The Importance of Calibration for Estimating Proportions from Annotations

Dallas Card
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
dcard@cmu.edu

Noah A. Smith
Paul G. Allen School of CSE
University of Washington
Seattle, WA, 98195, USA
nasmith@cs.washington.edu

Abstract

Estimating label proportions in a target corpus is a type of measurement that is useful for answering certain types of social-scientific questions. While past work has described a number of relevant approaches, nearly all are based on an assumption which we argue is invalid for many problems, particularly when dealing with human annotations. In this paper, we identify and differentiate between two relevant data generating scenarios (intrinsic vs. extrinsic labels), introduce a simple but novel method which emphasizes the importance of calibration, and then analyze and experimentally validate the appropriateness of various methods for each of the two scenarios.

1 Introduction

A methodological tool often used in the social sciences and humanities (and practical settings like journalism) is *content analysis* – the manual categorization of pieces of text into a set of categories which have been developed to answer a substantive research question (Krippendorff, 2012). *Automated* content analysis holds great promise for augmenting the efforts of human annotators (O'Connor et al., 2011; Grimmer and Stewart, 2013). While this task bears similarity to text categorization problems such as sentiment analysis, the quantity of real interest is often the *proportion* of documents in a dataset that should receive each label (Hopkins and King, 2010). This paper tackles the problem of estimating label proportions in a target corpus based on a small sample of human annotated data.

As an example, consider the hypothetical question (not explored in this work) of whether hate speech is increasingly prevalent in social media posts in recent years. "Hate speech" is a difficult-to-define category only revealed (at least initially) through human judgments (Davidson et al., 2017). Note that the goal would not be to identify individual instances, but rather to estimate a proportion, as a way of measuring the prevalence of a social phenomenon. Although we assume that trained annotators could recognize this phenomenon with some acceptable level of agreement, relying solely on manual annotation would restrict the number of messages that could be considered, and would limit the analysis to the messages available at the time of annotation.[1]

We thus treat proportion estimation as a *measurement* problem, and seek a way to train an instrument from a limited number of human annotations to measure label proportions in an unannotated target corpus.

This problem can be cast within a supervised learning framework, and past work has demonstrated that it is possible to improve upon a naïve classification-based approach, even without access to any labeled data from the target corpus (Forman, 2005, 2008; Bella et al., 2010; Hopkins and King, 2010; Esuli and Sebastiani, 2015). However, as we argue (§2), most of this work is based on a set of assumptions that we believe are invalid in a significant portion of text-based research projects in the social sciences and humanities.

Our contributions in this paper include:

- identifying two different data-generating scenarios for text data (*intrinsic* vs. *extrinsic* labels) and and establishing their importance to the problem of estimating proportions (§2);
- analyzing which methods are suitable for each setting, and proposing a simple alternative approach for extrinsic labels (§3); and
- an empirical comparison of methods that validates our analysis (§4).

[1] For additional examples see Grimmer et al. (2012), Hopkins and King (2010), and references therein.

Complicating matters somewhat is the fact that annotation may take place before the entire collection is available, so that the subset of instances that are manually annotated may represent a biased sample (§2). Because this is so frequently the case, all of the results in this paper assume that we must confront the challenges of *transfer learning* or *domain adaptation*. (The simpler case, where we can sample from the true population of interest, is revisited in §5.)

2 Problem Definition

Our setup is similar to that faced in transfer learning, and we use similar terminology (Pan and Yang, 2010; Weiss et al., 2016). We assume that we have a source and a target corpus, comprised of N_S and N_T documents respectively, the latter of which are not available for annotation. We will represent each corpus as a set of documents, i.e., $\boldsymbol{X}^{(S)} = \langle \boldsymbol{x}_1^{(S)}, ..., \boldsymbol{x}_{N_s}^{(S)} \rangle$, and similarly for $\boldsymbol{X}^{(T)}$.

We further assume that we have a set of K mutually exclusive categories, $\mathcal{Y} = \{1, \ldots, K\}$, and that we wish to estimate the proportion of documents in the target corpus that belong to each category. These would typically correspond to a quantity we wish to measure, such as what fraction of news articles frame a policy issue in a particular way, what fraction of product reviews are considered helpful, or what fraction of social media messages convey positive sentiment. Generally speaking, these categories will be designed based on theoretical assumptions, an understanding of the design of the platform that produced the data, and/or initial exploration of the data itself.

In idealized text classification scenarios, it is conventional to assume training data with already-assigned gold-standard labels. Here, we are interested in scenarios where we must generate our labels via an annotation process.[2] Specifically, assume that we have some annotation function, $\mathcal{A}$, which produces a distribution over the K mutually exclusive labels, conditional on text. Given a document, $\boldsymbol{x}_i$, the annotation process samples a label from the annotation function, defined as:

$$\mathcal{A}(\boldsymbol{x}_i, k) \triangleq p(y_i = k \mid \boldsymbol{x}_i). \tag{1}$$

Typically, the annotation function would represent the behavior of a human annotator (or group of annotators), but it could also represent a less controlled real-world process, such as users rating a review's helpfulness. Note that our setup *does* include the special case in which true gold-standard labels are available for each instance (such as the authors of documents in an authorship attribution problem). In such a case, $\mathcal{A}$ is deterministic (assuming unique inputs).

Given that our objective is to mimic the annotation process, we seek to estimate the proportion of documents in the target corpus expected to be categorized into each of the K categories, if we had an unlimited budget and full access to the target corpus at the time of annotation. That is, we wish to estimate $q^{(T)}$, which we define as:

$$q(y = k \mid \boldsymbol{X}^{(T)}) \triangleq \tfrac{1}{N_T} \textstyle\sum_{i=1}^{N_T} p(y_i = k \mid \boldsymbol{x}_i^{(T)}). \tag{2}$$

Given a set of documents sampled from the *source* corpus and L applications of the annotation function, we can obtain, at some cost, a labeled training corpus of L documents, i.e., $D^{\text{(train)}} = \langle (\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_L, y_L) \rangle$. Because the source and target corpora are not in general drawn from the same distribution, we seek to make explicit our assumptions about how they differ.[3] Past literature on transfer learning has identified several patterns of dataset shift (Storkey, 2009). Here we focus on two particularly important cases, linking them to the relevant data generating processes, and analyze their relevance to estimating proportions.

Two kinds of distributional shift. There are two natural assumptions we could make about what is constant between the two corpora. We could assume that there is no change in the distribution of text given a document's label, that is $p^{(S)}(\boldsymbol{x} \mid y) = p^{(T)}(\boldsymbol{x} \mid y)$. Alternately, we could assume that there is no change in the distribution of labels given text, i.e., $p^{(S)}(y \mid \boldsymbol{x}) = p^{(T)}(y \mid \boldsymbol{x})$. The former is assumed in the case of *prior probability shift*, where we assume that $p(y)$ differs but $p(\boldsymbol{x} \mid y)$ is constant, and the later is assumed in the case of *covariate shift*, where we assume that $p(\boldsymbol{x})$ differs but $p(y \mid \boldsymbol{x})$ is constant (Storkey, 2009).

These two assumptions correspond to two fundamentally different types of scenarios that we need to consider, which are summarized in Table 1. The first is where we are dealing with what we

[2] This could include gathering multiple independent annotations per instance, but we will typically assume only one.

[3] Clearly, if we make no assumptions about how the source and target distributions are related, there is no guarantee that supervised learning will work (Ben-David et al., 2012).

Label type	Intrinsic	Extrinsic
Data generating process	$\boldsymbol{x} \sim p(\boldsymbol{x} \mid y)$	$y \sim p(y \mid \boldsymbol{x})$
Assumed to differ across domains	$p(y)$	$p(\boldsymbol{x})$
Assumed constant across domains	$p(\boldsymbol{x} \mid y)$	$p(y \mid \boldsymbol{x})$
Corresponding distributional shift	Prior probability shift	Covariate shift

Table 1: Data generating scenarios and corresponding distributional properties.

will call *intrinsic* labels, that is labels which are inherent to each instance, and which in some sense precede and predict the generation of the text of that instance. A classic example of this scenario is the case of authorship attribution (e.g., Mosteller and Wallace, 1964), in which different authors are assumed to have different propensities to use different styles and vocabularies. The identity of the author of a document is arguably an intrinsic property of that document, and it is easy to see a text as having been generated conditional on its author.

The contrasting scenario is what we will refer to as *extrinsic* labels; this scenario is our primary interest. We assume here that the labels are not inherent in the documents, but rather have been externally generated, conditional on the text as a stimulus to some behavioral process.[4] We argue that this is the relevant assumption for most annotation-based projects in the social sciences, where the categories of interest do not correspond to pre-existing categories that might have existed in the minds of authors before writing, or affected the writing process. Rather, these are theorized categories that have been developed specifically to analyze or measure some aspect of the document's *effect* that is of interest to the researcher.

We won't always know the true distributional properties of our datasets, but distinguishing between intrinsic and extrinsic labels provides a guide. The critical point is that these two different labeling scenarios have different implications for robustness to distributional shift. In the case of extrinsic labels, especially when working with trained annotators, it is reasonable to assume that the behavior of the annotation function is determined purely by the text, such that $p(y \mid \boldsymbol{x})$ is unchanged between source and target, and any change in label proportions is explained by a change in the underlying distribution of text, $p(\boldsymbol{x})$. With intrinsic labels, by contrast, it *may* be the case that $p(\boldsymbol{x} \mid y)$ is the same for the source and the target, assuming there are no additional factors influencing the generation of text. In that case, a shift in the distribution of features would be fully explained by a difference in the underlying label proportions.

The idea that there are different data generating processes is obviously not new.[5] What is novel here, however, is asking how these different assumptions affect the estimation of proportions. Virtually all past work on estimating proportions has only considered prior probability shift, assuming that $p(\boldsymbol{x} \mid y)$ is constant.[6] Existing methods take advantage of this assumption, and can be shown empirically to work well when it is satisfied (e.g., through artificial modification of real datasets to alter label proportions in a corpus). We expect them to fail, however, in the case of extrinsic annotations, as there is no reason to think that the required assumption should necessarily hold.

By contrast, the problem of covariate shift is in some sense less of a problem because we directly observe $\boldsymbol{X}^{(T)}$. Since the annotation function is assumed to be unchanging, we could perfectly predict the expected label proportions in the target corpus if we could learn the annotation function using labeled data from the source corpus. The problem thus becomes how to learn a well-calibrated approximation of the annotation function from a limited amount of labeled data.

3 Methods

Given a labeled training set and a target corpus, the naïve approach is to train a classifier through any conventional means, predict labels on the target corpus, and return the relative prevalence of predicted labels. Following Forman (2005), we refer to this approach as **classify and count** (CC). If using a probabilistic classifier, averaging the predicted posterior probabilities rather than predicted labels will be referred to as **probabilistic classify and count** (PCC; Bella et al., 2010).

Both approaches can fail, however. In the case of intrinsic labels, this is because these approaches will not account for the shift in prior label prob-

[4] Fong and Grimmer (2016) also consider this process in attempting to identify the causal effects of texts.

[5] Peters et al. (2014) describe these, somewhat confusingly, as *causal* and *anti-causal* problems.

[6] For example, Hopkins and King (2010) argue that bloggers first decide on the sentiment they wish to convey and then write a blog post conditional on that sentiment.

ability, $p(y)$, which is assumed to have occurred (Hopkins and King, 2010). In the case of covariate shift, the difference in $p(\boldsymbol{x})$ will result in a model that is not optimal (in terms of classification performance) for the target domain. In both cases, there is also the problem of classifier bias or miscalibration. Particularly in the case of unbalanced labels, a standard classifier is likely to be biased, overestimating the probability of the more common labels, and vice versa (Zhao et al., 2017). Here we present a simple but novel method for extrinsic labels, followed by a number of baseline approaches against which we will compare. (See supplementary material for additional details.)

3.1 Proposed method: calibrated probabilistic classify and count (PCC$^{\text{cal}}$)

One simple solution, which we propose here, is to attempt to train a well-calibrated classifier. To be clear, calibration refers to the long-run accuracy of predicted probabilities. That is, a probabilistic classifier, $h_\theta(\boldsymbol{x})$, is well calibrated at the level μ if, among all instances for which the classifier predicts class k with a probability of μ, the proportion that are truly assigned to class k is also equal to μ.[7]

It has previously been shown (DeGroot and Fienberg, 1983; Bröcker, 2009) that any proper scoring rule (e.g., cross entropy, Brier score, etc.) can be factored into two components representing *calibration* and *refinement*, the later of which effectively measures how close predicted probabilities are to zero or one. Minimizing a corresponding loss function thus involves a trade-off between these two components.

Optimizing *only* for calibration is not helpful, as a trivial solution is to simply predict a probability distribution equal to the observed label proportions in the training data for all instances (which is perfectly calibrated on the labeled sample). The alternative we propose here is to train a classifier using a typical objective (here, regularized log loss) but use *calibration on held-out data* as a criterion for *model selection*, i.e., when we tune hyperparameters via cross validation. We refer to this method as **calibrated PCC** (PCC$^{\text{cal}}$). Specifically, we select regularization strength via grid search, choosing the value that leads to the lowest average calibration error across training / held-out splits. Of course, other hyperparameters could be included in model selection as well.

To estimate calibration error (CE) during cross-validation, we use an approximation due to Nguyen and O'Connor (2015), adaptive binning. In the case of binary labels, this is computed as:

$$\text{CE} \triangleq \frac{1}{B}\sum_{j=1}^{B}\left(\frac{1}{|\mathcal{B}_j|}\sum_{i\in\mathcal{B}_j} y_i - p_\theta(\boldsymbol{x}_i)\right)^2, \quad (3)$$

using B bins, where bin $\mathcal{B}_j$ contains instances for which $p_\theta(\boldsymbol{x}_i)$ are in the j^{th} quantile, where $p_\theta(\boldsymbol{x}_i)$ is the predicted probability of a positive label for instance i. For added robustness, we take the average of CE for $B \in \{3, 4, 5, 6, 7\}$.

In our experiments, we consider two variants of PCC: the first, PCC$^{\text{F}_1}$, which is a baseline, is tuned conventionally for classification performance, whereas the other (PCC$^{\text{cal}}$) is tuned for calibration, as measured using CE, but is otherwise identically trained. As a base classifier we make use of l_1-regularized logistic regression, operating on n-gram features.[8]

3.2 Existing methods appropriate for extrinsic labels

The idea of extrinsic labels has not been previously considered by past work on estimating proportions, but it is closely related to the problems of calibration and covariate shift. Here we briefly summarize two representative methods, which we consider as baselines (see supplementary material for details).

Platt scaling. One approach to calibration is to train a model using conventional methods and to then learn a secondary calibration model. One of the most common and successful variations on this approach is Platt scaling, which learns a logistic regression classifier on held-out training data, taking the scores from the primary classifier as input. This model is then applied to the scores returned by the primary classifier on the target corpus (Platt, 1999). To estimate proportions, the predicted probabilities are then averaged, as in PCC.

Reweighting for covariate shift. Although they are not typically thought of in the context of estimating proportions, several methods have been proposed to deal directly with the problem of covariate shift, including kernel mean matching and

[7] For example, a weather forecaster will be well-calibrated if it rains on 60% of days for which the forecaster predicted a 60% chance of rain, etc.

[8] More complex models could be considered, but we use logistic regression because it is a well-understood and widely applicable model that has been shown to be relatively well-calibrated in general (Niculescu-Mizil and Caruana, 2005).

its extensions (Huang et al., 2006; Sugiyama et al., 2011). Here, we consider the two-stage method from Bickel et al. (2009), which uses a logistic regression model to distinguish between source and target domains, and then uses the probabilities from this model to re-weight labeled training instances, to more heavily favor those that are representative of the target domain. The appeal of this method is that all unlabeled data can be used to estimate this shift.

3.3 Existing methods appropriate for intrinsic labels

As previously mentioned, virtually all of the past work on estimating proportions makes the assumption that $p(\boldsymbol{x} \mid y)$ is constant between source and target. Under this assumption, it can be shown that $p(y^{(\theta)} = j \mid y = k)$ is also constant for all j and k, where $y^{(\theta)}$ is the predicted label from h_θ, and y is the true (intrinsic) label. If these values were known, then the label proportions in the target corpus could be found by taking the model's estimate of label proportions in the target corpus, (CC), and then solving a linear system of equations as a post-classification correction. Although a number of variations on this model have been proposed, all are based on the same assumption, thus we take a method known as **adjusted classify and count** (ACC) as an exemplar, which directly estimates the relevant quantities using a confusion matrix (Forman, 2005). In the case of binary classification, this reduces to:

$$\hat{q}_{\text{ACC}}(y = 1 \mid \boldsymbol{X}^{(T)}) = \frac{\frac{1}{N_T}\sum_{i=1}^{N_T} y_i^{(\theta)} - \text{FPR}}{\text{TPR} - \text{FPR}}, \quad (4)$$

where $\text{FPR} = \hat{p}(y^{(\theta)} = 1 \mid y = 0)$ and $\text{TPR} = \hat{p}(y^{(\theta)} = 1 \mid y = 1)$ are both estimated using held-out data.

4 Experiments

For our experiments, we focus on the case of binary classification where the difference between the source and target corpora results from a difference in time—that is, the training documents are sampled from one time period, and the goal is to estimate label proportions on documents from a future time period. We include examples of both intrinsic and extrinsic labels to demonstrate the importance of this distinction to the effectiveness of different methods.

As described below, we create multiple subtasks from each dataset by using different partitions of the data. In all cases, we report absolute error (AE) on the proportion of positive instances, averaged across the subtasks of each dataset.

Although we do not have access to the true annotation function, we approximate the expected label proportions in the target corpus by averaging the available labels, which should be a very close approximation when the number of available labels is large (which informed our choice of datasets for these experiments). For a single subtask, the absolute error is thus evaluated as

$$\text{AE} = \left| \hat{q}(y = 1 \mid \boldsymbol{X}^{(T)}) - \tfrac{1}{N_T}\textstyle\sum_{i=1}^{N_T} y_i^{(T)} \right|. \quad (5)$$

For all experiments, we also report the AE we would obtain from using the observed label proportions in the training sample as a prediction (labeled "Train"). Although this does not correspond to an interesting prediction (as it only says the future will always look exactly like the past), it does represent a fundamental baseline. If a method is unable to do better than this, it suggests that the method has too much measurement error to be useful.

To test for statistically significant differences between methods, we use an omnibus application of the Wilcoxon signed-rank test to compare one method against all others, including a Bonferroni correction for the total number of tests per hypothesis. With 4 datasets, each with 2 sample sizes, comparing against 6 other methods this results in a significance threshold of approximately 0.001.

Finally, in order to connect this work with past literature on estimating proportions, we also include a side experiment with one intrinsically-labeled dataset where we have artificially modified the label proportions in the target corpus by dropping positive or negatively-labeled instances in order to simulate a large prior probability shift between the source and target domains.

4.1 Datasets

We briefly describe the datasets we have used here and provide additional details in the supplementary material. Note that although this work is primarily focused on applications in which the amount of human-annotated data is likely to be small, fair evaluation of these methods requires datasets that are large enough that we can approximate the expected label proportion in the target corpus using the available labels; as such, the following datasets were chosen so as to have a representative sample of sufficiently large intrinsically and extrinsically-labeled data, where documents were time-stamped, with label proportions that differ between time periods.

Media Frames Corpus (MFC): As a primary example of extrinsic labels, we use a dataset of several thousand news articles that have been annotated in terms of a set of broad-coverage framing dimensions (such as economics, morality, etc.). We treat annotations as indicating the presence or absence of each dimension, and consider each one as a separate sub-task. As with all datasets, we create a source and target corpus by dividing the datasets by year. Particularly for this dataset, it seems reasonable to posit that the annotation function was relatively constant between source and target, as the annotators worked without explicit knowledge of the article's date (Card et al., 2015).

Amazon reviews: As a secondary example of extrinsic labels, we make use of a subset of Amazon reviews for five different product categories, each of which has tens of thousands of reviews. For this dataset, we ignore the star rating associated with the review, and instead focus on predicting the proportion of people that would rate the review as helpful. Here we create separate subtasks for each product category by considering each pair of adjacent years as a source and target corpus, respectively (McAuley et al., 2015).

Yelp reviews: As a primary example of a large dataset with intrinsic labels, we make use of the Yelp10 dataset, treating the source location of the review as the label of interest. Specifically, we create binary classification tasks by choosing pairs of cities with approximately the same number of reviews, and again use year of publication to divide the data into source and target corpora, creating multiple subtasks per pair of cities.

Twitter sentiment: Finally, we include a Twitter sentiment analysis dataset which was collected and automatically labeled, using the presence of certain emoticons as implicit labels indicating positive or negative sentiment (with the emoticons then removed from the text). Because of the way this data was collected, and the relatively narrow time coverage, it seems plausible to treat the sentiment as an intrinsic label. As with the above datasets, we create subtasks by considering all pairs of temporally adjacent days with sufficient tweets, and treating them as a paired source and target corpora, respectively. (Go et al., 2009).

4.2 Results

The results on the datasets with extrinsic and intrinsic labels are presented in Figures 1 and 2, respectively.

As expected, the results differ in important ways between intrinsically and extrinsically labeled datasets, although there are some results which hold in all cases. In all settings, CC is worse on average than predicting the observed proportions in the training data (significantly worse for the Amazon and Twitter datasets), reinforcing the idea that averaging the predictions from a classifier will lead to a biased estimate of label proportions. This same finding holds for PCC$^{\text{F}_1}$ when the amount of labeled data is small ($L = 500$), suggesting that simply averaging the predicted probabilities is not reliable without a sufficiently large labeled dataset.

For the datasets with *extrinsic* labels, PCC$^{\text{cal}}$ performs best on average in all settings. For the MFC dataset, PCC$^{\text{cal}}$ is significantly better than all methods except Platt scaling when $L = 500$ and significantly better than all methods except reweighting and PCC$^{\text{F}_1}$ when $L = 2000$ (after a Bonferroni correction, as in all cases). As expected, ACC is actually worse on average than CC on the extrinsic datasets, presumably because of the mismatched assumptions. Reweighting for covariate shift offers mediocre performance in all settings, perhaps because, while it attempts to account for covariate shift, it may still suffer from miscalibration.

On the datasets with *intrinsic* labels, by contrast, no one method dominates the others. As expected, ACC does poorly when the amount of labeled data is small ($L = 500$); it does improve upon CC when $L = 4000$, but not by enough to do significantly better than other methods, perhaps calling into question the validity of the assumption that $p(\boldsymbol{x} \mid y)$ is constant in these datasets.

Surprisingly, both Platt scaling and PCC$^{\text{cal}}$ also offer competitive performance in the experiments with intrinsic labels. However, this is likely the case in part because the change in label proportions is relatively small from year to year (or day

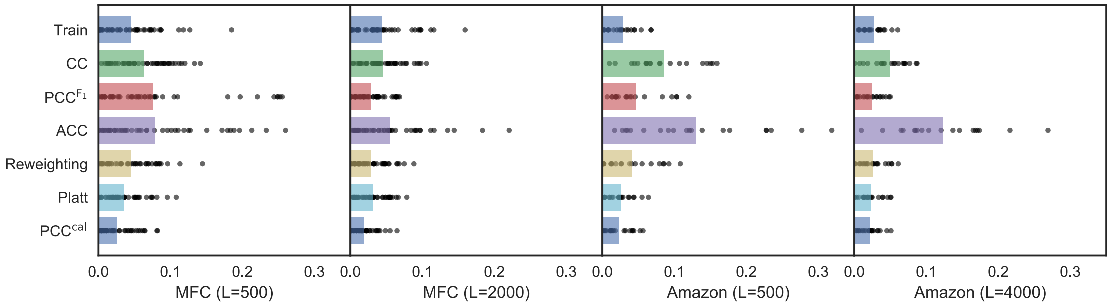

Figure 1: Absolute error (AE) on datasets with extrinsic labels. Each dot represents the result for a single subtask, and bars show the mean. PCC^{cal} (bottom row) performs best on average in all cases and is significantly better than most other methods on MFC.

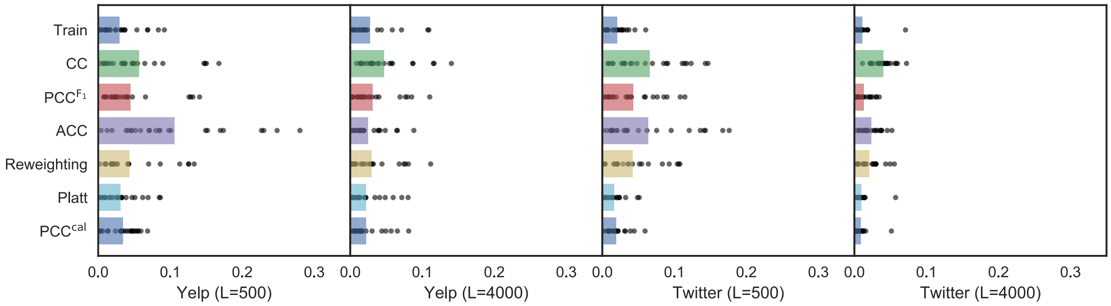

Figure 2: Absolute error (AE) on datasets with intrinsic labels. No method is significantly better than all others.

to day in the case of Twitter). This is illustrated by Figure 3, which presents the results of the side-experiment with artificially modified (intrinsic) label proportions using a subset of the Twitter data. These results confirm past findings, and show that ACC drastically outperforms other methods such as PCC^{F_1}, *if* we selectively drop instances so as to enforce a large difference in label proportions between source and target. This is the expected result, as ACC is the only method tailored to deal with prior probability shift (which is being artificially simulated). Unfortunately, its advantage is not maintained when the difference between source and target is small, which is the case for all of the naturally-occurring differences we found in the Yelp and Twitter datasets. Although past work has relied heavily on these sorts of simulated differences and artificial experiments, it is unclear whether they are a good substitute for real-world data, given that we mostly observed relatively small differences in practice.

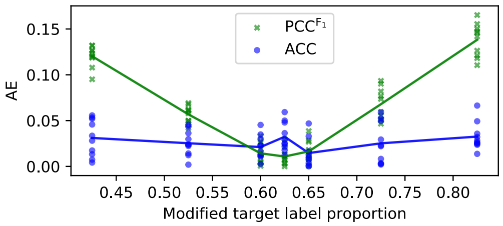

Figure 3: Absolute error (AE) for predictions on one day of Twitter data ($L = 5000$) when artificially modifying target proportions. The proportion of positive labels in the source corpus is 0.625. ACC performs significantly better given an large artificially-created difference in label proportions between source and target, but not when the difference is small.

Finally, we also tested the effect of using l_2 instead of l_1 regularization, but found that it tended to produce significantly worse estimates of proportions using CC and PCC^{F_1} on the datasets with extrinsic labels, and statistically indistinguishable results using other methods, suggesting that either type of regularization could serve as a basis for PCC^{cal} or Platt scaling.

5 Discussion

As anyone who has worked with human annotations can attest, the process of collecting annotations is messy and time-consuming, and tends to involve large numbers of disagreements (Artstein and Poesio, 2008). Although it is conventional to treat disagreements as *errors* on the behalf of some subset of annotators, this paper provides an alternative way of understanding these. By treating annotation as a stochastic process, conditional on text, we can explain not only the disagreements between annotators, but also the lack of self-consistency that is also sometimes observed. Although the assumption that $p(y \mid \boldsymbol{x})$ does not change is clearly a simplification, it seems reasonable when working with trained annotators. Certainly this assumption seems much better justified than the conventional assumption that $p(\boldsymbol{x} \mid y)$ is constant, since the latter does not account for differences in the distribution of text arising from differences in subject matter, etc.

Although we have demonstrated that using a method that is appropriate to the data generating process is beneficial, it is important to note that all methods presented here can still result in relatively large errors in the worst cases. In part this is due to the difficulty of learning a conditional distribution involving high-dimensional data (such as text) with only a limited number of annotations. Even with much more annotated data, however, previously unseen features could still have a potentially large impact on future annotations. Ultimately, we should be cautious about all such predictions, and always validate where possible, by eventually sampling and annotating data from the target corpus.

What if we can sample from the target corpus? Although there are many situations in which domain adaptation is unavoidable (such as predicting public opinion from Twitter in real time with models trained on the past), at least some research projects in the humanities and social sciences might reasonably have access to all data of interest from the beginning of the project, such as when working with a historical corpus. Although a full proof is beyond the scope of this paper, in this case, the best approach is almost certainly to simply sample a random set of documents, label them using the annotation function, and report the relative prevalence of each label (Hopkins and King, 2010).

Although this *simple random sampling* (SRS) approach ignores the text, it is an unbiased estimator with variance that can easily calculated, at least in approximation.[9] More importantly, because it is independent of the dimensionality of the data, it works well on high-dimensional data, such as text, whereas classification-based approaches will struggle. We can illustrate this by comparing SRS and PCC in simulation. Figure 4 shows the mean AE (averaged over 200 trials) for a case in which we know the true model (including the prior on the weights, and thus the appropriate amount of regularization) and only need to learn the values of the weights. Even in this idealized scenario, SRS remains better than PCC for all values of L. (See supplementary material for details).

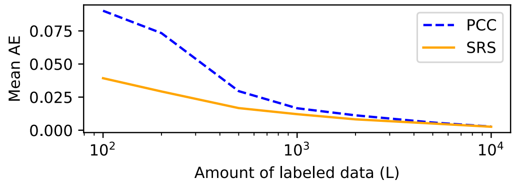

Figure 4: Comparison of SRS and PCC in simulation when we know the true model and sample from the target corpus (averaged over 200 repetitions).

Depending on the level of accuracy required, simply sampling a few hundred documents and labeling them should be sufficient to get a reasonably reliable estimate of the overall label proportions, along with an approximate confidence interval. Unfortunately, this option is only available when we have full access to the target corpus at the time of annotation.

Additional related work. There is a small literature on the problem of estimating proportions in a target dataset (see §1); as we have emphasized, almost all of it makes the assumption that $p(\boldsymbol{x} \mid y)$ is the same for both source and target. Moreover, most of the methods that have been proposed have been tested using relatively small datasets, or datasets where the target corpus has been artificially modified by altering the label proportions in the target corpus (as we did in the side experiment reported in Figure 3). It

[9] If we were sampling with replacement, the variance in the binary case would be given by the standard formula $\mathbb{V}[\hat{q}^{\text{SRS}}] = \frac{\bar{p}(1-\bar{p})}{L}$, where $\bar{p} = \frac{1}{N_T}\sum_{i=1}^{N_T} p(y_i = 1 \mid \boldsymbol{x}_i)$. This may not be possible, however, as annotators seeing a document for the second or third time would likely be affected by their own past decisions. Nevertheless, using this as the basis for a plug-in estimator should still be a reasonable approximation when the target corpus is large. Please refer to supplementary material for additional details.

seems unclear that this is a good simulation of the kind of shift in distribution that one is likely to encounter in practice. An exception to this is Esuli and Sebastiani (2015), who test their method on the RCV1-v2 corpus, also splitting by time. They perform a large number of experiments, but unfortunately, nearly all of their experiments involve only a very small difference in label proportions between the source and target (with the vast majority < 0.01), which limits the generalizability of their findings. Additional methods for calibration could also be considered, such as the isotonic regression approach of Zadrozny and Elkan (2002), but in practice we would expect the results to be very similar to Platt scaling.

Another line of work has approached the problem of aggregating labels from multiple annotators (Raykar et al., 2009; Hovy et al., 2013; Yan et al., 2013). That is, if we believe that some annotators are more reliable than others, it might make sense to try to determine this in an unsupervised manner, and give more weight to the annotations from the reliable annotators. This seems particularly appropriate when dealing with uncooperative annotators, as might be encountered, for example, in crowdsourcing (Snow et al., 2008; Zhang et al., 2016). However, with a team of trained annotators, we believe that honest disagreements could contain valuable information better not ignored.

Finally, this work also relates to the problem of *active learning*, where the goal is to interactively choose instances to be labeled, in a way that maximizes accuracy while minimizing the total cost of annotation (Beygelzimer et al., 2009; Baldridge and Osborne, 2004; Rai et al., 2010; Settles, 2012). This is an interesting area that might be productively combined with the ideas in this paper. In general, however, the use of active learning involves additional logistical complications and does not always work better than random sampling in practice (Attenberg and Provost, 2011).

6 Conclusions

When estimating proportions in a target corpus, it is important to take seriously the data generating process. We have argued that in the case of data annotated by humans in terms of categories designed to help answer social-scientific research questions, labels should be treated as *extrinsic*, generated probabilistically conditional on text, rather than as a combination of correct and incorrect judgements about a label *intrinsic* to the document. Moreover, it is reasonable to assume in this case that $p(y \mid \boldsymbol{x})$ is unchanging between source and target, and methods that aim to learn a well-calibrated classifier, such as PCC$^{\text{cal}}$, are likely to perform best. By contrast, if $p(\boldsymbol{x} \mid y)$ is unchanging between source and target, then various correction methods from the literature on estimating proportions, such as ACC, can perform well, especially when differences are large. Ultimately, any of these methods can still result in large errors in the worst cases. As such, validation remains important when treating the estimation of proportions as a type of measurement.

Acknowledgements We would like to thank Philip Resnik, Brendan O'Connor, anonymous reviewers, and all members of Noah's ARK for helpful comments and discussion, as well as XSEDE and Microsoft Azure for grants of computational resources used for this work.

References

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4):555–596. https://doi.org/10.1162/coli.07-034-R2.

Josh Attenberg and Foster Provost. 2011. Inactive learning?: Difficulties employing active learning in practice. *SIGKDD Explorations Newsletter* 12(2):36–41. https://doi.org/10.1145/1964897.1964906.

Jason Baldridge and Miles Osborne. 2004. Active learning and the total cost of annotation. In *Proceedings of EMNLP*.

Antonio Bella, Maria Jose Ramirez-Quintana, Jose Hernandez-Orallo, and Cesar Ferri. 2010. Quantification via probability estimators. In *IEEE International Conference on Data Mining*. https://doi.org/10.1109/ICDM.2010.75.

Shai Ben-David, Shai Shalev-Shwartz, and Ruth Urner. 2012. Domain adaptation – can quantity compensate for quality? *Annals of Mathematics and Artificial Intelligence* 70:185–202. https://doi.org/10.1007/s10472-013-9371-9.

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. 2009. Importance weighted active learning. In *Proceedings of ICML*. https://doi.org/10.1145/1553374.1553381.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. 2009. Discriminative Learning Under Covariate Shift. *Journal of Machine Learning Research* 10.

Jochen Bröcker. 2009. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society* 135(643):1512–1519.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of ACL*. https://doi.org/10.3115/v1/P15-2072.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM*.

Morris H. DeGroot and Stephen E. Fienberg. 1983. The comparison and evaluation of forecasters. *The Statistician: Journal of the Institute of Statisticians* 32:12–22.

Andrea Esuli and Fabrizio Sebastiani. 2015. Optimizing text quantifiers for multivariate loss functions. *ACM Trans. Knowl. Discov. Data* 9(4). https://doi.org/10.1145/2700406.

Christian Fong and Justin Grimmer. 2016. Discovery of treatments from text corpora. In *Proceedings of ACL*. https://doi.org/10.18653/v1/P16-1151.

George Forman. 2005. Counting positives accurately despite inaccurate classification. In *Proceedings of the European Conference on Machine Learning*.

George Forman. 2008. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery* 17(2):164–206. https://doi.org/10.1007/s10618-008-0097-y.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report.

Justin Grimmer, Solomon Messing, and Sean J. Westwood. 2012. How words and money cultivate a personal vote: The effect of legislator credit claiming on constituent credit allocation. *American Political Science Review* 106(4):703–719. https://doi.org/10.1017/S0003055412000457.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3):267–297. https://doi.org/10.1093/pan/mps028.

Daniel Hopkins and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54(1):220–247. https://doi.org/10.1111/j.1540-5907.2009.00428.x.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H. Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of NAACL*.

Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. 2006. Correcting sample selection bias by unlabeled data. In *Proceedings of NIPS*.

Klaus Krippendorff. 2012. *Content analysis: an introduction to its methodology*. SAGE.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of SIGIR*. https://doi.org/10.1145/2766462.2767755.

Frederick Mosteller and David L. Wallace. 1964. *Inference and Disputed Authorship*. Addison-Wesley publishing company, Inc. https://doi.org/10.1080/01621459.1963.10500849.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of EMNLP*.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of ICML*. https://doi.org/10.1145/1102351.1102430.

Brendan O'Connor, David Bamman, and Noah A. Smith. 2011. Computational text analysis for social science: Model assumptions and complexity. In *NIPS Workshop on Comptuational Social Science and the Wisdom of Crowds*.

S.J. Pan and Q. Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22(10):1345–1359. https://doi.org/10.1109/TKDE.2009.191.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. 2014. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research* 15:2009–2053.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74.

Piyush Rai, Avishek Saha, III Hal Daumé, and Suresh Venkatasubramanian. 2010. Domain adaptation meets active learning. In *Proceedings of NAACL*.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Anna K. Jerebko, Charles Florin, Gerardo Hermosillo, Luca Bogoni, and Linda Moy. 2009. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Proceedings of ICML*. https://doi.org/10.1145/1553374.1553488.

Burr Settles. 2012. *Active Learning*. Morgan & Claypool. https://doi.org/10.2200/S00429ED1V01Y201207AIM018.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*.

Amos J. Storkey. 2009. When training and test sets are different: Characterising learning transfer. In Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Candela Sugiyama Schwaighofer Lawrence Lawrence, editors, *Dataset Shift in Machine Learning*, MIT Press, chapter 1, pages 3–28.

Masashi Sugiyama, Makoto Yamada, Paul von Bünau, Taiji Suzuki, Takafumi Kanamori, and Motoaki Kawanabe. 2011. Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. *Neural Networks* 24(2). https://doi.org/10.1016/j.neunet.2010.10.005.

Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big Data* 3(1). https://doi.org/10.1186/s40537-016-0043-6.

Yan Yan, Rómer Rosales, Glenn Fung, Subramanian Ramanathan, and Jennifer G. Dy. 2013. Learning from multiple annotators with varying expertise. *Machine Learning* 95:291–327. https://doi.org/10.1007/s10994-013-5412-1.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of KDD*. https://doi.org/10.1145/775047.775151.

Jing Zhang, Xindong Wu, and Victor S. Sheng. 2016. Learning from crowdsourced labeled data: a survey. *Artif. Intell. Rev.* 46:543–576. https://doi.org/10.1007/s10462-016-9491-9.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the EMNLP*.

The Importance of Calibration for Estimating Proportions from Annotations: Supplementary Material

Dallas Card
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
dcard@cmu.edu

Noah A. Smith
Paul G. Allen School of CSE
University of Washington
Seattle, WA, 98195, USA
nasmith@cs.washington.edu

1 Experimental Details

For tuning the base classifier, we used grid search to choose the strength of regularization strength, testing 11 values from 0.01 to 1000. On each experiment, the training set was split into five random folds. A classifier was trained for each four-fifths of the data, using the remaining fifth as a validation set in each case. The validation set was used to choose regularization strength (using F_1 or calibration as a performance metric), as well as to estimate secondary models, such as ACC or Platt scaling. The predicted proportions from each of the five models (one for each development fold) were then averaged to produce the final estimate of proportions. Reweighting, Platt scaling, CC, and ACC were all based on the model trained using F_1. For Platt scaling, we do not regularize the secondary model, but instead replace the binary labels with smoothed target values, as suggested in the original paper (Platt, 1999). Because ACC can result in inadmissible values in extreme cases, we threshold its predictions to be in the range $[0, 1]$.

2 Datasets

Media Frames Corpus. For this dataset, we treat each framing dimension for each of three issues (immigration, same-sex marriage, and smoking) as a separate subtask. Because there are fewer labeled instances in this dataset than the others, we only create a single split into a source and target corpus for each subtask, treating the articles published before 2009 as a source corpus, and testing on articles from 2009–2012. Most documents in this dataset were annotated by two annotators, so we weight these inversely proportional to the number of annotators for each instance.

Amazon reviews. For this dataset, we made use of the 5-core subsets for five mid-sized product categories: 1) clothing, shoes and jewelry; 2) home and kitchen; 3) sports and outdoors; 4) toys and games; and 5) tools and home improvement, and treat the proportion of people rating the review as "helpful" as the target. For each category, we create separate subtasks by treating each pair of adjacent years in the range 2010–2014 as a source and target corpus (using the earlier year as the source and the later as the target). As with the MFC, we weight instances with multiple votes inversely proportional to the total number of votes per instance.

Yelp reviews. For this dataset, we used three pairs of cities with approximately the same numbers of reviews: Toronto and Scottsdale; Charlotte and Pittsburgh; and Tempe and Henderson. For each pair, we created multiple subtasks by treating each pair of adjacent years as a source and target corpus, respectively, for the years 2009–2017. We ignore the star rating, the title of the review and information about the author, and only consider the review text and location (as a label).

Twitter sentiment. For this dataset, we only make use of what is designated as the official training set (which is the vast majority of instances). Similar to the other datasets, we create subtasks by creating a source and target corpus from each pair of adjacent days for which both days had at least 4,000 tweets. Note that the tweets from after day 166 appear to be artificially biased (containing only positive or negative tweets), thus we exclude these from the analysis.

3 Simulation Details

To simulate a comparison of PCC and SRS when we are able to randomly sample instances to be labeled from the target corpus, we generate sparse binary data and sparse weights and then fit a model with the same form and hyperparameters to a subset of the data. Specifically, we use the following data generating process, for $i = 1, \ldots, N$ and $j = 1, \ldots, P$:

$$\begin{aligned} X_{ij} &\sim \text{Bernoulli}(p_x) \\ \beta_j &\sim \text{Laplace}(0, 1) \\ \beta_0 &\sim \mathcal{N}(0, 1) \\ p_i &= \text{Sigmoid}(X_{i,:} \cdot \beta + \beta_0) \\ y_i &\sim \text{Bernoulli}(p_i) \end{aligned}$$

We then fit this model to a subset of the data using an l_1-regularized logistic regression model with regularization strength equal to 1, and average the predicted probabilities over all instances (PCC), or simply average the observed labels in the subset (SRS). Figure 3 in the paper was made using values of $N = 20000$, $P = 10000$, and $p_x = 0.01$, averaged over 200 repetitions, varying the amount of labeled data available to the models.

4 Variance of Simple Random Sampling

As noted in the paper, if we were able to sample and annotate data from the target corpus *with replacement*, the variance of SRS for binary labels would be $\frac{\bar{p}(1-\bar{p})}{L}$, where $\bar{p} = \frac{1}{N_T}\sum_{i=1}^{N_T} p_i$, and $p_i = p(y_i = 1 \mid \boldsymbol{x}_i)$. In the case where we sample a random set of instances from the target corpus and annotate each one exactly once, the variance of the resulting estimate is somewhat more complicated, as there are two sources of randomness – the set of instances selected for annotation (A) and the labels returned by the annotation function (Y). Using the law of total variance, we have

$$\begin{aligned} &\mathbb{V}_{A,Y}[\hat{q}^{\text{SRS}}] \\ &\quad = \mathbb{E}_A[\mathbb{V}_Y[\hat{q}^{\text{SRS}} \mid A]] + \mathbb{V}_A[\mathbb{E}_Y[\hat{q}^{\text{SRS}} \mid A]]. \end{aligned} \tag{1}$$

Note that the first component in Equation (1) will be zero if $p_i = 0$ or $p_i = 1, \forall i$, and is maximized if $p_i = 0.5, \forall i$. Conversely, the second component is equal to zero if all p_i have the same value, and is maximized if the p_is are evenly split between $p_i = 0$ and $p_i = 1$. As such, there is a tradeoff between these two components.

We can further simplify the above terms as follows. First,

$$\begin{aligned} &\mathbb{E}_A[\mathbb{V}_Y[\hat{q}^{\text{SRS}} \mid A]] \\ &\quad = \mathbb{E}_A\left[\mathbb{V}_Y\left[\frac{1}{L}\sum_{i\in A} y_i \,\middle|\, A\right]\right] && (2) \\ &\quad = \mathbb{E}_A\left[\frac{1}{L^2}\sum_{i\in A} p_i(1-p_i)\right] && (3) \\ &\quad = \frac{1}{L}\frac{1}{N_T}\sum_{i=1}^{N_T} p_i(1-p_i) && (4) \\ &\quad = \frac{1}{L}\left(\bar{p} - \frac{1}{N_T}\sum_{i=1}^{N_T} p_i^2\right) && (5) \\ &\quad = \frac{1}{L}\left(\bar{p} - (S^2 + \bar{p}^2)\right) && (6) \\ &\quad = \frac{1}{L}\left(\bar{p}(1-\bar{p}) - S^2\right), && (7) \end{aligned}$$

where S^2 is the sample variance of the set of p_is in the target corpus. Similarly,

$$\begin{aligned} \mathbb{V}_A[\mathbb{E}_Y[\hat{q}^{\text{SRS}} \mid A]] &= \mathbb{V}_A\left[\mathbb{E}_Y\left[\frac{1}{L}\sum_{i\in A} y_i \,\middle|\, A\right]\right] && (8) \\ &= \mathbb{V}_A\left[\frac{1}{L}\sum_{i\in A} p_i\right]. && (9) \end{aligned}$$

For a sufficiently large L and N_T, we can approximate this with the central limit theorem for a finite population (Bellhouse, 2001), which gives us

$$\mathbb{V}_A[\bar{p}_A] \approx S^2\left(\frac{1}{L} - \frac{1}{N_T}\right). \tag{10}$$

When we only have access to a single label per instance, it is not possible to estimate S^2, but we can nevertheless combine the two parts above and use a standard plug-in estimator to approximate an upper bound on the variance of simple random sampling, $\hat{\mathbb{V}}[\hat{q}^{\text{SRS}}] \approx \frac{\bar{y}(1-\bar{y})}{L}$, where $\bar{y} = \frac{1}{L}\sum_{i\in A} y_i$, and empirically this produces a reasonable, if somewhat pessimistic estimate.

References

D. R. Bellhouse. 2001. The central limit theorem under simple random sampling. *The American Statistician* 55(4):352–357. https://doi.org/10.1198/000313001753272330.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74.