Unsupervised Document Classification with Informed Topic Models

Timothy A. Miller[1] and **Dmitriy Dligach**[2] and **Guergana K. Savova**[1]
[1] Boston Children's Hospital Informatics Program, Harvard Medical School, Boston, MA 02115
{firstname.lastname}@childrens.harvard.edu
[2]Department of Computer Science, Loyola University Chicago, Chicago, IL 60611
ddligach@luc.edu

Abstract

Document classification is an important and common application in natural language processing. Scaling classification approaches to many targets faces a bottleneck in acquiring gold standard labels. In this work, we develop and evaluate a method for using informed topic models to noisily label documents, creating a noisy but usable set of labels for training discriminative classifiers. We investigate multiple ways to train this noisy classifier, and the best performing method uses Wikipedia-seeded topic models to approximately label training instances without any supervision. We evaluate these methods on the classification task as well as in an active learning setting, in which they are shown to improve learning rates over traditional active learning.

1 Introduction

Document classification is a standard task in machine learning and natural language processing which has been studied extensively (Joachims, 1999; Sebastiani, 2002). For many instances of this problem, standard supervised machine learning methods are now sufficient, so that any given document classification problem may be considered an application or engineering task rather than an interesting research problem. Recent work related to this problem has come mainly from the machine learning community and has focused on a generalization of the task called multi-label classification, in which each instance has multiple categories that must be predicted (Tsoumakas and Katakis, 2007; Read et al., 2011). That work has been concerned with the problem of how to best make use of correlations between the different labels, and using that information to perform the classifications non-independently.

In contrast, the work here is concerned with the more practical problem of obtaining these labels, and particularly the issue that ad hoc classification targets require obtaining supervised training data from scratch. This problem may arise in any application area of natural language processing, but in the clinical domain this problem is potentially more pressing because expert annotators (physicians) are expensive and traditional cost-saving approaches such as crowdsourcing are not always viable due to privacy concerns.

A common use case for clinical document classification is physicians mining patient notes for diseases, then using genetic samples of that "virtual cohort" to do phenotype-genotype correlation studies. Billing codes have high recall but varying precision depending on the disease. Thus, machine learning and NLP applied to the narrative text in the clinical record are now often used as a solution to this problem.

Our approach to this task is to use the unsupervised method of topic modeling, specifically Latent Dirichlet Allocation (LDA), which can learn word probabilities for semantically coherent topics, and by providing informed priors, we can steer topics to categories of interest and use these word lists like features in a classifier. As a first step, we take advantage of the crowd-sourced knowledge

Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pages 83–91,
Berlin, Germany, August 12, 2016.

contained in Wikipedia to build a representation of the category of interest. We then use this category representation as an informed prior to LDA. This informed LDA algorithm then finds the topics that best satisfy the data given the priors, including both informed topics and traditional uninformed topics. In particular, we are able to guide the topic model to learn separate topics for similar categories if that is required by the categories we are interested in for classification.

The ability to extract pre-specified topics of varying granularity is interesting on its own, as it could be used for more guided data explorations of the kind that LDA is already in use for. But we can also use the output of this process to generate classifiers, by treating the occurrence of these topics in a document as a noisy label for that document. Given these noisy labels, we can immediately train a classifier, which performs much better than chance, without seeing a single gold standard training example.

Finally, we show that this has potential applications to active learning by using our noisy classifier's certainty estimates to select training examples, rather than first annotating a random seed set. This method results in faster learning rates than passive learning, standard active learning, and a baseline method that uses the Wikipedia-trained priors directly.

2 Background

2.1 Topic Modeling with Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a probabilistic unsupervised method for grouping tokens into a set of corpus-wide clusters. By setting parameters that constrain each document to use a subset of the clusters, frequently co-occurring words tend to get placed into the same clusters, and since distributionally similar words are often semantically similar, the result is that the clusters are often semantically coherent *topics*.

A document in LDA is represented as a bag of words. Each document has a probability distribution across K topic indices, and each topic is a global probability distribution across V words in the vocabulary. This leads to a generative story where the topic distribution for a document is drawn from a Dirichlet distribution, and each word is generated by first drawing a topic from the topic distribution, then drawing a word from the word distribution indexed by that topic. One common inference method for LDA is to use Markov Chain Monte Carlo sampling, which is an iterative algorithm where each variable of interest is sampled probabilistically. In LDA, the standard sampling algorithm is derived by integrating out the topic and word distributions from the joint probability, so that the only random variable left to sample is the topic assignment for each word. Each topic assignment is typically randomly initialized, then at each iteration a topic is sampled from the sampling equation (from Griffiths and Steyvers (2004)):

$$p(z_i = j|\mathbf{z_{-i}}, \mathbf{w}) \propto \frac{n^{w_i}_{-i,j} + \beta}{n^{(\cdot)}_{-i,j} + |V|\beta} \cdot \frac{n^{d_i}_{-i,j} + \alpha}{n^{d_i}_{-i,\cdot} + T\alpha} \quad (1)$$

where i indexes words in the corpus and j is an index into K topics. The first factor represents the probability of the given word being selected for this topic ($n^{w_i}_{-i,j}$ is the count of the word at position i in topic j). The second factor represents the probability of topic j being selected for a word in this document ($n^{d_i}_{-i,j}$ is the count of words in the same document as w_i with topic j). α and β are the hyper-parameters from the Dirichlet priors used to draw the probability distributions. While the Dirichlet distribution accepts a vector of hyper-parameters the size of the output distribution, in most work these hyper-parameters are symmetrical, and are set using intuition or experimentation. Low values of these parameters (≤ 1) encourage sparse distributions, and sparsity constraints give rise to the clustering behavior typical of LDA.

One limitation of standard LDA in practice is that it will not always make fine-grained distinctions, even if they are known to exist in the data. For example, in the 20 Newsgroups data set (described in Section 4.2), there are different topics for baseball and hockey, which share quite a bit of terminology (teams, games, scores, etc.) but users may wish or expect them to be separate. Running standard topic modeling on this corpus with number of topics $K = 25$ using the Mallet topic modeling framework (McCallum, 2002), we observe that one topic seems to have merged baseball and hockey terminology (top words in that topic: *game, team, year, play, games, hockey, season, players, ca, win, league, baseball, nhl*). Simply increasing the number of topics may solve the problem but will also have the general effect of making categories more specific, which may adversely affect other topics. This problem can also

be addressed by hierarchical models, in which topics that are higher in some hierarchy tend to model more general terms and lower topics are more specific. Hierarchical topic models (Blei et al., 2003) make use of a nested Chinese Restaurant Process where a word is a sample from a mixture between all the topics in a path from the root to a leaf node in a topic tree. Higher-level nodes will tend to be on more paths, and will thus be sampled more often and contain higher probability words. This method can then, for example, run on text without stop words removed and recover them as the top level of the hierarchy. One might imagine that for the baseball and hockey example, a hierarchical model would recover a higher-level sports topic with lower level topics specific to baseball and hockey.

Another method, Pachinko Allocation Model (Li and McCallum, 2006), generalized hierarchical topic models so that the topic hierarchy did not need to be a tree. This retains a hierarchy with higher and lower nodes corresponding to more or less general topics, but also allows for different words to be generated by different topic paths. While these hierarchical models have attractive properties, they are significantly more complex than standard LDA, which means they have more parameters, may take longer to train, and still may not recover topics of interest to a user.

Some relevant recent work in topic modeling has explored the importance of the prior values α and β. Wallach et al. (2009) developed optimization procedures for α and β and found that optimization of the document-topic prior α led to improved results, as measured by perplexity on held-out data. Jagarlamudi et al. (2012) found that priors on both α and β allowed them to incorporate information into the LDA inference, though they found that a more complex model structure was necessary to properly incorporate the information, which requires a more complex inference procedure.

Other relevant topic modeling work involves the augmentation of LDA-style models for labeling documents with multiple topics. Labeled LDA (Ramage et al., 2009) creates a topic for each label in a multi-label setting, and takes advantage of gold standard labels to learn topic distributions for each label. In the author-topic model (Rosen-Zvi et al., 2004), a document is generated by a set of authors, and an author is a distribution over topics. While both models are relevant to the multi-label classification problem, they both require gold standard labels, and we suspect that given gold standard labels discriminative classifiers will be superior.

3 Methods

Building on this existing work in topic modeling, we propose an extension to the LDA model that is able to find specific topics of interest, with minimal human effort. We call this method informed LDA, and the following sections will describe the method and how it can be used to train classifiers.

3.1 Building Informed Priors

We first build models for each of the target labels we are interested in. For this work, we use topics from two corpora, the 20 Newsgroups dataset mentioned above, as well as the 2008 i2b2 Challenge dataset[1], a set of 730 clinical discharge summaries labeled for multiple obesity-related diseases. Table 2 shows the 14 i2b2 labels we used for this work.

To build these models, we retrieved the Wikipedia article closest in meaning to each label. For most labels, there was an article with the exact title or a very similar title. We tokenized the articles and then TF-IDF (term-frequency/inverse document frequency) weighting was applied to these tokens (for the clinical articles we used an IDF derived from a sub-index of Wikipedia articles containing clinically relevant articles). The purpose of the TF-IDF reweighting is to down-weight commonly occurring words like those representing broad terms (especially in the clinical data, terms like ”disease” ”surgery” are not as informative as they are generally).

While in the present work the step of identifying the relevant Wikipedia article required a small amount of manual effort, there are many ways that it could be automated – for example, by querying Wikipedia or the Web with the category name and performing token counts over multiple retrieved articles. Performing this step manually and obtaining high quality models of each category allows for a purer evaluation of the more technically challenging downstream steps.

[1]The i2b2 Challenge datasets are publicly available with a Data Use Agreement at https://i2b2.org/NLP/DataSets/.

3.2 Informed LDA

The standard LDA sampling equation, Equation 1, has a single value of α and β, assuming symmetric Dirichlet priors. A simple extension of the sampling equation for arbitrary priors can be obtained by vectorizing $\vec{\alpha}$ and $\vec{\beta}$:

$$p(z_i = j|\mathbf{z_{-i}}, \mathbf{w}) \propto \frac{n^{w_i}_{-i,j} + \beta^{w_i}}{n^{(\cdot)}_{-i,j} + \sum_{i'=1}^{|V|} \beta^{i'}} \cdot \frac{n^{d_i}_{-i,j} + \alpha^j}{n^{d_i}_{-i,\cdot} + \sum_{j'=1}^{K} \alpha^{j'}} \quad (2)$$

Since each article has a different length, the prior vectors are first normalized so that all informed priors have equal strength. The $\vec{\beta}$ parameters are then filled in by these normalized weights. For token values that are not in the article, we use a default value of 0.01 in the prior.

To do inference using this model, we modified the source code of Mallet (McCallum, 2002), allowing for arrays of priors and modifying the sampling equation as described above. The number of topics K was set to 30, as this value gave reasonable results during preliminary experiments with standard LDA. This means that, in contrast to methods like Labeled LDA, not all topics are associated with a label – 16 of our topics were informed and the remaining 14 are uninformed, allowing the model to fit other topics in the data that we may not be currently interested in.

3.3 Creating a Bronze Standard

After running inference on the informed LDA model, the output of interest is the empirical estimates of document-topic probability – frequencies of each topic in each document. For example, the output may say that in document 0, the topic for *asthma* accounted for 3% of the tokens. Our goal is to use these values to assign noisy labels to each document for the value of that cluster category. We call this set of labels a *bronze standard*, in contrast to the *gold standard* of expert-generated labels.[2]

There are many ways one might go about converting topic frequencies to labels. For binary classification, as in the i2b2 data, one could set a threshold value and give all documents with topic frequencies above that threshold a positive label. Possible thresholds include 0 and $1/K$. We found that thresholds allowed for too much variation, and led to some severely skewed label distributions, so that the next stage classifier may have only a few positive examples to work with. Even if this approximates a true distribution, it is probably not enough data points to find a signal in the features, and so the resulting classifier probably will not be useful.

Another option is, for each informed topic, sort all documents by that topic's frequency, and then split the data at the median frequency value into the *true* and *false* classes, so that the classifier gets training data with no skew in its distribution. We found that this method was the most reliable across labels and does not require fitting any parameter.

For multi-way classification, as in the 20 Newsgroups data, we use as the bronze label the topic whose document-topic probability was the maximum of all the informed topics. To simplify further, this is just the topic that accounts for the greatest number of words in the document.

3.4 Building a Classifier

However bronze labels are obtained, they can now be used in the typical way to train a classifier. The feature representation may also be varied. We will describe classifier settings in detail in the Evaluation, but we experimented with a variety of classifiers. The representation used here is bag of words for a document.

3.5 Active Learning with Bronze and Gold Labels

While this technique may have value as a low-cost low-accuracy classifier, we suspect that it might have additional value as an input to other systems. One such potential application is as an input to an active learning-based annotation system, as a way of obtaining gold standard labels to achieve optimal classification performance. Active learning is an annotation technique that has a classifier in the loop – instead of labeling examples randomly, examples are selected for labeling based on some notion of usefulness, such as classifier uncertainty (see Settles (2010) for an excellent overview of active learning). To initialize the active learning classifier, however, a small set of "seed" examples are randomly selected to be labeled.

Here, instead of using a seed set, we use our unsupervised classifiers from the start of the annotation process. Using this bronze-trained classifier,

[2] *Silver standard* is already used to describe huge automatically labeled datasets (Rebholz-Schuhmann et al., 2010).

we get a probability distribution across categories for every instance in the training data. We use uncertainty sampling to select the next instance, which in the two-class case means selecting the instance whose classification probability is closest to 0.5. The bronze standard label is then replaced with the gold standard label (simulating annotation of the instance) and the classifier is re-trained. This process repeats until every instance in the training data has its gold label uncovered. To test this method experimentally (as in Section 4), we would also have a set of held out data, and every time we train a classifier we would evaluate it on this held out data.

The active learning method just described differs from standard active learning in that there is no longer a breakdown into initial seed set and a pool set from which examples are drawn, but rather we have a mixed gold/bronze training set. Since the gold labels are more reliable, we give them a higher weight relative to the bronze labels, so the classifier can treat them differently.

4 Evaluation

To evaluate the effectiveness of this method, we will start with one brief qualitative evaluation to inspect the topics found, and then proceed to two quantitative evaluations. The first evaluation attempts to get a preliminary look at how the informed LDA method works on topics that are superficially similar, to gauge how adding information can guide the model to make difficult distinctions. The first quantitative evaluation is a simple set of unsupervised classification experiments. We build a bronze standard for the 20 Newsgroups and i2b2 data sets, then train classifiers for each category and evaluate the classifier. The second quantitative evaluation examines the use case of active learning. Our experiment uses the unsupervised classifiers from the previous experiment to evaluate whether active learning can be made even faster by using those classifiers to select examples at the start of the active learning procedure, when the gold standard training data is still quite small.

4.1 Qualitative Inspection of Similar Topics

Table 1 shows the results of inspecting a few sports-related topics from the 20 Newsgroups corpus. This is to simply see if this method can address the issue discussed in Section 2, the conflation of similar topics. The first column shows the words in a sports-related topic using standard LDA. It clearly finds words related to both hockey and baseball, with no other topics containing any significant amount of hockey or baseball content.

LDA	Baseball	Hockey
game	year	team
team	baseball	game
year	hit	hockey
play	san	play
games	win	canada
hockey	team	games
season	season	toronto
players	runs	nhl
ca	league	cup
win	game	players
league	won	division
baseball	lost	season
nhl	games	gary

Table 1: Comparison of topic words in similar topics with standard LDA (first column) and informed LDA (last two columns).

In contrast, the last two columns show the informed topics for baseball and hockey using informed LDA. In addition to the sport names there are additional terms that are discriminative, including *hit*, *runs*, and *wins* (a pitching statistic) for baseball, and *canada*, *nhl*, and *cup* for hockey. Informed LDA also did not have any other topics containing significant amount of hockey or baseball content. This kind of evaluation is of limited use, but it does verify that the algorithm is able to find closely aligned topics.

4.2 Experimental Configuration

The data sets used for evaluation are the 20 Newsgroups data set[3] and the 2008 i2b2 Challenge data set described above. The 20 Newsgroups data set contains around 11,000 training documents, partitioned into 20 topics, which are used as labels for the documents. These include labels such as *alt.atheism* for atheism-related conversations, *sci.crypt* for cryptography-related discussions, and so forth. While each document may have multiple "topics" in the strict semantic sense, it will have one topic *label* – in other words, a classifier must choose a single category from 20 possibilities.

The 2008 i2b2 Challenge data consists of clin-

[3]This data set can be downloaded here: http://qwone.com/ jason/20Newsgroups/

ical discharge summaries from patients at an obesity clinic. This data contains 730 notes in the training set, with each note being labeled for 16 disease categories, with both textual and intuitive labels.[4] We use the more challenging intuitive label set, which did not require explicit confirmation of a diagnosis in the text. We discard two labels, hypertriglyceridemia and venous insufficiency, after preliminary work on the training set indicated that those two labels could not be learned satisfactorily even with fully supervised approach. The likely cause of the difficulty is that these two categories contained the fewest number of positive examples, an important issue but one we will have to reserve for future work. In contrast to the 20 Newsgroups data, in the i2b2 data the labels are not mutually exclusive, so we frame the task as 14 binary classification problems.

We used the Weka machine learning toolkit (Hall et al., 2009) during development, and evaluated many different classifiers on both datasets, including Adaboost, support vector machines, logistic regression, and naive bayes. We use the Adaboost algorithm (Freund and Schapire, 1996) with decision stumps as the weak learner for the i2b2 data. For the 20 Newsgroups data we used a support vector machine with linear kernel for the classification experiment and switched to Naive Bayes for the active learning experiments for speed reasons. Besides being relatively accurate, using boosting with decision trees has the beneficial property that the models it builds have some degree of transparency, which clinical researchers appreciate.

For the first experiment we evaluate the effectiveness of informed LDA on generating labels that can train a classifier. We compare first to a random labeling baseline (labeled *RandL*), that generates a random labeling, trains a classifier with those labels, and then uses it to classify the training set. This is not intended to be a competitive baseline, as much as it is a check to set a lower bound on what kind of performance we would get if informed LDA labeling had no signal whatsoever. We also compare to a standard random classifier (*RandC*) which is based on a recall of 0.5 and a precision of the category's prevalence. This baseline is important for the binary classifier to make sure our classifier is learning more than just how to do random guessing based on our evenly split labels. In the main experimental condition (*Bronze*), we use informed LDA to generate a bronze standard label set for the training data as described in Section 3.3, train a classifier with those labels and evaluate it on those same examples from the training set. The upper bound we compare against is a 5-fold cross-validation of the training set using gold labels.

The next experiment examines the usefulness of these unsupervised classifiers in an augmented active learning scheme described in Section 3.5. We use the two baselines of passive learning and standard active learning. The passive learning baseline is equivalent to just plotting a learning curve for a machine learning problem with random ordering of the instances. The active learning baseline uses an initial seed set of 25 examples from within the pool set. We use uncertainty sampling to select the next example, which uses the example which has the smallest difference in probability estimates between the two most likely classes.

The condition we are testing is labeled *Bronze*. This condition does not use a seed set, but starts with a classifier trained on the entire bronze-labeled pool set. Learning proceeds by finding examples in the pool set that the current iteration of the classifier is uncertain about and uncovering the gold label (i.e. simulating annotation). This means that, in the active learning curve, the x-axis, which traditionally indicates the size of the training data used to train the classifier, now indicates the number of gold instances in the training data (the remaining instances still have bronze labels).

We give gold and bronze instances different weights to reflect varying quality of the labels. This weight is used in calculating the cost function during training – a higher weight on gold labels means the classifier will try harder to get gold-labeled instances correct. Here we use a weight of 0.1 for bronze-labeled instances and a weight of 1.0 for gold-labeled instances.

4.3 Results

Table 2 shows the results of the 14 binary classifiers on the i2b2 data. The random labeling gives rise to a classifier that never obtains an F1 score better than 0.11. The bronze labeling, performs much better than the RandL classifier, with a low performance of 0.26 (for depression) and a high performance of 0.83 (for diabetes). The bronze-

[4] In actuality, not every note is labeled for every category, but most are.

Category	RandL	RandC	Bronze	CV
Asthma	0.02	0.20	0.47	0.91
CAD	0.05	0.54	0.66	0.91
CHF	0.07	0.50	0.75	0.86
Depression	0.07	0.29	0.26	0.77
Diabetes	0.06	0.58	0.83	0.95
Gallstones	0.02	0.22	0.33	0.83
GERD	0.04	0.33	0.39	0.77
Gout	0.05	0.21	0.42	0.88
HC	0.08	0.51	0.56	0.83
HTN	0.06	0.62	0.67	0.96
OA	0.10	0.26	0.27	0.66
Obesity	0.07	0.46	0.56	0.97
OSA	0.02	0.22	0.28	0.91
PVD	0.11	0.25	0.37	0.76

Table 2: F1 scores for traditional supervised classifier (CV) vs. unsupervised classifier trained using informed LDA (Bronze), classifiers trained with random labels (RandL), and a classifier that makes random guesses (RandC). (CAD=Coronary Artery Disease, CHF=Congestive Heart Failure, GERD=Gastroesophageal Reflux Disease, HC=Hypercholesterolemia, HTN=Hypertension, OA=Osteoarthritis, OSA=Obstructive Sleep Apnea, PVD=Peripheral Vascular Disease)

	RL	RC	Bronze	CV
Accuracy	0.05	0.05	0.64	0.85

Table 3: Multi-way classifier accuracy on the 20 Newsgroups dataset using random labels (RL), a random classifier (RC), bronze labels obtained from informed LDA (Bronze) and a supervised cross-validation (CV).

trained classifier also outperforms the RandC random classifier in 13 out of 14 categories, by an average of approximately 12 points F1 score. Cross-validation using the gold standard can be very accurate, ranging from 0.66 to 0.96.

There are a few interesting things to point out from these results. First, our analysis of the errors shows that the classifiers trained by the bronze labeling did not systematically favor either precision or recall. A linear regression with the Gold score as the independent variable and the Bronze score as the dependent variable shows that the Gold score is a statistically significant predictor of the Bronze score ($p = 0.01$), but with so few data

	Active Learning		
Disease	Passive	Active	Bronze
Asthma	469.6	486.1	**500.1**
CAD	455.5	462.7	**469.6**
CHF	415.3	422.3	**435.1**
Depression	371.7	**414.5**	410.4
Diabetes	491.7	503.2	**510.8**
Gallstones	400.7	450.8	**457.0**
GERD	317.3	350.0	**360.2**
Gout	477.4	506.1	**519.1**
HC	372.4	389.4	**398.9**
Hypertension	460.1	**476.2**	465.1
Osteoarthritis	305.6	328.7	**349.9**
Obesity	490.4	501.9	502.0
OSA	463.6	487.7	**495.8**
PVD	363.0	**390.9**	372.2
Average Curve	451.0	473.8	**479.8**

Table 4: Performance of augmented active learning on 14 categories from the 2008 i2b2 Challenge data .

	Passive	Active	Bronze
ALC	7203	7469	**7678**

Table 5: Performance of augmented active learning on 20-way classifier for 20 Newsgroups data. Unit is Area Under the Learning curve (ALC).

points the exact nature of this effect is not clear.

Table 3 shows the results of the three classifiers on the Newsgroups data. For this multi-category experiment we use accuracy as the metric instead of F1 score. Here the accuracy of the RandL and RandC are both quite low, at 0.05. Bronze labeling can train a classifier that attains an accuracy of 0.64. The Gold labeling gives us an approximate ceiling performance of 0.85.

Table 4 shows the area under the active learning curve (ALC) for 14 categories in the i2b2 data under three conditions. Both the active learning and the bronze-augmented active learning outperform passive learning on all 14 categories. In 11 of the 14 categories the bronze-augmented version is superior to traditional active learning. We also averaged the curves together and computed the average learning curve, for which the bronze-augmented algorithm is again optimal.

Figure 1 shows the average learning curve across i2b2 category labels. The x-axis has been truncated at 100 instances to clarify the region

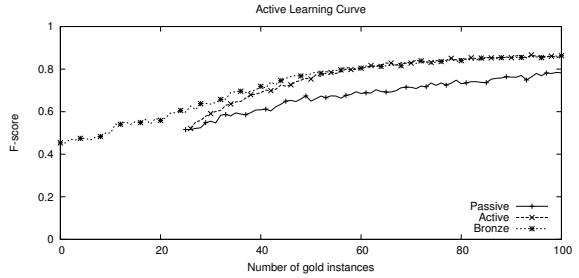

Figure 1: Average active learning curve across 14 disease categories.

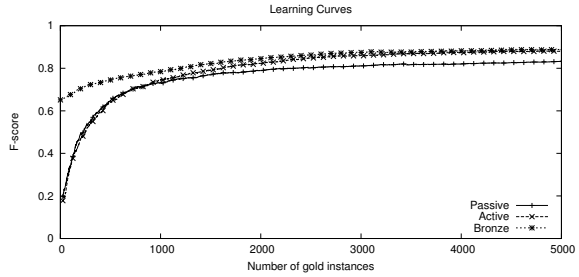

Figure 2: Active learning plot for 20 Newsgroups data.

where there is a clear distinction. Qualitatively, the distinction between passive and active is quite clear – this much is not surprising, given previous success in active learning. While the bronze curve shows an advantage up to maybe 30 instances, it quickly converges with the active curve.

Table 5 shows the Area Under the Learning curve (ALC) results of active learning on the 20 Newsgroups data. Active learning again beats passive learning, and the augmented version using bronze labels performs best. The learning curves for all three conditions are in Figure 2, truncated to 5000 examples to highlight the area showing the most difference. Here the bronze label-based version of active learning seems to have a clearer advantage than in the i2b2 corpus.

4.4 Discussion and Future Work

One aspect that deserves further mention is that of class prevalence and skew. The decision to assign bronze labels on the i2b2 corpus with an even class distribution was vastly superior to any thresholding that was attempted. However, we should note that in the i2b2 data we used, the prevalence is relatively high for most categories. Diabetes, for example, is present in 70% of the patients here, while gout is present in 13%.

Evaluating unsupervised methods on supervised tasks is tricky. Our experiments here focused on the training set of each corpus rather than following the default train/test splits. Our primary concerns here were evaluating whether this method had any promise at all, and that it was applicable to more than one corpus. One could argue that future work should develop and tune the methods on the training data and then evaluate them on the test set. However, the very nature of this method breaks the traditional training/test model because tuning on the training data is already cheating relative to how the method would actually be applied on unlabeled data.

We are not sure that this problem has any perfect solutions, but we suggest that evaluating on as many different corpora as possible will be the best validation for this method. In this work, we tried to do that by starting on i2b2 data and then moving to the 20 Newsgroups data. Doing this helped us understand how informed priors need to be modified based on the size of the corpus.

One sticking point to portability with this method is the choice of classifier. We could have chosen a single classifier to stick with across corpora but then if one is particularly weak for a given corpus (e.g., SVM performed poorly on i2b2), it is less clear how much credit to assign the bronze labels for the performance. One possible solution to this issue is to require a much smaller sample of gold-labeled validation set if validated performance is strictly necessary.

One final point is that the classifier trained on bronze training labels probably would not generalize to a new corpus very well. This is not much of a problem, because the idea of the method is that for a new corpus one should generate new bronze labels using informed LDA on that data set. This does raise the question of what the difference would be between two classifiers trained on different corpora but with the same topic label, and whether there is some way of extracting additional information from comparing the decisions of these different classifiers on new data.

5 Conclusion

This work has shown that informed topic models seeded with topic information from Wikipedia can be used to train classifiers that perform much better than random. These classifiers are given no gold standard information and yet obtain results that may be useful in some applications. We show

that in active learning this method can improve learning rate for many categories. This method may be beneficial in domains where a large number of classifiers are required and state of the art performance is not necessary.

Acknowledgements

Research reported in this publication was supported by National Institute for General Medical Sciences (NIGMS) and National Library of Medicine (NLM) of the National Institutes of Health under award number R01GM114355 (HealthNLP) and U54LM008748 (i2b2). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Yoav Freund and Robert E Schapire. 1996. Experiments with a new boosting algorithm. In *ICML*, volume 96, pages 148–156.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213. Association for Computational Linguistics.

Thorsten Joachims. 1999. Making large scale svm learning practical. In B. Schlkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. Universität Dortmund.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584. ACM.

Andrew McCallum. 2002. Mallet: A machine learning for language toolkit.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359.

Dietrich Rebholz-Schuhmann, Antonio José Jimeno Yepes, Erik M Van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. 2010. CALBC silver standard corpus. *Journal of bioinformatics and computational biology*, 8(01):163–179.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Burr Settles. 2010. Active learning literature survey. Technical report, University of Wisconsin–Madison.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Hanna M Wallach, David M Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *NIPS*, volume 22, pages 1973–1981.