.80	2.03	2.71	4.06	3.80	4.71
	MS	6.92	-1.43	-14.99	6.28	6.31	6.74	3.67	4.60	5.78	6.48	3.57	4.09
	T2D	4.12	6.91	5.86	1.61	2.37	3.82	0.19	0.65	1.90	2.36	5.96	2.65
	Average	1.88			4.14			2.49			4.27		
Average Improvement, %	**CD**	0.91	2.31	2.19	0.32	1.01	1.50	0.29	0.68	0.97	0.56	1.57	1.43
	UC	0.60	-0.47	-1.57	0.79	1.07	1.38	0.50	0.51	0.63	1.11	0.98	1.19
	MS	1.98	-0.22	-3.79	1.84	1.83	1.92	1.15	1.37	1.62	1.81	1.02	1.19
	T2D	1.05	1.76	1.56	0.39	0.53	1.04	0.01	0.10	0.54	0.65	1.58	0.77
	Average	0.53			1.14			0.70			1.15		

Table 4. Difference in AUC and average improvement of semi-supervised learning over the supervised baseline.

Discussion

Incorporating unlabeled data into the model using basic EM algorithm shows very promising results, as evidenced by Figure 1 (top left). For Crohn's Disease dataset (Figure 2, top left), the fully supervised model trained on 10 labeled examples is only 74% accurate. With 3000 *unlabeled* examples added, the model receives a large performance boost: the accuracy is now 85%. It takes more than 300 more labeled examples for the supervised model to achieve the same level of performance. Clearly semi-supervised learning has a great potential, offering a large reduction in the amount of manual annotation that is required to train an accurate model.

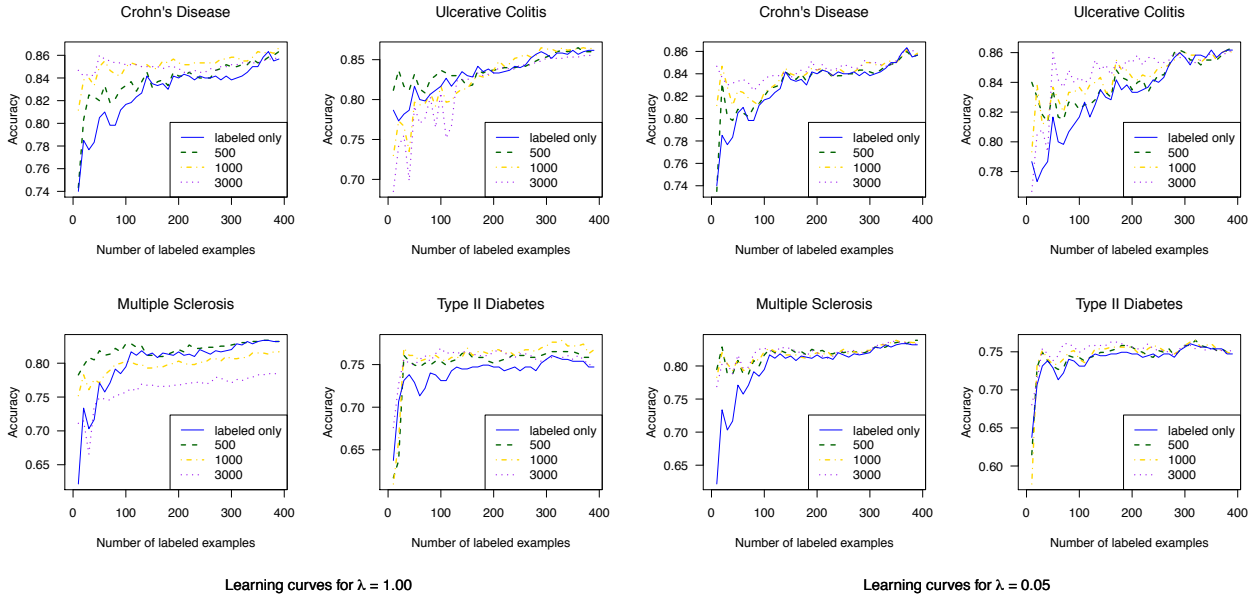

Figure 2. Learning curves for $\lambda = 1.00$

Figure 3. Learning curves for $\lambda = 0.05$

Why does EM algorithm work? The top ten features (Table 3, left) learned by the supervised model are simply too general to reliably identify a Crohn's Disease patient: notice the "Pain" and "Prednisone" features ("Prednisone" is a drug used to treat many inflammatory diseases). On the other hand, adding unlabeled data (Table 3, right) leads to a feature set that is much more specific to Crohn's Disease: "Pain" no longer appears on the list; Humira, Cimzia, Natalizumab, Certolizumab pegol, and Pentasa are all medications that are used to treat this condition. With few labeled examples, there is apparently not enough signal to capture the high number of drug names. However, by adding data that cluster with the positive labeled examples, the drug name features are found to be highly discriminative. Thus, unlabeled data helps to obtain better model parameter estimates.

However, the behavior of the remaining three phenotypes in Figure 2 paints a more complex picture. To remind the reader, Figure 2 presents the performance of the models where labeled and unlabeled instances are both given the same weight. While injecting a small amount of unlabeled data (500 unlabeled examples) improves the model performance, adding a larger amount (1000 and 3000 examples) is detrimental to classification accuracy for two out of four datasets we tried (Ulcerative Colitis and Multiple Sclerosis). This effect of unlabeled data is likely caused by violated model assumptions. When a large number of unlabeled examples is present, model parameter estimation is strongly influenced by the counts from the unlabeled data. The model essentially performs unsupervised clustering, using the labeled examples mostly to assign cluster memberships. When the most probable two-class clustering of the data does not match the annotated class boundaries, the unlabeled data can skew parameter estimation, potentially even leading to worse estimates than with labeled data only.

Introducing the λ parameter that downweights the unlabeled data leads to more accurate model parameter estimation. As we decrease λ, injecting additional unlabeled examples should more thoroughly describe the space we are classifying in without swamping the highly valuable gold standard instances. This results in more consistent improvements over the supervised baseline. Observe that basic EM (Figure 1, top left) did not seem to take full advantage of unlabeled data: increasing the amount of unlabeled data from 500 to 3000 resulted in a reduction in accuracy. By lowering the value of λ (Figure 1, bottom row), EM makes increasingly better use of unlabeled data, pushing the 3000 curve higher. At $\lambda = 0.05$ (Figure 1, bottom left; Figure 3) as the amount of unlabeled data is increased to 3000 examples, the semi-supervised models beat the supervised baseline for most training set sizes.

Applying our lambda selection heuristic results in a similar behavior (Figure 1, top center; Figure 4), although the overall gains over the supervised baseline are smaller both in terms of the AUC (2.49 vs. 4.14 for $\lambda = 0.05$) and the average improvement in accuracy (0.70 vs. 1.14 percent average improvement).

Using 10-fold cross-validation for lambda selection (Figure 1, top right; Figure 5) brings about overall gains that are comparable to using a small lambda, both in terms of AUC and the average improvement in accuracy. However, cross-validation does not appear to consistently pick the best lambda. For example, for the Crohn's Disease dataset, it is possible for the semi-supervised models to do better if $\lambda = 1.00$ is selected for all amounts of unlabeled data we evaluated, in terms of both AUC and average accuracy improvements. For Multiple Sclerosis (see Table 4), cross-validation does not seem to take full advantage of the unlabeled data: the model that has access to 500 unlabeled examples outperforms the models that have access to more unlabeled data. At the same time for $\lambda = 0.05$ the best result is achieved by the model that has 3000 unlabeled examples; that model also outperforms all models selected via cross-validation in terms of both AUC (6.74) and average accuracy improvement (1.92). Surprisingly, even the simple lambda selection heuristic makes better use of unlabeled data than cross-validation: the models that utilize the heuristic, outperform the models where lambda was selected through cross-validation when 1000 and 3000 unlabeled examples are included.

Overall, whether lambda is computed using our simple heuristic, using 10-fold cross-validation, or simply set to some small value, the unlabeled data has the strongest positive effect on model performance when the amount of labeled data is relatively small. For $\lambda = 0.05$ (Figure 1, bottom left) the best semi-supervised model reaches its top performance at around 150 labeled examples. It takes more than twice as many labeled examples for the supervised model to reach the same level of performance. Eventually, as the number of labeled examples in the training set grows, the models are capable of finding good parameter estimates from the labeled data alone and the learning curves for the supervised and semi-supervised cases become very similar.

These findings are relevant to scaling up phenotype algorithm development for projects such as eMERGE, PGRN, and now BD2K. Domain expert time is always limited, and if a maximum performance can be achieved by

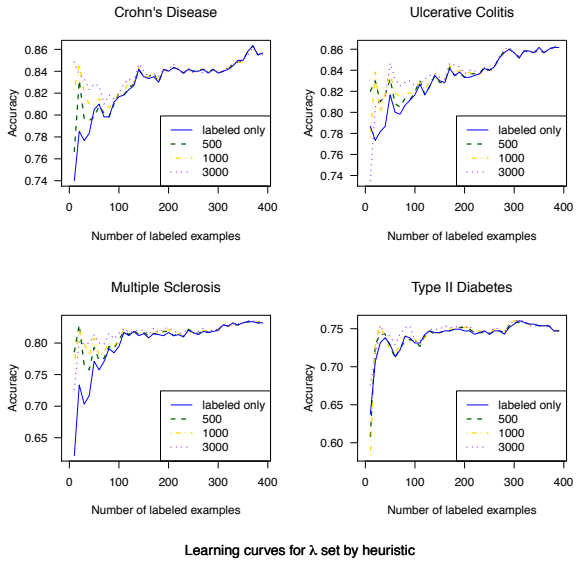

Figure 4. Learning curves for λ set by heuristic.

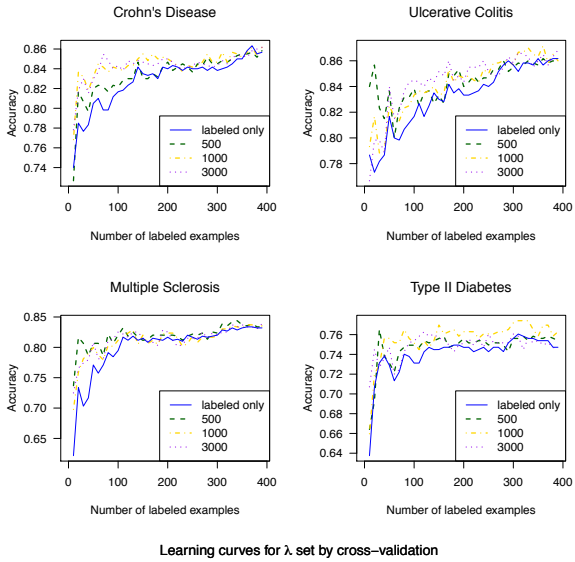

Figure 5. Learning curves for λ set by cross-validation.

combining a small seed of label instances (100-200) with unlabeled data, that would lead to gains in efficiency. If

the domain expert time is plentiful, then large amounts of labeled data would be preferable as the completely supervised model will reach accurate parameter estimates.

Conclusion

Even though for some datasets the basic version of EM did not succeed in beating the supervised baseline for larger amounts of unlabeled data, it appears to be possible to overcome this shortcoming. Introducing the lambda parameter that downweights the unlabeled data leads to more consistent gains over the supervised baseline. Clearly, semi-supervised learning can be of high importance in practice. Are there any practical lessons we can learn from the experiments we presented?

To answer this question, let us first identify two properties of a model that makes use of unlabeled data that a practitioner of semi-supervised learning would find desirable. First, a semi-supervised model should not perform worse than the corresponding supervised baseline. Clearly, if adding unlabeled data is as likely to lead to performance deterioration as to improvement, it is too risky to use in practice. Second, the semi-supervised model should make a full use of unlabeled data that is available to it. Out of two models, we prefer the one that can "squeeze" more out of unlabeled data, resulting in larger gains over the supervised baseline.

In light of these properties, we can draw several lessons from our experiments. It appears that semi-supervised learning of the type we discussed in this work has a higher chance of success if augmented EM is used. A priori, the most obvious approach to setting its lambda parameter would be to use n-fold cross-validation. Empirically, simply setting lambda conservatively (i.e. to some small value) works a little better, possibly due to the sensitivity of semi-supervised learning to the amount of labeled data (n-fold cross-validation has to allocate some portion of the labeled data as a validation set for parameter evaluation). The simple lambda selection heuristic we introduced also has some desirable properties over n-fold cross-validation: it often results in more consistent improvements over the supervised baseline and for larger training set sizes it results in performance level that is no worse than the supervised baseline. At the same time, it is much more efficient, which may be important for rapid development of multiple phenotyping models. Finally, the heuristic does not require the implicit supervision of a human-generated list of acceptable weights as in cross-validation. If the size of the unlabeled data were increased by an order of magnitude, the cross-validation procedure may require manual adjustments to its inputs to find a suitable value.

For future work, we are planning to examine the behavior of semi-supervised learning under the scenario where an order of magnitude more unlabeled data is available to the models. We will also investigate the use of Bayesian inference methods for model parameter estimation such as Gibbs sampling.

Acknowledgements

The study was funded by U54LM008748 (i2b2), 1R01GM103859-01A1 (PGx), 1U24CA184407-01 (DeepPhe), and R01GM090187 (ShARe).

References

1 Shivade C, Raghavan P, Fosler-Lussier E, *et al.* A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 2013;**21**:221–30. doi:10.1136/amiajnl-2013-001935

2 McCarty CA, Chisholm RL, Chute CG, *et al.* The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med. Genomics. 2011;**4**:13. doi:10.1186/1755-8794-4-13

3 Long RM, Berg JM. What to expect from the Pharmacogenomics Research Network. *Clin Pharmacol Ther* 2011;**89**:339–41. doi:10.1038/clpt.2010.293

4 Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J. Am. Med. Informatics Assoc. 2012;**19**:181–5. doi:10.1136/amiajnl-2011-000492

5 Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B* 1977;**39**:1–38. doi:10.1.1.133.4884

6 Nigam K, McCallum AK, Thrun S, *et al.* Text Classification from Labeled and Unlabeled Documents using EM. *Mach Learn* 2000;**39**:103–34. doi:10.1023/A:1007692713085

7 Nigam K, McCallum A. Semi-Supervised Text Classification Using EM. In: *Semi-Supervised Learning*. 2006. 33–54.

8 Garla V, Taylor C, Brandt C. Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management. *J Biomed Inform* 2013;**46**:869–75. doi:10.1016/j.jbi.2013.06.014

9 Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. *J Biomed Inform* 2003;**36**:414–32. doi:10.1016/j.jbi.2003.11.002

10 Lin C, Canhao H, Miller T, *et al.* Feature engineering and selection for rheumatoid arthritis disease activity classification using electronic medical records. In: *ICML Workshop on Machine Learning for Clinical Data Analysis*. 2012.

11 Lin C, Karlson EW, Canhao H, *et al.* Automatic Prediction of Rheumatoid Arthritis Disease Activity from the Electronic Medical Records. *PLoS One* 2013;**8**.

12 Xia Z, Bove R, Cai T, *et al.* Leveraging Electronic Health Records for Research in Multiple Sclerosis. In: *European Committee for Treatment and Research in Multiple Sclerosis (ECTRIMS)*. 2012.

13 Savova GK, Masanz JJ, Ogren P V, *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;**17**:507–13. doi:10.1136/jamia.2009.001560

14 Liao KP, Cai T, Gainer V, *et al.* Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res* 2010;**62**:1120–7. doi:10.1002/acr.20184

15 Ananthakrishnan AN, Cai T, Cheng S, *et al.* Improving Case Definition of Crohn's Disease and Ulcerative Colitis in Electronic Medical Records Using Natural Language Processing: a Novel Informatics Approach. *Gastroenterology* 2012;**142**:S – 791.

16 Wu ST, Juhn YJ, Sohn S, *et al.* Patient-level temporal aggregation for text-based asthma status ascertainment. *J Am Med Inform Assoc* 2014;:1–9. doi:10.1136/amiajnl-2013-002463

17 McCallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification. In: *AAAI/ICML-98 Workshop on Learning for Text Categorization*. 1998. 41–8. doi:10.1.1.46.1529

18 Ananthakrishnan AN, Gainer VS, Perez RG, *et al.* Psychiatric co-morbidity is associated with increased risk of surgery in Crohn's disease. *Aliment Pharmacol Ther* 2013;**37**:445–54.

19 Ananthakrishnan A, Gainer V, Cai T, *et al.* Similar risk of depression and anxiety following surgery or hospitalization for Crohn's disease and ulcerative colitis. *Am J Gastroenterol* 2013.