

Ballpark Learning: Estimating Labels from Rough Group Comparisons

Tom Hope^(✉) and Dafna Shahaf

The Hebrew University of Jerusalem, Jerusalem, Israel
tom.hope@mail.huji.ac.il, dshahaf@cs.huji.ac.il

Abstract. We are interested in estimating individual labels given only coarse, aggregated signal over the data points. In our setting, we receive sets (“bags”) of unlabeled instances with constraints on label proportions. We relax the unrealistic assumption of known label proportions, made in previous work; instead, we assume only to have upper and lower bounds, and constraints on bag differences. We motivate the problem, propose an intuitive formulation and algorithm, and apply our methods to real-world scenarios. Across several domains, we show how using only proportion constraints and no labeled examples, we can achieve surprisingly high accuracy. In particular, we demonstrate how to predict income level using rough stereotypes and how to perform sentiment analysis using very little information. We also apply our method to guide exploratory analysis, recovering geographical differences in twitter dialect.

1 Introduction

In many classification problems, labeled instances are often difficult, expensive, or time-consuming to obtain. Unlabeled instances, on the other hand, are easier to obtain, but it is harder to use them for classification. Semi-supervised learning [6] addresses this problem, using unlabeled instances together with a small amount of labeled instances to improve performance.

We are interested in a learning setting where few, if any, labeled instances exist. Instead, we only know some coarse, aggregated signal over the data points. In particular, our instances are divided into sets (or *bags*), and we are given some aggregate information about the bags; for example, we might know that one bag has a higher percentage of positive-label instances than another.

There is recent interest in the task of estimating the labels of *individual* instances given aggregate information, due to the many real-world scenarios in which such information is available. In particular, aggregate information (e.g., summary statistics) is often published for sensitive data, when one cannot publish individual statistics. Being able to estimate individual labels from such data has important implications regarding privacy and data anonymization.

Constraining class proportions of unlabeled data has been shown to be useful for semi-supervised learning [20, 25, 27]. Under this setting, we are given sets of unlabeled instances with known label proportions (for example, one bag has 30% positive instances and 70% negative instances).

We believe that the assumption of known proportions is unrealistic, and limits the applicability of such methods. For example, suppose we want to classify Twitter users by political orientation. We have some information about the users (for example, the text of their tweets), but no explicit political affiliation to use as labels. We could, however, use the commonly-known fact that political orientation is correlated with geographic location. Thus, we can construct bags of users based on their geographic location: bags would correspond to states whose residents predominantly vote for the Republican Party (red states) or Democratic Party (blue states).

Estimating the proportion of Democrats on Twitter is hard, even using location information. Previous election data or polls are unlikely to accurately reflect the behavior of Twitter users. Instead of assuming known proportions, we propose a setup where our input is much weaker: we only know some constraints on bag proportions and on differences between bags. In other words, users from red states do not necessarily vote for the Republicans, but it is safe to expect to see more Republicans in the red-state bags. It is also reasonable to assume that, say, at least 10% of Blue-state users are Democrats. Using only this type of weak, “ballpark” estimates, we would like to be able to classify individual users.

Figure 1 demonstrates this idea. Our input includes approximate information on label proportions in some bags (left) and pairwise comparisons between bags (middle) or sets of bags (right). Our contributions are as follows:

- We extend the Learning from Labeled Proportions setting by proposing a new, more realistic scenario in which label proportions in each bag are not assumed to be known, but rather some constraints on them. We suggest various domains that lend themselves to this setting.
- We propose a simple and intuitive bi-convex problem formulation and an efficient algorithm, including a novel form of cross-validation.
- We apply our algorithm to real data, perform sentiment analysis of movie reviews from a very coarse signal, and predict income using stereotypes.
- We demonstrate the use of our method for exploratory analysis. We find vernacular difference in geo-tagged tweets by incorporating expressive constraints such as “Alabama > Florida > New York”.
- Our algorithm is designed to use when human labeling resources are scarce. Despite the simplicity of our methods, we achieve high accuracy with a very modest amount of input, and considerably loose (or misspecified) constraints.



Fig. 1. We are given bags of instances and rough estimates about label proportions and differences between bags. Here, the purple bag has at least 50% positive instances, more than the red bag (but the magnitude of the difference is uncertain). (Color figure online)

2 Problem Formulation

We begin by formalizing our setting and problem. Consider a set of N training instances $\mathcal{X}_N = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$. Each \mathbf{x}_i has a corresponding *unknown* label $y_i^* \in \{-1, 1\}$. In addition, we could be given a (possibly empty) set of L labeled training instances $\mathcal{X}_L = \{\mathbf{x}_{N+1}, \mathbf{x}_{N+2}, \dots, \mathbf{x}_{N+L}\}$ with known binary labels y_i , where typically the vast majority of our instances are unlabeled: $N \gg L$. In addition, we are given a set of K subsets of \mathcal{X} , which we call *bags*:

$$\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_K\}, \mathcal{B}_k \subseteq \mathcal{X}_N \cup \mathcal{X}_L.$$

Note that bags \mathcal{B} may overlap, and do not have to cover all training instances \mathcal{X}_N . Let p_k be the proportion of positive-labeled instances in bag \mathcal{B}_k :

$$p_k = |\{i : i \in \mathcal{B}_k, y_i^* = 1\}| / |\mathcal{B}_k| \quad (1)$$

(where y_i^* is replaced with y_i for instances $\mathbf{x}_i \in \mathcal{X}_L$). Previous work [20] tackled the case of *known* label proportions, suggesting that precise proportions could be estimated using sampling. However, obtaining accurate estimates could be costly or impractical (e.g., for bags with high label skew). In this work we do not assume to know p_k . Rather, we are given weaker prior knowledge, in the form of constraints on proportions. We allow constraints of the following forms:

- **Lower and upper bounds** on bag proportions: $l_k \leq p_k \leq u_k$
- **Bag difference bounds**: $0 \leq l_{k_{12}} \leq p_{k_1} - p_{k_2} \leq u_{k_{12}}$

We are especially interested in the case where very little information is known: constraints are *loose*, and specified only for a small subset of the bags.

Our goal is to predict a label for each \mathbf{x}_i , using a function $f(\mathbf{x}) = \text{sign}(\mathbf{w}^T \varphi(\mathbf{x}))$, where \mathbf{w} is a weight vector and $\varphi(\cdot)$ is a feature map (to simplify notation we drop a bias term \mathbf{b} by assuming a vector $\mathbf{1}_{N+L}$ is appended to the features). To attain the classification goal, we use a maximum-margin approach. Let \mathcal{R} be the subset of \mathcal{B} for which we have upper and/or lower bounds. Let \mathcal{D} be the set of tuples $(\mathcal{B}_{k_1}, \mathcal{B}_{k_2})$ for which we have difference bounds. To solve this problem we directly model the latent variable \mathbf{y}^* – the vector of unknown labels $y_i^* \in \{-1, 1\}$, in an alternating optimization approach.

Noting that (1) can be written as $p_k = \frac{\sum_{i \in \mathcal{B}_k} y_i^*}{2|\mathcal{B}_k|} + \frac{1}{2}$, we formulate the following bi-convex optimization problem:

$$\begin{aligned} \underset{\mathbf{y}, \mathbf{w}, \xi}{\text{argmin}} & \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{C}{N} \sum_{i=1}^N \max(0, 1 - y_i \mathbf{w}^T \varphi(\mathbf{x}_i)) + \frac{C_L}{L} \sum_{j=N+1}^{N+L} \xi_j \\ \text{s.t.} & -1 \leq y_i \leq 1 \quad \forall i \in 1, \dots, N \\ & y_j \mathbf{w}^T \varphi(\mathbf{x}_j) \geq 1 - \xi_j \quad \forall j \in \{N+1, \dots, N+L\} \\ & \xi_j \geq 0 \quad \forall j \\ & l_k \leq \hat{p}_k \leq u_k \quad \forall \{k : \mathcal{B}_k \in \mathcal{R}\} \\ & l_{k_{12}} \leq \hat{p}_{k_1} - \hat{p}_{k_2} \leq u_{k_{12}} \quad \forall \{k_1 \neq k_2 : (\mathcal{B}_{k_1}, \mathcal{B}_{k_2}) \in \mathcal{D}\}, \end{aligned} \quad (2)$$

where $\hat{p}_k = \frac{1}{2|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} y_i + \frac{1}{2}$ is the estimated positive label proportion in bag \mathcal{B}_k , l_k (or u_k) can be 0 (1) if not given as input, and analogously for difference bounds $l_{k_{12}}$ ($u_{k_{12}}$). C and C_L are cost hyperparameters for unlabeled and labeled instances, respectively. Intuitively, the second term in the objective function helps find a weight vector \mathbf{w} accurately predicting \mathbf{y} , and constraints ensure that we find an assignment to \mathbf{y} that satisfies proportions constraints. C_L controls how much weight we give to our labeled instances versus our prior knowledge on \mathcal{B} . In our experiments we do not use any labeled instances, thus $C_L = 0$.

3 Algorithm

We have formalized our problem as a bi-convex optimization problem – holding either \mathbf{w} or \mathbf{y} fixed, we get a convex problem. We thus propose the following intuitive alternating algorithm to solve it.

– For a fixed \mathbf{w} , solve for \mathbf{y} :

$$\begin{aligned} \operatorname{argmin}_{\mathbf{y}} \frac{1}{N} \sum_{i=1}^N \max(0, 1 - y_i \mathbf{w}^T \varphi(\mathbf{x}_i)) \\ \text{s.t. } -1 \leq y_i \leq 1 \quad \forall i \in 1, \dots, N \\ l_k \leq \hat{p}_k \leq u_k \quad \forall \{k : \mathcal{B}_k \in \mathcal{R}\} \\ l_{k_{12}} \leq \hat{p}_{k_1} - \hat{p}_{k_2} \leq u_{k_{12}} \quad \forall \{k_1 \neq k_2 : (\mathcal{B}_{k_1}, \mathcal{B}_{k_2}) \in \mathcal{D}\}, \end{aligned} \quad (3)$$

– For a fixed \mathbf{y} , solve w.r.t \mathbf{w} :

$$\begin{aligned} \operatorname{argmin}_{\mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{C}{N} \sum_{i=1}^N \max(0, 1 - y_i \mathbf{w}^T \varphi(\mathbf{x}_i)) + \frac{C_L}{L} \sum_{j=N+1}^{N+L} \xi_j \\ \text{s.t. } y_j \mathbf{w}^T \varphi(\mathbf{x}_j) \geq 1 - \xi_j \quad \forall j \in \{N+1, \dots, N+L\} \\ \xi_j \geq 0 \quad \forall j \end{aligned} \quad (4)$$

Intuitively, the first step finds an assignment to \mathbf{y} that is “close” to predictions made by applying weights \mathbf{w} , and also satisfies proportions constraints. The second step re-adjusts \mathbf{w} . Our alternating algorithm for this bi-convex problem is thus guaranteed to descend, decreasing the objective in every iteration.

In practice, we replace \mathbf{y} with $\text{Sign}(\mathbf{y})$ ($\text{Sign}(\cdot)$ applied elementwise) in order to use efficient off-the-shelf SVM solvers (See Fig. 2). Empirically, in most cases we observed that \mathbf{y} were very close to either 1 or -1 .

To start off the alternation, we need to initialize \mathbf{w} . Specific label proportions constraints are handled by modeling the latent \mathbf{y}^* directly, which is only possible in our alternating scheme once a vector \mathbf{w} is fixed. Thus, we start the alternating optimization process by first solving the following simple convex program, which uses only the partial order between bags. Let the set of pairwise orderings \mathcal{P} be the set of all tuples $(\mathcal{B}_{k_1}, \mathcal{B}_{k_2})$ such that $p_{k_1} \geq p_{k_2}$. To find our initial \mathbf{w} we solve:

$$\begin{aligned}
 & \underset{\mathbf{w}, \xi}{\operatorname{argmin}} \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{1}{|\mathcal{P}|} \sum_{p=1}^{|\mathcal{P}|} \xi_p + \frac{C_L}{L} \sum_{j=N+1}^{N+L} \xi_j \\
 \text{s.t. } & y_j \mathbf{w}^T \varphi(\mathbf{x}_j) \geq 1 - \xi_j \quad \forall j \in \{N+1, \dots, N+L\} \\
 & \mathbf{w}^T \frac{1}{|\mathcal{B}_{k_1}|} \sum_{i \in \mathcal{B}_{k_1}} \varphi(\mathbf{x}_i) \geq \mathbf{w}^T \frac{1}{|\mathcal{B}_{k_2}|} \sum_{i \in \mathcal{B}_{k_2}} \varphi(\mathbf{x}_i) - \xi_p \\
 & \forall \{k_1 \neq k_2 : (\mathcal{B}_{k_1}, \mathcal{B}_{k_2}) \in \mathcal{P}\},
 \end{aligned} \tag{5}$$

The second constraint in Problem 5 amounts to representing bags with their (\mathbf{w} -weighted) mean in feature-space. Note that in order for a bag \mathcal{B}_k to be well-approximated by its mean in feature-space, \mathcal{B}_k should induce a low-variance distribution over bag instances. This is a strong assumption, but yields a simple quadratic program easy to solve quickly with standard solvers, and empirically leads to good starting points in parameter-space. We additionally note that when $C_L = 0$ (no labels), we recover as a special case the

Input: $\mathbf{x}, \mathcal{R}, \mathcal{D}, C$

1. **Init** \mathbf{w}^0 : $\mathbf{w}^0 \leftarrow$ Solution to (5)
2. **Repeat**
 - (a) Solve (3) for \mathbf{y}^t w.r.t \mathbf{w}^{t-1}
 - (b) Solve an SVM problem for \mathbf{w}^t w.r.t $\operatorname{Sign}(\mathbf{y}^t)$ and cost parameter C

until $\frac{\|\mathbf{w}^t - \mathbf{w}^{t-1}\|_2}{\|\mathbf{w}^{t-1}\|_2} \leq 10^{-5}$

Return \mathbf{w}

Fig. 2. Alternating algorithm

Multiple-Instance (MI) ranking problem proposed in the image-retrieval framework of [13], albeit with a different objective (we are interested in classifying instances rather than learning to rank bags). We note that in Problem 3, we impose hard constraints on label proportions. Certain sets of constraints could, of course, be infeasible. In this case, a practitioner might adjust the constraints, or simply make them soft (by adding slack variables).

Optimizing C . In practice, we need to tune hyperparameter C . This is typically done with cross-validation (CV) grid search, measuring performance on held-out data. However, standard CV is impossible here, as we have no labeled examples.

We thus develop a novel variant of CV, suited for our setting. We run K -fold CV, splitting each bag \mathcal{B}_k into training and held-out subsets. The intuition is that the label proportion in uniformly-sampled subsets of a bag is similar to the proportion p_k in the entire bag. For each split we run Algorithm 2 on training bags, and then compute by how much constraints are violated on *held-out* bags. More formally, we compute the average deviations from bounds, $\max(\hat{p}_k - u_k, 0)$, $\max(l_k - \hat{p}_k, 0)$ for \hat{p}_k the estimated label proportion in the held-out subset of bag k . We do so over a grid, and select the C with lowest average violation.

4 Evaluation

In order to evaluate our algorithm, we prepared the following datasets:

- **Movie Reviews:** The Movie Reviews dataset [17] contains 1000 positive and 1000 negative movie reviews written before 2002. The task is to classify the sentiment of movie reviews as positive or negative.
- **Census:** The Adult dataset [1] (48842 instances) is from the Census bureau. The task is to predict whether a given adult makes more than \$50,000 a year based on attributes such as education, hours of work per week, etc.

For each of the classification tasks described, we run 10-fold cross-validation and report average results (note that labels are used only for testing). For text classification tasks, feature map $\varphi(\cdot)$ is the standard TF-IDF features.

We formed bags corresponding to the different tasks (see below), demonstrating the wide applicability of the setting and our approach. In order to test our method’s robustness we used approximate constraints, at times violating the true underlying proportions.

Baselines. To the best of our knowledge, no other method aims to solve the problem of Sect. 2. Thus, we compare ourselves to three natural baselines.

- **“High vs. low”:** One reasonable approach in our setting is to create two sets of instances: The “high” set contains instances from bags with the highest label proportions, and the “low” set – from bags with the lowest proportions. The idea is to pretend all instances in the “high” set are positive, and in the “low” set – negative, and learn a classifier with the noisy labels. To make the baseline stronger, we use grid search to optimize hyper-parameter C (chosen from a commonly used grid for SVM C values, $[10^{-4}, 10^{-3}, \dots, 10^3, 10^4]$, with 10-fold cross-validation and selecting C with best average). To counter the class-imbalance created, we apply a weighted SVM.
- **Supervised SVM:** Our method does not need labeled instances, but instead uses weaker, aggregate information. To show how many labels are needed to obtain comparable results to our method, we report SVM results over a labeled training set (note that this information is not available to our algorithm). We use grid search to optimize hyper-parameter C as above.
- **Learning from labeled proportions:** For the census data set, we compare our method’s performance to results reported in [20] using known label proportions with various algorithms. Note that our method does not have access to the exact label proportions.

For our method, we select C using the constraint-violation approach described in the previous section.

We run the procedure for a maximum of 200 iterations, with convergence typically occurring long before. A typical iteration (for one value of C , one CV split) took at most a few seconds on a standard laptop. Our data is available on <https://github.com/ttthhh/ballpark.git>.

4.1 One-Word Classifier

Our first task is to classify sentiment of movie reviews. Our goal is not to compete with the host of previous sentiment-analysis algorithms [18] in terms of accuracy, but rather to provide a light-weight tool when very little information and

resources are available: a “poor-man’s” classifier. In this section, we show how we are able to obtain good results while assuming very scarce prior knowledge with simple, clean tools.

We envision a practitioner who knows a very simple fact – that reviews containing the word “great” are more likely to be positive than negative, but far from exclusively: many positive reviews do not use the word “great”, and some negative reviews do use it (“horrific performance by a usually great actor”).

We construct three bags: $\mathcal{B}_{\text{great}}$, $\mathcal{B}_{\text{good}}$, \mathcal{B}_{bad} , each containing reviews with the corresponding word in them (note the bags are not necessarily disjoint). For the three bags created on training set instances (10-fold CV) we find that $|\mathcal{B}_{\text{great}}| \approx 700$, $|\mathcal{B}_{\text{good}}| \approx 630$, $|\mathcal{B}_{\text{bad}}| \approx 160$, $p_{\text{great}} \approx 0.6$, $p_{\text{good}} \approx 0.45$, $p_{\text{bad}} \approx 0.25$.

For simplicity, we assume no labels are given, but the practitioner has rough estimates for proportions. This information could come from a sample or from domain knowledge. In our experiment, we assume an upper bound on the bag with the highest proportion and a lower bound for each bag. We used a weak bound for each bag, underestimating it by 50%. We also assumed that $p_{\text{great}} > p_{\text{good}} > p_{\text{bad}}$. Again we use a weak bound, overestimating the real difference by 33%. In Sect. 4.3 we explore how the tightness of the constraints affects accuracy, showing our method is robust to loose constraints.

For the “*high vs. low*” baseline, we take bag $\mathcal{B}_{\text{great}}$, $\mathcal{B}_{\text{good}}$ as the positive class, and \mathcal{B}_{bad} as the other. As seen in Table 1, our method outperforms this naive baseline, and competes with supervised SVM trained on considerable amounts of labeled examples. Given fewer labels, supervised SVM is inferior to our label-free method: providing SVM with 25 labeled instances leads to accuracy of 0.51, 50 labels to accuracy of 0.63, and 75 labels increases accuracy to 0.69.

To test stability, we run the same experiment using different words to create the bags. The results are similar. Table 1 shows the results using “excellent”, “nice”, and “terrible”. To make sure the classifier is not learning our input words, we test removing these words (e.g., “good”) from the documents. In our experiment, the removal reduced accuracy by less than 1%.

Table 1. Movie results for different sets of bags based on different choices of words. Our method outperforms the naive SVM baseline, and rivals a supervised SVM with a considerable number of labels.

Method	$\mathcal{B}_{\text{great}}, \mathcal{B}_{\text{good}}, \mathcal{B}_{\text{bad}}$	$\mathcal{B}_{\text{excellent}}, \mathcal{B}_{\text{nice}}, \mathcal{B}_{\text{terrible}}$
Bag constraints	0.71	0.73
“high vs. low” SVM	0.52	0.55
Supervised SVM	100 labels (0.71)	100 labels (0.71)

4.2 Learning from Stereotypes

In this section we simulate a scenario frequently occurring in practice. We have a large sample of individuals, and would like to predict their level of income

using socio-demographic information. One variable that is known to be correlated with income is *education level*. This information is difficult to obtain (budgetary constraints, privacy issues, respondents’ reluctance etc.) and is available only for a small sub-sample. In addition, we have no labels – individuals with known income. We do have ballpark-estimations on income proportions for different education levels, and the difference between them (based on an earlier census, expert assessments or other external sources).

In our first experiment we construct bags based on education level: $\mathcal{B}_{\text{Masters}}$, $\mathcal{B}_{\text{Bachelors}}$, $\mathcal{B}_{\text{Some-college}}$, $\mathcal{B}_{\text{High-School}}$. Over 20-fold CV (size of training set ≈ 1220) we find that $|\mathcal{B}_{\text{Masters}}| \approx 90$, $|\mathcal{B}_{\text{Bachelors}}| \approx 265$, $|\mathcal{B}_{\text{Some-college}}| \approx 360$, $|\mathcal{B}_{\text{High-School}}| \approx 520$, $p_{\text{Masters}} \approx 0.55$, $p_{\text{Bachelors}} \approx 0.42$, $p_{\text{Some-college}} \approx 0.19$, $p_{\text{High-School}} \approx 0.16$.

We use similar constraints to the previous section, but remove all lower bounds on bags, thus incorporating even less prior information than before. For the *SVM using “high vs. low”* baseline, we use $\mathcal{B}_{\text{Masters}}$, $\mathcal{B}_{\text{Bachelors}}$ as one class, and $\mathcal{B}_{\text{Some-college}}$, $\mathcal{B}_{\text{High-School}}$ as the other.

We start with basic features: age, gender, race. After assigning individuals to education bags, we discard education features from the data – we assume not to have this information at test time (only for a small sub-sample available for training). We do retain those features for the *Supervised SVM* baseline. Our method achieves cross-validation accuracy of 0.74, while the baseline achieves 0.57. Supervised SVM, even with 1000 labeled examples, only reaches 0.71.

We also experiment with using less bags (removing “Masters”), and with an expanded feature set (age, race, gender, hours-per-week, capital-gain, capital-loss). See Table 2 for results. Here too, our method outperforms the baseline, and rivals supervised SVM with 900 labels.

Of course, we are not limited to using bags based on only education level. Another well-known correlation is between gender and income. Thus, we can also slice the data into bags based on education *and* gender. In another experiment we create 6 bags, $\mathcal{B}_{\text{Bachelors+Female}}$, $\mathcal{B}_{\text{Some-college+Female}}$, $\mathcal{B}_{\text{High-School+Female}}$, $\mathcal{B}_{\text{Bachelors+Male}}$, $\mathcal{B}_{\text{Some-college+Male}}$, $\mathcal{B}_{\text{High-School+Male}}$. There are stark differences in label proportions between the groups, notably in favor of males.

For the SVM using *“high vs. low”* baseline, we try two different class assignments. We start from Bachelors vs. everyone else. (It could seem more natural to take, for example, $\mathcal{B}_{\text{Bachelors+Male}}$ as the “high” bag and $\mathcal{B}_{\text{High-School+Female}}$ as “low”, but this results in too small a sample). The baseline performed relatively well (Table 2) due to good class separation. However, when we tried females vs. males, performance of our method remained stable (with highest accuracy), but the baseline suffered a drastic drop (Table 2). This highlights the difficulty of using this baseline when using multiple bags based on richer information: it is not immediately clear how to create two well-separated classes. On the other hand, our method naturally compares groups based on given constraints.

More Baselines. Finally, we report classification accuracy on the same dataset, taken from [20]. The authors create two artificial bags, one retaining original label proportions and another containing only one class. With these bags, their method

Table 2. Census results for different sets of bags. Our method outperforms the naive SVM baseline, and rivals a supervised SVM with many labeled examples.

Method	Education bags	Edu + Gender
Bag constraints	0.75	0.77
“high vs. low” SVM (Bachelors vs. other)	0.52	0.6
“high vs. low” SVM (Female vs. Male)	-	0.38
Supervised SVM	0.75 (900 labels)	0.77 (900 labels)

(using known proportions) achieved 0.81 accuracy. They also report results for Kernel Density Estimation (0.75), Discriminative Sorting – a supervised method (0.77), MCMC sampling (0.81), and a baseline of predicting the major class (0.75). Our method achieves comparable performance despite having much less information on label proportions, fewer features, and using more realistic bags.

4.3 Sensitivity Analysis

In this section we give a short demonstration of how the tightness of constraints could affect model performance. We create artificial bags and vary the tightness of some constraints, reporting accuracy. This is a preliminary study, serving to illustrate some of the different factors that come into play.

We use the **20 Newsgroups** dataset [2] containing approximately 20,000 posts across 20 different newsgroups. Some of the newsgroups are closely related (e.g., *comp.sys.ibm.pc.hardware* and *comp.sys.mac.hardware*), while others are further apart (*rec.sport.hockey* and *sci.space*). The task is to classify messages according to the newsgroup to which they were posted.

We assume predefined bags and vary constraints on label proportions within and between bags. We do not use any labeled data at training time.

We examine three binary classification tasks, between different categories of posts: *space* vs. *medicine*, *ibm.pc* vs. *mac*, and *hockey* vs. *baseball*. For each of these binary classification tasks, we create six bags of training instances $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_6\}$. The sizes of each bag are $|\mathcal{B}_1| = |\mathcal{B}_2| = 200, |\mathcal{B}_3| = |\mathcal{B}_4| = 50, |\mathcal{B}_5| = |\mathcal{B}_6| = 100$. We thus use only 650 instances in this case – about half of the 1187 in the training set. The real label proportions within each bag are $p_1 = p_2 = 0.5, p_3 = p_4 = 0.3, p_5 = p_6 = 0.2$.

We test the effects of three different types of constraints, corresponding to common types of aggregate information:

- **Upper bounds on bag proportions:** Let k_{max} be the index of the bag with the highest proportion. We assume an upper bound multiplicative factor only on this bag: $p_{k_{max}} \times u_m$, where we control factor u_m .
- **Lower bounds:** For each true p_k , we take as a lower bound $l_p * p_k$.
- **Bag difference bounds:** For each true $p_{k_1} - p_{k_2}$ such that $p_{k_1} \geq p_{k_2}$, we lower-bound the difference with $l_d \times (p_{k_1} - p_{k_2})$.

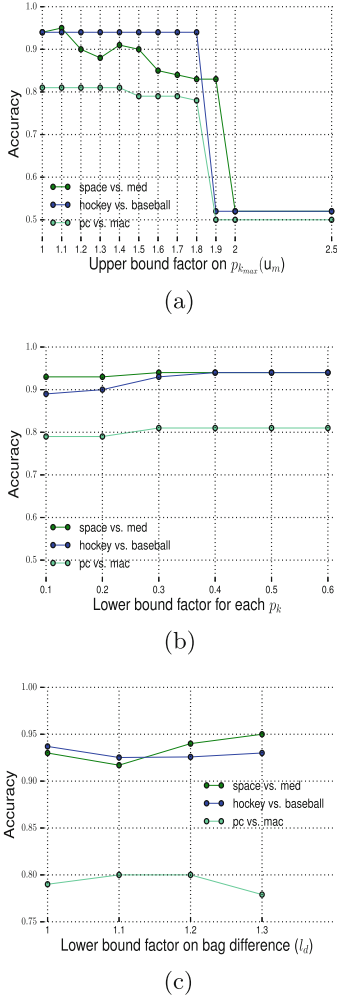


Fig. 3. Constraint effects. Accuracy results on a validation set: (a) Varying upper-bound factor on highest p_k (b) Varying individual lower-bound factor (c) Varying lower bound on bag differences. Results remain fairly robust (with fluctuations due to small-sample noise). The graph stops abruptly where constraints are no longer feasible.

4.4 Simulation Study

To further test the behavior of our algorithm, we conduct simulation studies on synthetic data. We use the built-in simulation function *make_classification*

Figure 3 shows the results of our experiments. In our initial setting, we take a fairly loose configuration of constraints to test our method’s flexibility: $l_d = 1$ (no lower bound at all for bag differences), $l_p = 0.5$, and $u_m = 1$. In each experiment we vary one factor, keeping the others fixed: (a) upper bound on $p_{k_{max}}$, (b) individual lower bound, (c) lower bound on bag differences.

Notable in Fig. 3 is the overall robustness of the method to misspecified constraints. As u_m is gradually increased, performance remains overall stable for a long stretch (Fig. 3a). However, when u_m reaches extremely large values, the upper bound on $p_{k_{max}}$ becomes too loose (reaching 1) and robustness collapses. Increasing the lower bound on individual p_k slightly improves results, by tightening constraints (Fig. 3b). Results remain fairly robust to overestimating the lower bound on bag differences by increasing l_d , with fluctuations due to small-sample noise (Fig. 3c). The graph stops abruptly at $l_d = 1.3$ since beyond that point constraints are no longer feasible.

Finally, we compare results to the baselines of the previous section. For our method, we fix $u_m = 1, l_p = 0.5, l_d = 1.33$ (with no upper bound on bag differences, as in previous sections). For the SVM using “high vs. low” baseline, we take bags $\mathcal{B}_1, \mathcal{B}_2$ as one class, and $\mathcal{B}_5, \mathcal{B}_6$ as the other (adding $\mathcal{B}_3, \mathcal{B}_4$ led to inferior results). Our method outperforms this naive baseline, and also competes with supervised SVM trained on considerable amounts of labeled examples (Table 3). Given fewer labels, supervised SVM was inferior to our label-free method.

Table 3. 20 newsgroups results.

Method	med-space	pc-mac	baseball-hockey
Bag constraints	0.94	0.81	0.94
“high vs. low” SVM	0.82	0.62	0.64
Supervised SVM	110 labels (0.93)	95 labels (0.78)	140 labels (0.94)

provided in python package scikit-learn [19] to generate data for a binary classification problem. We create three equally-sized bags of instances $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$ for our training set, with label proportions p_1, p_2, p_3 , respectively. We vary bag sizes $|\mathcal{B}_k|$ and proportions p_k , as well as the number of features ($n_features$), number of informative features ($n_informative$), and class separation ($class_sep$).

We apply our cross-validation procedure to select \mathcal{C} , using 3 folds. We observe some typical behaviors, such as accuracy improvement with growing sample size. For instance, fixing $n_features=20$, $n_informative = 1$ and $p_1 = 0.4, p_2 = 0.3, p_3 = 0.2$, mean accuracy increases from 0.65 with $|\mathcal{B}_k| = 500$, to 0.77 with $|\mathcal{B}_k| = 1000$.

Accuracy suffered with smaller gaps between bag proportions p_k . However, with increasing sample size our algorithm got better at handling minuscule differences between p_k . For example, fixing $p_1 = 0.4, p_2 = 0.35, p_3 = 0.33$, mean accuracy increases from 0.6 with $|\mathcal{B}_k| = 500$ to 0.65 with $|\mathcal{B}_k| = 1000$, and further increases to 0.67 with $|\mathcal{B}_k| = 1500$.




Finally, we expect that labeled instances can improve performance, helping to counter bags that are very noisy or constraints that are not sufficiently tight. Preliminary experiments suggest that labeled instances can improve accuracy, but a comprehensive study of this effect is beyond the scope of this paper.

5 Exploratory Analysis

In previous sections we tackled classification problems with a clear objective. In this section our users have no specific classification in mind, but rather are interested in exploring the data. A sub-field within clustering allows users to guide the formation of clusters, usually in the form of pairwise constraints on instances (forcing data points to belong to the same cluster or to different clusters). A recent approach uses a maximum-margin framework [28], which extends the supervised large margin theory (such as SVMs) to an unsupervised setting. Similarly, we adapt our method to the exploratory setting. Rather than using instance-level constraints on cluster membership, we use *ranking* constraints based on prior knowledge – or on hypotheses we would like to explore.

We used the **Geo-tagged tweets** dataset, containing 377616 messages from 9475 geo-located microblog users over one week in March 2010 [9]. The user base is likely dominantly composed of teens and young adults (as some of the examples below will make clear). We combine all tweets for each user, and reverse-geocode the GPS coordinates to obtain the corresponding state.

Table 4. Geo-tagged tweets. For each set of geographic constraints, we show some of the top positive and negative words resulting from running our method.

Constraints	Positive terms	Negative terms
 French > English: Quebec > Texas	je, est, et, le, pour	Houston, Texas, dal- las, bro, tryna, boo
 East Coast > West Coast: CA > NY, CA > PA, WA > NY, WA > PA	hella, coo, fasho, af, la, cali, san, washing- ton	deadass, niggas, skool, wassup, dis, dat, philly, crib, lml, nah, dey, den
 Ranking by religiosity: Alabama > Florida > New York	thank, easter, pray, road, trip, drove, loving, relationship, spring, folks, happy, dreams, laugh, friend	mad, bitches, neva, dis, dat, niggas,ova, spanish, girls, crazy, party, fun, high, dead

The dataset was used in [9] to analyze regional dialects. The authors used a cascading topic model to model geographic topic variation. The observed output of the generative process includes the texts and GPS coordinates of each user. We pursue this line of exploration too, but rather than positing a generative model of language, we investigate how various constraints on differences between geographic locations interact with dialect.

In Table 4, we show some of the constraints we explored and the resulting top positive and negative words. We start with a simple check with two bags $\mathcal{B}_{\text{Quebec}} \succeq \mathcal{B}_{\text{Texas}}$, combining tweets from Quebec and Texas, respectively. We discover obvious differences in language, with strong positive weights corresponding to French words and negative weights to English.

An ordinary classifier would likely recover similar results, as would standard unsupervised clustering algorithms. However, our method allows to pursue richer, more *expressive* constraints. First, we look into the difference between the East Coast and West Coast by imposing pairs of constraints such as $\mathcal{B}_{\text{California}} \succeq \mathcal{B}_{\text{New York}}$, $\mathcal{B}_{\text{California}} \succeq \mathcal{B}_{\text{Pennsylvania}}$. We recover various results previously highlighted by [9], such as the use of the slang terms “fasho” (for sure) “coo” (cool), “hella” in the West Coast, and “deadass”, “wassup” and “niggas” in the East Coast. Our results agree with findings by [9, 10], as well as suggest some potential new findings.

Finally, we look at a set of more expressive constraints, aiming to recover difference based on religiosity (or at least sociological confounders). We take states from the top, middle and bottom of a list of US states ranked by percentage of self-reported religiosity¹, and build sets of constraints that reflect this ordering. For instance, in Table 4, we show results for $\mathcal{B}_{\text{Alabama}} \succeq \mathcal{B}_{\text{Florida}} \succeq \mathcal{B}_{\text{New York}}$. Note that using such information in a standard classifier is unnatural. It is not

¹ https://en.wikipedia.org/wiki/List_of_U.S._states_by_religiosity.

clear how to construct the classes, and different splits could lead to very different results. Again, this artificial splitting is not required by our method.

We removed terms not in the wordnet [11] lexicon to mitigate the effects of local vernacular and highlight deeper differences. The differences in language are quite striking. As we traverse from Alabama to Florida to New York, discourse shifts from words such as “glad”, “loving”, “happy”, “dreams”, “easter” and “pray”, to words including “mad”, “bitches”, “crazy”, “party”, “fun”, “high” and other more profane content we spared from the reader. Similar results were obtained for other state tuples (e.g., Texas instead of Alabama).

Note that our method can be used for formulating new hypotheses. To test the hypotheses, more experiments (and often more data collection) are needed. We leave it up to sociologists to provide deeper interpretation of these results.

Our goal in this section was to use coarse prior information (in the form of relative rankings) for exploring a dataset. We note that the problem could be tackled with other approaches, such as topic models or classification. However, classification models assume a much stronger discriminative pattern or signal than taking a softer, weakly-supervised approach that seeks a direction (weight vector \mathbf{w}) along which one bag of instances is ranked higher than another. While clustering with pairwise memberships constraints is well-studied, we demonstrate clustering with expressive pairwise *ranking* constraints over sets. Many real-world settings naturally lend themselves to this formulation.

6 Discussion and Criticism

One clear practical issue with our method is the source of the constraints. We have illustrated several real-world cases where it is plausible to attain rough constraints on label proportions within and between groups of instances. In previous work [20], it is suggested that practitioners could sample from bags of instances to estimate label frequencies (e.g., in spam classification tasks). However, accurate estimations might require extensive sampling, exacting high costs. We thus relax this rather strong assumption, and propose that in many cases, it is possibly enough to get rough estimates. For example, after sampling 10 instances, we might observe 9 positives and only one negative, and rather conservatively declare “ \mathcal{B} should have more than 50% positives”. This sort of statement could of course be made more rigorous with probabilistic considerations (e.g., confidence intervals). We have demonstrated that even with considerably mis-specified constraints, we are still able to achieve good performance across various domains.

Furthermore, external sources of knowledge could be used to construct these constraints, such as previous surveys. In many cases taking exact figures from surveys (such as political polls) and expecting them to accurately reflect the distribution in new data is not realistic. This is the case, for instance, when looking at national political polls and wishing to extrapolate from them to new very different socio-demographic slices, such as Twitter users. Here too, we could use this external knowledge to *approximately* guide our model, rather than dictate precise hard proportions the model should match.

7 Related Work

There is a large body of work that is related to our problem.

Multiple Instance Learning. The field of *Multiple Instance Learning* (MIL) generally assumes instances come in “bags”, each associated with a label modeled as a function of latent instance-level labels, which can be seen as a form of weak supervision. MIL methods vary by the assumptions made on this function. For a comprehensive review of assumptions and applications, see [7, 12]. Most work in MIL focuses on making bag-level predictions rather than for individual instances. Recently, [15] used a convolutional neural network to predict labels for sentences given document-level labels.

Learning from Proportions. A niche within MIL which has seen growing interest and is closely-related to this paper, is concerned with predicting instance-level labels from known label proportions given for each bag. [20] assume to be given bags of unlabeled examples, each bag with *known* label proportions. Their method is based on estimating bag-means using given label proportions. The authors provide examples for scenarios in which such information could be available. In [21], the authors represent each bag with its mean, and model the known class proportions based on this representative “super-instance” with an SVM method, showing superior performance over [20]. In [27], instance-level labels are explicitly modeled to overcome issues the authors raise with representing bags with their means, such as when data distribution has high variance. The fundamental property these and other approaches share is that bag proportions are assumed be known or easily estimated, an assumption we relax.

Classification with Weak Signals. We applied our model to the problem of text classification when little or no labels are available but only a weaker signal. A vast amount of literature has tackled similar scenarios over the years, using tools from semi-supervised [6, 14] active [16, 22, 24] and unsupervised [3] learning. Druck et al. [8] apply generalized expectation feature-labeling (GE-FL) approaches, using “labeled features” given by an oracle that encode knowledge such as “the word puck is a strong indicator of hockey”. In practice, a Latent Dirichlet Allocation (LDA) [5] topic model is applied to the data to select top features per topic, for which a user provides labels. [23] propose a semi-supervised + active-learning method, with a human-in-the-loop who provides both feature-level and instance-level labels. We are also able to use labeled instances to refine the learning process, allowing for a trade-off between the user’s trust in the (typically few) labeled instances available, and prior knowledge on bag proportions.

Similar to the above work on learning from labeled proportions, [25] considers a classification problem with no access to labels for individual training examples, but only average labels over subpopulations. They frame the problem as weakly-supervised clustering. When using our method for exploratory analysis, it can also be seen as a weakly-supervised clustering algorithm, using information on partial ordering between bags rather than assuming known proportions, within a max-margin framework (somewhat akin to clustering using maximum-margin as in [28]). The seminal work of [26] uses side-information for clustering in the form

of pairwise constraints on cluster membership (pairwise similarity). Much work has since been done along these lines. We incorporate pairwise constraints in our maximum-margin approach, though with pairs representing bags of instances, and partial ordering with respect to relative label proportions.

Robust Optimization. Finally, robust optimization [4] research deals with uncertainty-affected optimization problems, by optimizing for the *worst-case* value of parameters. Because of its worst-case design, robust optimization can do poorly when the constraints are not tight. Our method, on the other hand, is designed to handle rough estimates and loose constraints.

8 Conclusions and Future Work

In this paper we proposed a new learning setting where we have bags of unlabeled instances with loose constraints on label proportions and difference between bags. Thus, we relax the unrealistic assumption of known bag proportions.

We formalized the problem as a bi-convex optimization problem and proposed an efficient algorithm. We showed how, surprisingly, our classifier performs well using very little input. We also demonstrated how the algorithm can guide exploratory classifications.

We have empirically studied the behavior of our algorithm under different types of constraints. One direction for future work is to analytically understand, for instance, how constraint tightness affects performance, obtain convergence guarantees, and provide generalization error bounds. This, in turn, could perhaps lead to better algorithms with theoretical justifications.

Finally, the relative-proportions setting is very natural, and can be found in various domains. We believe that this line of work will have interesting implications regarding privacy and anonymization of data – in particular, the amount of information one can recover using only weak, aggregated signals.

Acknowledgments. The authors thank the anonymous reviewers and Ami Wiesel for their helpful comments. Dafna Shahaf is a Harry&Abe Sherman assistant professor, and is supported by ISF grant 1764/15 and Alon grant.

References

1. <https://archive.ics.uci.edu/ml/datasets/Adult/>
2. <http://qwone.com/~jason/20Newsgroups/>
3. Aggarwal, C.C., Zhai, C.X.: A survey of text clustering algorithms. In: Aggarwal, C.C., Zhai, C.X. (eds.) *Mining Text Data*, pp. 77–128. Springer, New York (2012)
4. Ben-Tal, A., El Ghaoui, L., Nemirovski, A.: *Robust Optimization*. Princeton University Press, Princeton (2009)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003)
6. Chapelle, O., Schölkopf, B., Zien, A., et al.: *Semi-supervised Learning*. MIT Press, Cambridge (2006)

7. Cheplygina, V., Tax, D., Loog, M.: On classification with bags, groups, sets. arXiv preprint [arXiv:1406.0281](https://arxiv.org/abs/1406.0281) (2014)
8. Druck, G., Mann, G., McCallum, A.: Learning from labeled features using generalized expectation criteria. In: SIGIR 2008, pp. 595–602 (2008)
9. Eisenstein, J., Brendan, O., Smith, N., Xing, E.P.: A latent variable model for geographic lexical variation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Cambridge, MA (2010)
10. Eisenstein, J., Smith, N.A., Xing, E.P.: Discovering sociolinguistic associations with structured sparsity. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (2011)
11. Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998)
12. Foulds, J., Frank, E.: A review of multi-instance learning assumptions. *Knowl. Eng. Rev.* **25**, 125 (2010)
13. Hu, Y., Li, M., Yu, N.: Multiple-instance ranking: learning to rank images for image retrieval. In: Proceedings of CVPR, p. 18 (2008)
14. Joachims, T.: Transductive inference for text classification using support vector machines. In: ICML 1999, pp. 200–209 (1999)
15. Kotzias, D., Denil, M., de Freitas, N., Smyth, P.: From group to individual labels using deep features. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2015 (2015)
16. Li, L., Jin, X., Pan, S.J., Sun, J.-T.: Multi-domain active learning for text classification. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1086–1094. ACM (2012)
17. Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity. In: Proceedings of ACL, pp. 271–278 (2004)
18. Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Found. Trends Inf. Retrieval* **2**(1–2), 1–135 (2008)
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
20. Quadrianto, N., Smola, A.J., Caetano, T.S., Le, Q.V.: Estimating labels from label proportions. *J. Mach. Learn. Res.* **10**, 2349–2374 (2009)
21. Rueping, S.: SVM classifier estimation from group probabilities. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010) (2010)
22. Settles, B.: Active learning literature survey. University of Wisconsin, Madison, 52(55-66):11
23. Settles, B.: Closing the loop: fast, interactive semi-supervised annotation with queries on features and instances. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1467–1478 (2011)
24. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.* **2**, 45–66 (2002)
25. Wager, S., Blocker, A., Cardin, N.: Weakly supervised clustering: learning fine-grained signals from coarse labels. *Ann. Appl. Stat.* **9**(2), 801–820 (2015)
26. Xing, E.P., Jordan, M.I., Russell, S.J., Ng, A.Y.: Distance metric learning with application to clustering with side-information. In: NIPS 2003. MIT Press (2003)
27. Yu, F., Liu, D., Kumar, S., Jebara, T., Chang, S.: α -SVM for learning with label proportions. In: ICML 2013 (2013)
28. Zhou, G.-T., Lan, T., Vahdat, A., Mori, G.: Latent maximum margin clustering. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (eds.) *Advances in Neural Information Processing Systems 26*, pp. 28–36 (2013)