Dependence and Model Selection in LLP: The Problem of Variants

Gabriel Franco
Boston University
Boston, USA
gvfranco@bu.edu

Mark Crovella
Boston University
Boston, USA
crovella@bu.edu

Giovanni Comarela
Federal University of Espírito Santo
Vitória, Brazil
gc@inf.ufes.br

ABSTRACT

The problem of Learning from Label Proportions (LLP) has received considerable research attention and has numerous practical applications. In LLP, a hypothesis assigning labels to items is learned using knowledge of only the proportion of labels found in predefined groups, called *bags*. While a number of algorithmic approaches to learning in this context have been proposed, very little work has addressed the model selection problem for LLP. Nonetheless, it is not obvious how to extend straightforward model selection approaches to LLP, in part because of the lack of item labels. More fundamentally, we argue that a careful approach to model selection for LLP requires consideration of the dependence structure that exists between bags, items, and labels. In this paper we formalize this structure and show how it affects model selection. We show how this leads to improved methods of model selection that we demonstrate outperform the state of the art over a wide range of datasets and LLP algorithms.

CCS CONCEPTS

• **Computing methodologies** → **Cross-validation**; **Semi-supervised learning settings**.

KEYWORDS

Learning from Label Proportions; Hyperparameter Selection; Weakly Supervised Learning

ACM Reference Format:
Gabriel Franco, Mark Crovella, and Giovanni Comarela. 2023. Dependence and Model Selection in LLP: The Problem of Variants. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3580305.3599307

1 INTRODUCTION

The problem of Learning from Label Proportions (LLP) seeks to build a classifier $f : \mathcal{X} \mapsto \mathcal{Y}$ that maps items $\mathbf{X} \in \mathcal{X}$ to labels $Y \in \mathcal{Y}$. However, unlike the standard supervised learning problem, in LLP the data available for learning f does not consist of $(\mathbf{X}, Y)$ pairs. Rather, what is available is an assignment of each item to a group (called a 'bag'), and knowledge of the *proportion* of positive labels among the items in each bag.

KDD '23, August 6–10, 2023, Long Beach, CA, USA.

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599307

The LLP problem has received considerable research attention and has numerous practical applications. For example, it has been used to generate fine-grained predictions of public opinion [7], to aid in embryo selection during assisted reproduction [13], as a tool in industrial quality control [42], and to infer the demographics of Twitter users [2].

However, although many algorithms for LLP have been proposed (see §3), the problem of hyperparameter setting, ie *model selection,* has rarely been addressed (to our knowledge, the only work devoted to hyperparameter selection in LLP is [12]). This is surprising, because it is not obvious how to extend standard hyperparameter selection strategies to use with LLP algorithms. In particular, standard approaches to hyperparameter selection involve holding out a portion of the labeled data to test generalization ability. However, in the LLP problem, data items are not individually labelled, and so held-out data can have unknown properties.

In this paper we take a close look at the challenge of model selection in LLP. We argue that to put model selection on firm ground, it is necessary to consider aspects of the LLP problem that have typically been overlooked in prior work. In particular, we show that model selection is strongly affected by *dependence relations* between bags (B), item features ($\mathbf{X}$), and labels (Y). In other words, while the standard supervised learning problem involves only the dependence of labels on features, the LLP problem additionally involves a large set of additional potential dependence relationships between features, labels, and bags.

Hence, the first contribution of this work is to introduce a *taxonomy* of LLP problem variants. Our taxonomy is based on the dependence and conditional dependence present between B, $\mathbf{X}$, and Y. We show that prior work has rarely considered the impacts of such dependence relations in LLP, and no prior work has dealt with LLP variants in a systematic way. This new taxonomy of LLP variants provides a framework for classifying LLP problems, for defining LLP benchmarking strategies, and for guiding model selection when presented with a new LLP problem instance.

Our second contribution is to demonstrate the value of our LLP taxonomy by using it to derive new LLP model selection strategies. Understanding model selection in LLP is important, because all algorithms proposed to date for solving LLP incorporate hyperparameters. However, very rarely in the literature has the method of hyperparameter selection been precisely stated, which is problematic. For example, we show through extensive experiments that a given model on a given dataset will typically have radically different generalization ability depending on the method used for holding out data during model selection.

Finally, we evaluate our new model selection strategies and show their superiority over the state of the art (ie, [12]). Our experiments study performance across over 100 LLP test cases (encompassing 4 LLP problem variants) to which we apply 3 LLP algorithms, each

using 4 model selection strategies. In all we report the results of 954 distinct experiments, which are each repeated several times to enable comparisons with statistical significance. Our results show that the new LLP model selection strategies derived from our taxonomy are superior to the state of the art in 90% of the statistically distinguishable cases. Furthermore, our results show that understanding the LLP variant at hand provides direct insight into the relative benefit of different model selection strategies, and so guides their deployment in practice.

2 LLP AND ITS VARIANTS

2.1 Basic Definitions

We consider the LLP problem involving N items taken from feature space $\mathcal{X}$, and $L \leq N$ bags. For simplicity, we consider a binary classification problem, although the taxonomy we will develop does not depend on this assumption. A problem instance is given by a set of pairs $D = \{(\mathbf{x}_i, b_i), i = 1, \ldots, N\}$ and a vector $\mathbf{p} \in [0,1]^L$, with $\mathbf{x}_i \in \mathcal{X}$ and $b_i \in \{1, \ldots, L\}$. To define $\mathbf{p}$, we assume that each $\mathbf{x}_i$ is associated to an (unknown) label $y_i \in \{0, 1\}$. Then

$$p_\ell = \frac{|\{i \,|\, b_i = \ell, y_i = 1\}|}{|\{i \,|\, b_i = \ell\}|}.$$

In other words, p_ℓ is the proportion of 1's in bag ℓ.

The challenge posed by LLP is that the vector $\mathbf{p}$ is known, but the individual labels $\{y_i\}$ are not. Despite the lack of individual labels, the goal of LLP is to accurately infer a classification function assigning labels to items. We note that in the standard formulation of LLP, the bag of an item b_i is not an input to the classification function.

We assume a general model for data generation. In particular, the $\{(\mathbf{x}_i, y_i, b_i),\ i = 1, \ldots, N\}$ are i.i.d. observations of a random vector $(\mathbf{X}, Y, B)$ with distribution $P_{\mathbf{X},Y,B}(\mathbf{x}, y, b)$. We will be concerned with independence relations between $\mathbf{X}$, Y, and B; we denote independence of random variables P and Q as $P \perp\!\!\!\perp Q$, and independence of P and Q given W as $P \perp\!\!\!\perp Q \,|\, W$.

We note that our generative model allows for any logically possible dependence relationships between $\mathbf{X}$, Y, and B. However, in what follows, we often assume that $Y \not\perp\!\!\!\perp \mathbf{X}$, as independence of $\mathbf{X}$ and Y would make the learning problem uninteresting.

2.2 Taxonomizing LLP

Our starting point is to observe that when bags are brought into the dependence picture, a large number of cases arise. We can consider the variety of possible dependence structures using the following tableau:

$$\begin{array}{lll} \mathbf{X} \perp\!\!\!\perp Y? & \mathbf{X} \perp\!\!\!\perp B? & Y \perp\!\!\!\perp B? \\ \mathbf{X} \perp\!\!\!\perp Y \,|\, B? & \mathbf{X} \perp\!\!\!\perp B \,|\, Y? & Y \perp\!\!\!\perp B \,|\, \mathbf{X}? \end{array}$$

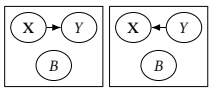
One may answer these six dependence questions in 64 possible ways. However, not all 64 are logically consistent; as it happens, there are 18 logically consistent cases, which can be derived from basic properties of conditional dependence [8].[1] Of the 18 consistent cases, in 8 cases we have that $\mathbf{X}$ is independent of Y. As mentioned above, we consider cases in which $\mathbf{X} \perp\!\!\!\perp Y$ as uninteresting from a learning perspective and so leave them outside the scope of our subsequent analysis.

The remaining 10 cases represent consistent and potentially interesting settings in which LLP problems may arise. To aid in interpreting these cases, we seek generative models that can give rise to the dependence structures. For this purpose, we turn to the framework of directed graphical models (DGMs) [14]. A directed graphical model expresses how a joint distribution may be factored in terms of marginal and conditional probabilities. As such, a DGM involving $\mathbf{X}$, Y, and B implies a particular dependence structure among the variables and can be used as a tool for understanding how particular dependence structures may arise.

As an example, consider the tableau:

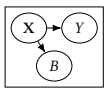
$$\begin{array}{lll} \mathbf{X} \not\perp\!\!\!\perp Y & \mathbf{X} \perp\!\!\!\perp B & Y \perp\!\!\!\perp B \\ \mathbf{X} \not\perp\!\!\!\perp Y \,|\, B & \mathbf{X} \perp\!\!\!\perp B \,|\, Y & Y \perp\!\!\!\perp B \,|\, \mathbf{X} \end{array}$$

This dependence structure arises in two possible DGMs:



Using these DGMs, it is easy to understand this case as the setting in which items are assigned to bags at random. The fact that the two DGMs are different illustrates that different generative processes may give rise to the same dependence structure, and so aids in analysis.

There exist dependence structures that are not implied by any DGM. To see the reason for this, consider the DGM:



One implication of this DGM is that, in general, $Y \not\perp\!\!\!\perp B$. However it is possible for a DGM with this structure to result in $Y \perp\!\!\!\perp B$ in special cases. For example, if Y and B are based on orthogonal features of $\mathbf{X}$, or depend on orthogonal regions of the support of $\mathbf{X}$, then Y will be independent of B.[2] But because this is a special case, we must in general conclude that $Y \not\perp\!\!\!\perp B$. Hence the following dependence structure, while logically consistent and therefore possible, represents a special case that is not expected to occur in practice:

$$\begin{array}{lll} \mathbf{X} \not\perp\!\!\!\perp Y & \mathbf{X} \not\perp\!\!\!\perp B & Y \perp\!\!\!\perp B \\ \mathbf{X} \not\perp\!\!\!\perp Y \,|\, B & \mathbf{X} \not\perp\!\!\!\perp B \,|\, Y & Y \perp\!\!\!\perp B \,|\, \mathbf{X} \end{array}$$

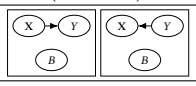
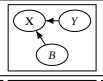
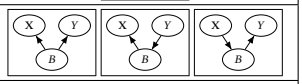
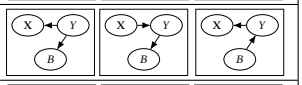
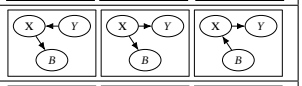
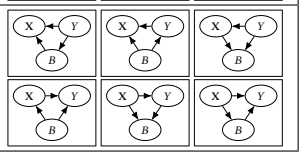
Direct enumeration shows that there are 25 possible directed graphical models involving $\mathbf{X}$, Y, and B. These 25 DGMs imply (in their general forms) 7 out of the 10 logically consistent dependence structures.

Putting all the facts above together, we can form a complete taxonomy of LLP problems. This taxonomy is shown in Table 1. The Table shows all the logically consistent dependence structures that may arise in LLP problems, which we call *LLP variants*. For interpretive purposes, it additionally exhibits any directed graphical models that can give rise to each variant. The variants that can arise

[1] Two facts serve to reduce the 64 combinations to the 18 logically possible: (a) $X \perp\!\!\!\perp Y \,|\, Z,\ X \perp\!\!\!\perp Z \,|\, Y \rightarrow X \perp\!\!\!\perp Y,\ X \perp\!\!\!\perp Z$ (as long as joint distributions are everywhere positive) and (b) $X \perp\!\!\!\perp Y \,|\, Z,\ X \not\perp\!\!\!\perp Z \,|\, Y \rightarrow X \not\perp\!\!\!\perp Z$. Derivations are given in the Appendix.

[2] See, eg., https://terrytao.wordpress.com/2014/06/05/when-is-correlation-transitive/

Table 1: Taxonomy of LLP Variants.

Dependence Structure	DGMs
Uninteresting $\mathbf{X} \perp\!\!\!\perp Y$ (8 Cases)	6 (not shown)
Naive LLP $\mathbf{X} \not\perp\!\!\!\perp Y$ $\quad \mathbf{X} \perp\!\!\!\perp B$ $\quad Y \perp\!\!\!\perp B$ $\mathbf{X} \not\perp\!\!\!\perp Y \mid B$ $\quad \mathbf{X} \perp\!\!\!\perp B \mid Y$ $\quad Y \perp\!\!\!\perp B \mid \mathbf{X}$	
$\mathbf{X} \not\perp\!\!\!\perp Y$ $\quad \mathbf{X} \perp\!\!\!\perp B$ $\quad Y \perp\!\!\!\perp B$ $\mathbf{X} \not\perp\!\!\!\perp Y \mid B$ $\quad \mathbf{X} \not\perp\!\!\!\perp B \mid Y$ $\quad Y \not\perp\!\!\!\perp B \mid \mathbf{X}$	
$\mathbf{X} \not\perp\!\!\!\perp Y$ $\quad \mathbf{X} \perp\!\!\!\perp B$ $\quad Y \not\perp\!\!\!\perp B$ $\mathbf{X} \not\perp\!\!\!\perp Y \mid B$ $\quad \mathbf{X} \perp\!\!\!\perp B \mid Y$ $\quad Y \not\perp\!\!\!\perp B \mid \mathbf{X}$	
Label Collider $\mathbf{X} \not\perp\!\!\!\perp Y$ $\quad \mathbf{X} \perp\!\!\!\perp B$ $\quad Y \not\perp\!\!\!\perp B$ $\mathbf{X} \not\perp\!\!\!\perp Y \mid B$ $\quad \mathbf{X} \not\perp\!\!\!\perp B \mid Y$ $\quad Y \not\perp\!\!\!\perp B \mid \mathbf{X}$	
$\mathbf{X} \not\perp\!\!\!\perp Y$ $\quad \mathbf{X} \not\perp\!\!\!\perp B$ $\quad Y \perp\!\!\!\perp B$ $\mathbf{X} \not\perp\!\!\!\perp Y \mid B$ $\quad \mathbf{X} \not\perp\!\!\!\perp B \mid Y$ $\quad Y \perp\!\!\!\perp B \mid \mathbf{X}$	
Feature Collider $\mathbf{X} \not\perp\!\!\!\perp Y$ $\quad \mathbf{X} \not\perp\!\!\!\perp B$ $\quad Y \perp\!\!\!\perp B$ $\mathbf{X} \not\perp\!\!\!\perp Y \mid B$ $\quad \mathbf{X} \not\perp\!\!\!\perp B \mid Y$ $\quad Y \not\perp\!\!\!\perp B \mid \mathbf{X}$	
Cross-Bag LLP $\mathbf{X} \not\perp\!\!\!\perp Y$ $\quad \mathbf{X} \not\perp\!\!\!\perp B$ $\quad Y \not\perp\!\!\!\perp B$ $\mathbf{X} \perp\!\!\!\perp Y \mid B$ $\quad \mathbf{X} \not\perp\!\!\!\perp B \mid Y$ $\quad Y \not\perp\!\!\!\perp B \mid \mathbf{X}$	
Simple LLP $\mathbf{X} \not\perp\!\!\!\perp Y$ $\quad \mathbf{X} \not\perp\!\!\!\perp B$ $\quad Y \not\perp\!\!\!\perp B$ $\mathbf{X} \not\perp\!\!\!\perp Y \mid B$ $\quad \mathbf{X} \perp\!\!\!\perp B \mid Y$ $\quad Y \not\perp\!\!\!\perp B \mid \mathbf{X}$	
Intermediate LLP $\mathbf{X} \not\perp\!\!\!\perp Y$ $\quad \mathbf{X} \not\perp\!\!\!\perp B$ $\quad Y \not\perp\!\!\!\perp B$ $\mathbf{X} \not\perp\!\!\!\perp Y \mid B$ $\quad \mathbf{X} \not\perp\!\!\!\perp B \mid Y$ $\quad Y \perp\!\!\!\perp B \mid \mathbf{X}$	
Hard LLP $\mathbf{X} \not\perp\!\!\!\perp Y$ $\quad \mathbf{X} \not\perp\!\!\!\perp B$ $\quad Y \not\perp\!\!\!\perp B$ $\mathbf{X} \not\perp\!\!\!\perp Y \mid B$ $\quad \mathbf{X} \not\perp\!\!\!\perp B \mid Y$ $\quad Y \not\perp\!\!\!\perp B \mid \mathbf{X}$	

from a directed graphical model (in the most general interpretation) are given names that we use for discussion in what follows. As mentioned in the previous paragraphs, the cases where no DGM (in its most general interpretation) implies a particular dependence structure are special cases that are not expected to occur often in practice; as a result we do not name them nor treat them further in the following. For example, the special case discussed above appears as the sixth row in the Table.

3 PRIOR WORK

We now examine prior work on LLP in light of the problem taxonomy laid out in the previous section. We make two main observations. First, we observe that prior work has rarely considered the existence or nature of dependence relationships between $\mathbf{X}$, Y, and B. However, much prior work has involved settings in which non-trivial dependence relationships are nonetheless present in experimental data. Second, we observe that prior work has rarely discussed specifics of model selection for the LLP problem. In those cases where model selection procedures are clearly specified, we find that those procedures seem to be based on critical, unstated assumptions about the dependence structure of the problem. These two observations are important because, as we will show in subsequent sections, the dependence structure of the LLP variant at hand has a significant impact on the relative success of different model selection procedures.

3.1 Data Used in Prior Studies

We can distinguish three classes of data used in prior studies. First, *synthetic data* is data generated manually or by sampling standard distributions. Data features $\mathbf{X}$, labels Y, and assignments to bags B are manually determined. Second, *semi-synthetic* data is real-world data (typically a standard machine learning dataset) that starts with pre-existing data items $(\mathbf{X}, Y)$ and assigns them to bags B via some synthetic, deterministic, or probabilistic process. Third, *organic* datasets are data in which data items and bag proportions both arise through the measurement of some real-world properties. In the organic case, true labels Y may not be known. For example, a typical organic dataset consists of individual-level demographic data (for features $\mathbf{X}$) combined with opinion polling which defines bags B (geographic regions) and proportions $\mathbf{p}$ (poll results). Note that in the papers reviewed below, there is rarely, if ever, any discussion of dependence or correlation between data items, labels, and bags.

Synthetic. A few studies use purely synthetic datasets; these are usually small, and are usually used along with other kinds of datasets [21, 24, 26, 28, 45]. Bags and proportions in these datasets are defined by the authors in order to highlight what they would like to show. As such, a wide range of dependence structures exists in these datasets.

Semi-synthetic. The vast majority of studies use semi-synthetic data. These approaches start with pre-existing data without bag assignments; often, these are from the UCI[3] or LibSVM[4] collections [4–6, 12, 24, 28, 29, 31, 32, 34–36, 38, 40, 45, 47, 48], other pre-existing collections [1, 19, 22, 26, 27, 30, 35, 44], or from image classification datasets [3, 4, 9, 16, 18–21, 28, 39–41, 43, 46, 47]. By assigning bags to these pre-existing items a classification dataset is turned into an LLP problem instance.

The most common approach to create bags from pre-existing data is to partition the data items into bags. It is also common that the sizes of bags in a dataset constructed this way are all equal or nearly equal. This is how bags are constructed in [3–6, 9, 16, 18–21, 27–32, 34, 36, 38–41, 43–48]. Often assignments are made to bags so as to achieve a particular set of bag proportions, e.g. as in [26, 27]. Partitioning data creates dependence between B and $(\mathbf{X}, Y)$. When partitions are purely random, the variation of $P(\mathbf{X}, Y)$ across bags may be small; but in many cases the partitions are chosen to achieve some properties (eg, a particular proportion profile $\mathbf{p}$) in which case there is likely a strong dependence between $\mathbf{X}$ and B, and between Y and B.

An exception to the partition approach is taken in [47], which generates bags in a manner that makes the instances conditionally independent given the bag. In this approach, a feature is used as the bag attribute, but items are placed into bags by sampling with replacement from the set of items having a specific feature value. However, dependence is still likely to exist between bags and items, and bags and labels.

[3] https://archive.ics.uci.edu/ml/datasets.php
[4] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

Another approach used to create bags from a pre-existing dataset is to choose a feature as a bag, and then remove this feature from the dataset [4, 22, 24, 47]. Alternatively, the clustering algorithm k-means was used to assign items to bags in [24, 35, 43]. In both cases we expect a strong dependence between $\mathbf{X}$ and B, and between Y and B.

As can be seen, in most or all of the semi-synthetic datasets, we expect that $\mathbf{X} \not\perp B$ and $Y \not\perp B$. It follows that the problem variants represented lie in the bottom four rows of Table 1, and in many cases will fall into the 'Hard' problem variant.

Organic. Finally, some papers used datasets that fit in the LLP setup directly. The authors in [7] studied the candidate preference in the US using Web browsing activity data. Furthermore, [13] studied the embryo implantation prediction problem using a dataset collected by the Unit of Assisted Reproduction of the Hospital Donostia, in Spain. Organic datasets are also standard in the ecological inference literature, eg, [10]. A more complicated process that also relies on empirically observed proportions is described in [2].

In all these cases, data is equipped *a priori* with bags, i.e., these are natural instances of the LLP problem. As mentioned in [47], in such cases we expect that $(\mathbf{X}, Y)$ will not be independent of B.

3.2 Prior Approaches to LLP Model Selection

As we will show below, different LLP variants suggest differences in model selection strategies. However, prior work does not in general discuss model selection in detail.

First, a number of papers do not describe their model selection methods at all [2, 39], perform no model selection [7, 21, 41], make an analysis of the relationship between performance and hyperparameters [16, 18, 19, 22], or state only that they used 'grid search' [3, 26, 27], or 'cross-validation' [46, 47]. A majority of papers simply state that they use 'k-fold' cross validation [5, 6, 24, 28–30, 32, 34–36, 38, 40, 44, 45, 48]. However, as we discuss in § 5, in the LLP setting it is not obvious what is meant by 'k-fold' cross validation, and there are a number of qualitatively different approaches, yielding different results, that fall under the general rubric of 'k-fold' cross validation.

As mentioned above, the only previous study that compares model selection methods for LLP is [12]. That paper considered a number of strategies for holding out validation data during model selection. The authors in [12] identified the best performing approach, in which full bags are allocated among folds to maintain as well as possible the overall proportion of positive instances in the entire dataset ($P(Y = 1)$). They show that this outperforms a number of other methods [13, 43] that involve holding out whole bags as validation data. Accordingly, we compare against this approach as state of the art in our experiments, where we term it the *full bag k-fold* model selection strategy.

In summary, we conclude that the vast majority of prior LLP studies, in addition to not considering data dependence structure, also do not provide detail on the model selection strategies used; and those that do provide detail form validation data using full bags. In the next section, we will look at the dependence structures of LLP variants in detail and note the implications of variants for model selection and for learning strategies.

4 CHARACTERISTICS OF VARIANTS

In this section we examine LLP variants and discuss how their properties impact learning approaches. For purposes of discussion, in each case we assume that a problem instance consists of data that has been drawn from a distribution $P_{\mathbf{X},Y,B}$ that follows the dependence structure of the variant. When we make reference to the Bayes optimal classifier for a given bag ℓ, we mean

$$C_\ell^{\text{Bayes}}(x) \triangleq \arg\max_{r \in \{0,1\}} P(Y = r \mid X = x, B = \ell).$$

In what follows, due to space constraints, we do not discuss all the variants listed in Table 1. Instead, we focus on the four variants that have strong implications in terms of learning: *Naive*, *Simple*, *Intermediate*, and *Hard*. These are the variants that we explore in our experimental results (§ 6). We provide a brief discussion of the other variants in the Supplemental Material (SM).

4.1 Naive LLP

In a Naive LLP problem instance (Table 1, 2nd row), the bag of an item does not depend on the item itself, its label, or any function of these two. A number of implications follow. First, the Bayes optimal classifier for a given bag does not depend on the bag. Second, the proportions p_ℓ across all bags ℓ should be similar (i.e., variations in p_ℓ should not be statistically significant). Finally, given the last two points, a model selection strategy that trains the model in one set of bags and validates using other bags can be appropriate for selecting proper hyperparameters. That is, full-bag strategies such as used in [12, 13, 43] could be effective model selection methods.

For example, many prior studies use Naive problem instances [3–6, 9, 16, 18–21, 27–32, 34, 36, 38–41, 43–48], in which items are randomly assigned to bags. Using random assignment, we expect bags (B) to be independent of both features ($\mathbf{X}$) and labels (Y).

4.2 Simple LLP

As can be seen in the DGMs (Table 1, 9th row), the Simple LLP case can be thought of as a non-independent assignment of labels to bags, in which items are correlated with bags, but only through the labels Y. For example, among the positive instances there is no correlation between bag and item, and likewise for the negative instances.

This has implications for learning and model selection. Unlike the Naive variant, the proportions p_ℓ may vary significantly across bags, allowing for additional information that can be used in the learning process. Additionally, methods that use full bags in model selection may be ineffective, as the Bayes optimal classifier may be affected by bag proportions. To see that, consider, for instance, a case where p_ℓ is close to 1, then the best choice for C_ℓ^{Bayes} may be assigning label 1 to every instance in the bag ℓ regardless of the instances' feature values.

An example of Simple LLP occurs when items are assigned to bags so as to achieve specific proportions in each bag, without any other consideration. This strategy is used, for example, in [26, 27].

4.3 Intermediate LLP

In Intermediate LLP (Table 1, 10th row), items are associated with bags in a non-independent fashion. In this variant, the correlation

between bags and labels is only through the assignment of items to bags. As such, different bags will 'see' significantly different samples of $\mathbf{X}$ – ie, $P(\mathbf{X} \mid B = b_\ell) \neq P(\mathbf{X})$. This implies that models trained on data taken from a subset of bags may not generalize well to the entire distribution of $\mathbf{X}$.

As a result, Intermediate LLP is the second case we have considered so far in which model selection procedures that train on whole bags are undesirable.

An example of an Intermediate LLP instances occurs when items are assigned to bags based on features ($\mathbf{X}$), without any other consideration. As an example, [24] uses k-means clusters as bags, creating a dependence relationship between the features ($\mathbf{X}$) and the bags (B).

4.4 Hard LLP

Like Intermediate LLP, in Hard LLP (Table 1, 11$^{\text{th}}$ row) item-bag associations are not independent. As a result, models trained on data from a subset of bags may not generalize well to the entire distribution of $\mathbf{X}$.

More generally, Hard LLP represents the only case (other than Label Collider) in which labels are correlated directly with both $\mathbf{X}$ and B. This variant therefore implies that the Bayes optimal classifier for items in one bag is different than the Bayes optimal classifier for items in another bag.

Thus Hard LLP is challenging because, as discussed in § 2.1, the standard LLP problem seeks to find a single hypothesis that accurately assigns a label to an item without knowledge of the bag of the item. However, in the Hard LLP setting, the best that can be hoped is to learn a function that interpolates in some sense between the various classifiers that are optimal for each bag. Put another way, in the Hard setting it would be desirable to change the problem definition to include the bag of an item as a feature input to the classification function – but of course this would no longer be (standard) LLP.

For example, in a political science study like [7], we expect that the distribution of types of voters ($\mathbf{X}$) to regions (B) will not be random, and furthermore that the type of voter ($\mathbf{X}$) that supports a given candidate (Y) can vary across regions (B). these relationships result in an instance of the Hard LLP variant.

4.5 Identifying the LLP Variant

The taxonomy we detail here can be used in a number of ways.

As detailed in § 3.1, the great majority of algorithmic studies of the LLP problem make use of datasets in which the labels $\{y_i\}$ are known. Access to labels is important in validating an LLP learning method, since accurate prediction of individual labels is the goal in LLP (§ 2.1). In those settings, the dependence relations between $\mathbf{X}$, Y and B can be empirically tested, and the LLP variant at hand can be determined. We believe it is important to perform such tests in studies that use datasets to validate new LLP methods, because of the implications of dependence structure in data on the information available to the LLP method. Further, as we show in the next section, the choice of strategies for model selection depends on the LLP variant at hand.

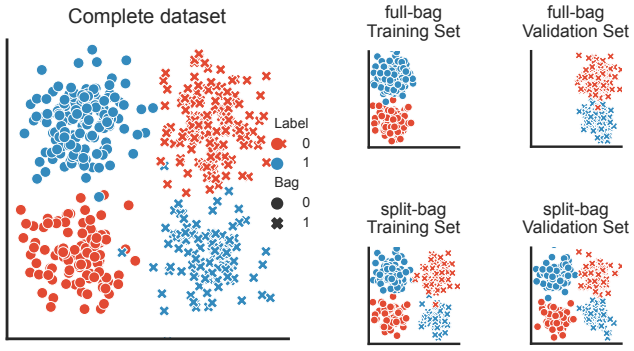


Figure 1: Left: An instance of the LLP problem with $N = 500$ (250 elements per bag) and $\mathbf{p} = [0.6, 0.4]^T$. Right: Comparison of possible training and validation sets when using *full-bag* and *split-bag* strategies.

In studies using 'organic' datasets (see § 3.1), where labels are not available, we still believe that the LLP taxonomy may be fruitfully used to understand the nature of an LLP variant at hand. It is common to reason about potential correlations in data in advance of analyzing the data (eg, in political science and economics). Further, one may posit the existence of causal relationships between variables that lead to dependence structures among $\mathbf{X}$, B, and Y. For example, as noted above, in the case of a study using polling, one would expect certain dependence relationships to exist between voters ($\mathbf{X}$), regions (B) and preference for a given candidate (Y).

5 NEW APPROACHES TO LLP MODEL SELECTION

To motivate the new approaches we develop for model selection, in Figure 1 (left) we show an instance of a difficult LLP problem. In this example, we have 500 items in two bags, and the dependence structure is that of Hard LLP (as detailed in § 4.3).

Standard model selection methods dictate that we must obtain several pairs of training and validation sets from the data, and use these sets to perform hyperparameter tuning, thus avoiding overfitting [33]. However, in the LLP problem setting, the question of how to construct training and validation sets is more complex than in traditional supervised learning, because of the absence of labels on items. As discussed in §3, prior work (when it specifies a strategy) makes use of *full-bag* strategies. Full-bag strategies retain some bags to train the classifier and use the others as validation data. Figure 1 (right) shows why a full-bag strategy may not be a good idea, in general. More specifically, the Bayes optimal classifier for bag 0 will not be suitable for bag 1.

Based on the above arguments, we propose a new family of strategies for model selection in LLP, which we term *split-bag*. These strategies allow information from all bags to be present in both training and validation sets, and we argue they are more appropriate for general LLP settings. An example of this strategy is shown in Figure 1 (right) as well.

To make our suggestion concrete, we start from a general strategy for model selection in LLP, (details in Algorithm 1), which

Algorithm 1: `LLP-Model-Selection`

Data: Integer K; LLP problem instance, i.e., $D = \{(\mathbf{x}_i, b_i)\}_{i=1}^{N}$ and $\mathbf{p}$; LLP algorithm $\mathcal{M}$, which depends on hyperparameters vector θ; and set of candidate hyperparameters vectors Θ
Result: Best hyperparameters vector

1 $(D_{\text{Tr},1}, D_{\text{V},1}), \ldots, (D_{\text{Tr},K}, D_{\text{V},K}) \leftarrow$ `Train-Validation`(D, K)
2 **for** $\theta \in \Theta$ **do**
3 **for** $k \leftarrow 1$ **to** K **do**
4 $C_{k,\theta} \leftarrow$Train a model with $\mathcal{M}$ on $D_{\text{Tr},k}$, $\mathbf{p}$, and θ
5 $\hat{\mathbf{y}}_{k,\theta} \leftarrow$ Use $C_{k,\theta}$ to predict labels from instances in $D_{k,\text{V}}$
6 $\hat{\mathbf{p}}_{k,\theta} \leftarrow$ Compute the predicted bags proportions from $\hat{\mathbf{y}}_{k,\theta}$
7 $\text{Error}_{k,\theta} \leftarrow$ `Error`$(\hat{\mathbf{p}}_{k,\theta}, \mathbf{p})$
8 $\text{Error}_{\theta} \leftarrow \frac{1}{K}\sum_{k=1}^{K} \text{Error}_{k,\theta}$
9 **return** $\arg\min_{\theta\in\Theta} \text{Error}_{\theta}$

Algorithm 2: `Split-Bag-Shuffle`

Data: The set of instances and associated bags, $D = \{(\mathbf{x}_i, b_i)\}_{i=1}^{N}$; an integer K; and $\beta \in (0, 1)$, the ratio between validation and traning sets sizes

1 **for** $\ell \leftarrow 1$ **to** L **do**
2 $B_\ell \leftarrow \{(\mathbf{x}, b) \in D \mid b = \ell\}$
3 **for** $k \leftarrow 1$ **to** K **do**
4 **for** $\ell \leftarrow 1$ **to** L **do**
5 $B_{\ell,k} \leftarrow$ random sample, **without replacement**, of size $\beta|B_\ell|$ from B_ℓ
6 $B'_{\ell,k} \leftarrow B_\ell \setminus B_{\ell,k}$
7 $D_{\text{Tr},k} \leftarrow \bigcup_{1\le\ell\le L} B'_{\ell,k}$
8 $D_{\text{V},k} \leftarrow \bigcup_{1\le\ell\le L} B_{\ell,k}$
9 **return** $(D_{\text{Tr},1}, D_{\text{V},1}), \ldots, (D_{\text{Tr},K}, D_{\text{V},K})$

follows the strategy for model selection commonly used in standard supervised learning. The first step is to build from the input data K (training, validation) pairs of sets. The next step is, for each combination of (training, validation) pairs and candidate hyperparameter settings, to train a classifier using the training data, use the classifier to predict the labels from instances in the validation set, and compute the proportion of positive predictions in each bag. The next step is to compare the estimated and real proportions in order to measure the classifier's error. Finally, errors are averaged across all training-validation pairs to determine the hyperparameters yielding the best model.

It is important to note that splitting bags between training and validation sets can introduce sampling error. Because per-item labels are not available, the label proportions in split bags are not known exactly. Sampling error is a concern when the proportions in the training or validation sets differ significantly from those of the original data enough to jeopardize the hyperparameter selection process. Sampling error probability is reduced as bag sizes grow, (lessening with the square root of the number of items per bag) but can not be completely avoided. We return to the question of sampling error at the end of the paper.

The split-bag strategy raises two important questions. In order to fully specify model selection methods, one must first determine how to form training and validation sets from the data (Line 1 in Algorithm 1), and second determine how to compare the *real* and *predicted* bags proportions (Line 7 in Algorithm 1).

We address the first question in the next three sections, where we present a family of *split-bag* strategies for generating training and validation sets: *split-bag K-fold*, *split-bag shuffle*, and *split-bag bootstrap*. With regard to the second question, in this work, we use the *mean absolute error*, i.e., given vectors $\mathbf{p}$ and $\mathbf{p}'$,

$$\text{Error}(\mathbf{p}', \mathbf{p}) = \frac{1}{L}\sum_{\ell=1}^{L} |p'_\ell - p_\ell|. \tag{1}$$

Equation (1) exposes a fundamental challenge when approaching LLP. Since instances' labels are unknown, cross-validation techniques commonly used for model selection in supervised learning cannot be used. This is because one cannot compute standard evaluation metrics (e.g., accuracy and F-score) over the validation sets.

Therefore, we follow related work and propose to evaluate models based on the bags proportions that their predictions yield. Even though this approach seems a natural choice, it is important to recognize its limitations. In particular, the correct label assignment will match the bags proportions, but there may be (many) incorrect label assignments that nearly match the proportions as well.

Finally, we recognize that there may be more sophisticated ways to compare the vectors $\mathbf{p}$ and $\mathbf{p}'$ than by using Equation (1). For instance, one may consider for each bag its size, as well as how far from $\frac{1}{2}$ its proportion is. We leave this investigation for future work.

5.1 Split-bag K-fold

The idea of split-bag K-fold is similar to the usual K-fold for supervised learning [33]. However, instead of partitioning the whole dataset, the partition is conducted over each bag.

Let B_ℓ be the subset of instances belonging to bag ℓ, i.e., $B_\ell = \{(\mathbf{x}, b) \in D \mid b = \ell\}$ and let's assume that $|B_\ell| \ge K$. First, split-bag K-fold partitions each bag, B_ℓ, into K equal-sized subsets (prior shuffle is recommended), $B_{\ell,1}, \ldots, B_{\ell,K}$. Then

$$D_{\text{Tr},k} = \bigcup_{\substack{1\le\ell\le L \\ j\neq k}} B_{\ell,j} \quad \text{and} \quad D_{\text{V},k} = \bigcup_{1\le\ell\le L} B_{\ell,k},$$

for $k = 1, \ldots, K$. Hence, a usual K-fold is applied to each bag, and then, the bag-folds are aggregated in order to create the training and validation sets.

5.2 Split-bag Shuffle

Similarly to split-bag K-fold, we propose split-bag shuffle inspired by the usual shuffle-split method used for model selection in supervised learning. The procedure is described in Algorithm 2. Again, the main idea is to apply the usual shuffle-split in a per-bag fashion in order to create training and validation sets. However, instead of partitioning each bag and rotating over each partition (i.e., fold), we draw a random sample from the bag with fraction β of elements for validation, and we leave the remaining for training.

5.3 Split-Bag Bootstrap

Our third method is a natural extension of the split-bag shuffle method. The idea is to draw random samples *with replacement* in order to create the per-bag training and validation subsets. More specifically, Line 5 of Algorithm 2 is modified so that the random sample is drawn with replacement, and in Line 6, we define $B'_{\ell,k}$ as a random sample, with replacement as well, of size $(1-\beta)|B_\ell|$ from B_ℓ.

The motivation for incorporating bootstrap into the LLP model selection problem is to counter sampling effects on the per-bag proportions of positives instances in the training and validation sets. Consider, for instance, the split-bag shuffle procedure presented in Algorithm 2. If the proportion of positive instances in $B_{\ell,k}$ deviates significantly from p_ℓ, the same will happen to $B'_{\ell,k}$, since the composition of the latter depends on the former. Considering that such an effect can propagate to $D_{\mathrm{Tr},k}$ and $D_{\mathrm{V},k}$ (for some values of k), one may end up with training and validation sets having per-bag proportions far from $\mathbf{p}$. This deviation will incorrectly inform the learning and model selection algorithms with vector $\mathbf{p}$, making the overall LLP problem even harder.

When we use bootstrap in Lines 5 and 6 of Algorithm 2, we ensure that the composition of $B'_{\ell,k}$ is independent of $B_{\ell,k}$. Hence, the chances of having both sets (i.e., training and validation) with bag proportions far from $\mathbf{p}$ are lessened, which is particularly important for LLP problem instances having (many) small bags.

6 EVALUATION

In this section, we provide empirical evidence showing the practical importance of understanding LLP variants. More specifically, our experiments show that previously proposed model selection methods can work well for some LLP variants (e.g., Naive LLP), but they are unable to adapt to more complicated scenarios (e.g., Hard LLP).[5]

6.1 Datasets

In contrast to most previous work (as discussed in §3) in this study we pay close attention to the dependence relationships among bags, items, and labels in our test data. In particular, we generate test data in which the dependence structure follows either the Naive, Simple, Intermediate, or Hard LLP variants. To do so, we start from a *base* dataset which is a standard classification dataset. We then allocate items to bags according to a procedure that induces the desired dependence relation as given in Table 1. Details of the bag generation process are given in SM §A.4. Note that when we generate different LLP variants for a given base dataset, we keep the number of bags, sizes of bags, and bag label proportions approximately equal across the different variants.

We consider three classes of base datasets: *tabular*, *object images*, and *digit images*. The *tabular* class consists of datasets designated Adult, Default Credit Card, Covertype, and Census KDD (sources and properties shown in Table 2 in SM § A.3). The *object images* class consists of pairs of object comparisons from the CIFAR-10 repository; we consider 12 pairs of comparisons such as 'Airplane' vs 'Automobile.' Likewise the *digit images* class consists of 12 pairs of handwritten digit comparisons from the MNIST repository. From each base dataset we generate four variants (Naive, Simple, Intermediate and Hard) resulting in a total of $4\times 28 = 112$ datasets which we refer to as *test cases*.

6.2 Experimental Setup

To explore whether algorithmic complexity plays a role in the choice of model selection methods, we choose three algorithms from the literature that have differing numbers of hyperparameters. First, we consider the algorithm used in [7] which is an EM-like method based on logistic regression, which we denote EM/LR; it has only a single hyperparameter C. Next, we consider alter-∝SVM with Linear Kernel from [47], which has two hyperparameters C and C_p. Finally, we consider the Laplacian Mean Map ($\mathrm{LMM}_{G,s}$) algorithm from [24], which we denote LMM; it has three hyperparameters σ, λ, and γ.

For each of the 112 test cases, we apply the three algorithms; and for each of those combinations we perform model selection using each of the four approaches: *split-bag k-fold*, *split-bag bootstrap*, *split-bag shuffle*, and the baseline approach *full-bag k-fold*. However, we found alter-∝SVM has a very high runtime (an average of 15 hours for a single run on our smallest test cases using a 28 core CPU) and hence we were unable to run it on large test cases; as a result we ran alter-∝SVM on a subset of 24 test cases, also omitting the *split-bag k-fold* model selection strategy for alter-∝SVM. Each combination of (test case, algorithm, model selection method) constitutes a *experiment*. Our results span 968 experiments of which 720 experiments employed one of our new model selection approaches, and the remaining 248 experiments used the full-bag approach as a baseline for comparison.

For each experiment, we randomly hold out 25% of the data for testing and use the other 75% for training and validation. After removing the testing data, we recompute the bag proportions (p_ℓ) for the remaining data. In this way, we emulate the real-world setting in which one has the true bag proportions for the input data. The remaining data is then divided into training and validation data according to the model selection strategy. Note however, that after splitting into training and validation sets, bag proportions in those sets are no longer known exactly. Each such experiment is run 30 times with different random seeds, and we ensure apples-to-apples comparisons by ensuring the train/test splits are identical across the four model selection approaches.

For EM/LR, the parameter C is tuned from the set $\{10^{-2}, 10^{-1}, \cdots, 10^{3}\}$. For the alter-∝SVM parameters, C and C_p, we use grid search over values in $\{10^{-2}, \ldots, 10^{3}\}$. For LMM, the parameters are tuned using grid search over $\lambda \in \{0, 10^0, 10^1, 10^2\}, \gamma \in \{10^{-2}, 10^{-1}, 10^0\}$, and $\sigma \in \{2^{-2}, 2^{-1}, 2^0\}$. Once hyperparameter(s) are chosen, we refit the model using the whole training data making use of the best hyperparameter(s). Then, we compute the accuracy of the resulting classifier on the test set.

The above strategy results in 27,600 trained models, each of which yields a value for accuracy on the test data (10 replications were performed for alter-∝SVM). We compare each new model selection strategy against the baseline, and test statistical significance

[5] All datasets and code to reproduce our results are available at https://github.com/gaabrielfranco/llp-variants-kdd

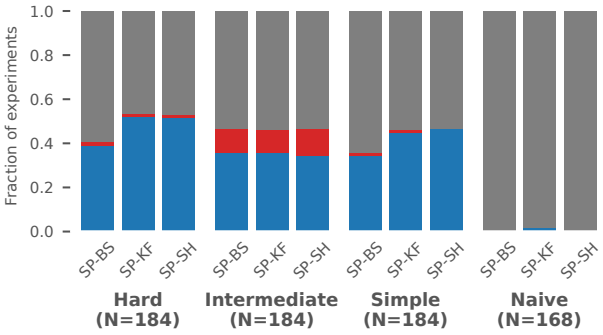


Figure 2: Fraction of experiments comparing new vs baseline, for LLP variant and model selection method. Blue denotes statistically significant superiority of the new method over the baseline. Red denotes cases where the baseline is significantly better than the new method. Gray denotes cases where the new method and the baseline are statistically indistinguishable. SP-BS = *split-bag bootstrap*; SP-KF = *split-bag k-fold*; SP-SH = *split-bag shuffle*.

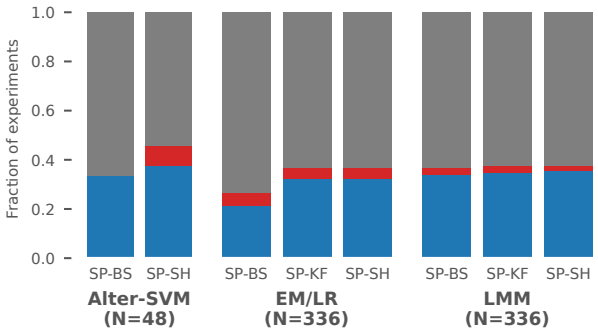


Figure 3: Fraction of experiments comparing new (split-bag) vs baseline (full-bag), for LLP algorithm and model selection method. Same color and labeling scheme as Figure 2.

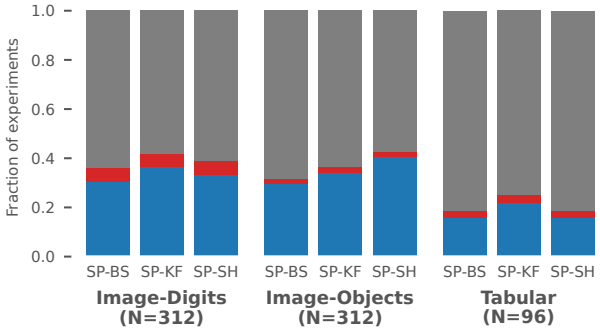


Figure 4: Fraction of experiments comparing new (split-bag) vs baseline (full-bag), for dataset type and model selection method. Same color and labeling scheme as Figure 2.

of the difference in means between each set of 30 replications, using a t-test at the 0.05 significance level.

6.3 Results

Our first set of results demonstrate that (a) *our new approaches to LLP model selection significantly outperform the state of the art*; and (b) *the relative performance of model selection approaches depends critically on the LLP variant at hand.* Both of these results are evident from Figure 2. The figure shows, for each model selection approach, the fraction of experiments in which the new approach outperforms the state of the art baseline to a statistically significant degree (in blue).

With respect to (a), the figure shows that in almost every case where a significant difference exists, the new model selection approach outperforms the baseline. In fact, *in 90% of the cases where a significant difference exists, the new model selection approach is superior to the baseline.*

With respect to (b), the figure shows that there is dramatic difference between LLP variants Hard, Intermediate, and Simple on one hand, and the Naive variant on the other hand. In the case of the Naive variant, neither full-bag (baseline) nor split-bag (new) model selection methods are superior; for this variant, one could reasonably use either model selection method. However, for the cases that inspired the development of the new model selection methods (Simple, Intermediate and Hard, as explained in § 4), it is clear that any of the split-bag methods are preferable to the baseline.

Next, we ask whether the superiority of the new methods is dependent on the complexity (ie, number of hyperparameters) of the algorithm used. Figure 3 shows our results broken down by algorithm. The figure shows that whether one is using a model with numerous hyperparameters (LMM) or few hyperparameters (EM/LR) split-bag model selection is generally superior, in cases where there is a statistically significant difference.

We next ask whether the nature of the base datasets used affects our conclusions. For example, images may involve more complex decision boundaries than tabular datasets. Since full-bag model selection methods train on only portions of the feature space (for the Hard and Intermediate variants) it may be more important to use split-bag methods on datasets with more complex decision boundaries. Figure 4 shows some evidence of this effect. It shows that split-bag methods are more likely to outperform full-bag methods on the test cases from the image classes (*digits images* and *object images*) than from the tabular classes of data.

Furthermore, to demonstrate that our model selection strategies indeed select different hyperparameters than previous (full-bag) strategies, we look at a single example: the CIFAR-10-Grey (Airplane vs. Automobile) Hard LLP dataset. In Figure 5 we show heatmaps of the hyperparameters chosen for each strategy, with the hyperparameters chosen by *full-bag k-fold* on the left of the figure. The remaining three columns show that the split-bag strategies generally choose hyperparameter values that are quite different from the values chosen by the full-bag strategy.

Finally, to show the effect sizes behind the results in Figures 2 to 4 we show in Figure 6 the distribution of the difference between split-bag and full-bag accuracies (for cases in which the difference was statistically significant). The figure shows that when split-bag

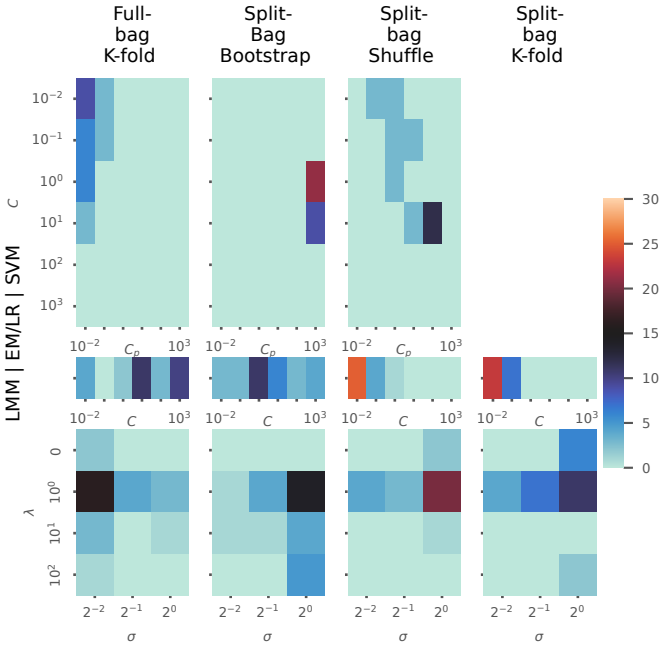


Figure 5: Frequency, over 30 executions, of selected hyperparameters for the CIFAR-10-Grey (Airplane vs. Automobile) Hard LLP dataset.

methods outperform full-bag methods, the magnitude of improvement is typically large compared the alternative case.

Discussion. In conclusion we make a number of observations about model selection in LLP. First, we note that discrepancy between the expected bag proportions p_ℓ and the actual bag proportions can arise in split-bag methods due to sampling. For small bag sizes, this can be an important effect that does not arise in full-bag strategies (because they are not sampled). However, relative error in sampling-induced deviation from expected proportions shrinks as $1/\sqrt{n}$ and we find that for bag sizes in the low thousands or more, sampling error does not in general prevent split-bag strategies from dominating full-bag approaches. A further means for mitigating this effect comes in choosing how much of each bag to allocate to training versus validation data. In Algorithm 2, this is controlled by parameter β. We find that setting β near 1/2 (which is what we use our experiments) also improves overall performance since it reduces sampling error in p_ℓ. In fact, there are opportunities for algorithm improvement in the proper selection of β. We leave this problem for future work.

The picture that emerges is one in which large datasets (leading to large bags) with complex dependencies (e.g. Intermediate or Hard variants) and complex decision boundaries (making out-of-bag generalization difficult) present the greatest opportunity for performance improvement through the use of split-bag model selection strategies. Specifically, for all non-Naive variants, there is benefit from using split-bag strategies over the *full-bag k-fold* in the hyperparameter selection process. Among split-bag strategies, *split-bag k-fold* and *split-bag shuffle* generally outperform *split-bag bootstrap*. However, further investigation is needed to decide the best split-bag strategy in individual cases, and we leave this problem for future work.

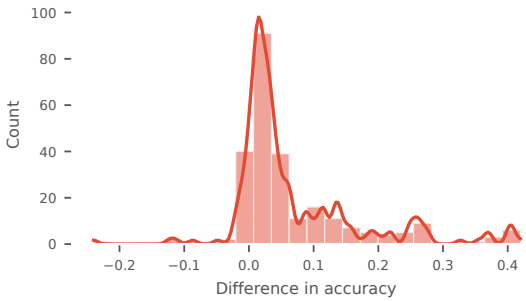


Figure 6: Distribution of differences between split-bag and full-bag accuracies, for cases showing a significant difference.

7 CONCLUSIONS

In this paper, we have looked at the LLP problem from a perspective that has not been taken in previous work, by putting a spotlight on the dependence relationships that can exist between items, bags, and labels. As a result, our first contribution is a taxonomy of LLP variants, a characterization of the main properties of each variant, and an analysis of the implications regarding hardness of learning and model selection.

Our taxonomy can be interesting for any researcher studying or applying LLP methods. However, the taxonomy may be of special interest for those proposing new techniques. Ideally, new algorithms, for the LLP problem itself and for model selection strategies in LLP, should work across a range of LLP variants, or at least, the readers should be aware of untested variants or situations where the algorithms are known to perform badly. For these controlled scenarios, statistical tests for independence and conditional are available (e.g., [37]), and they may aid researchers in the task of choosing the correct variant in Table 1.

Reviewing prior work in LLP, we note that in many cases, the dependence structure of the problem instances considered, and strategies used for model selection, are unclear. This fact and the implications of our taxonomy motivate our second contribution, a family of LLP-driven model selection techniques. We conduct extensive experiments, and our results show the superiority of our new methods for model selection over the state of the art. Over nearly a thousand distinct experiments, spanning a range of LLP algorithms, problem variants, and data types, our methods outperform the state of the art in 90% of the statistically significant cases. We analyze the advantages of our new methods, and conclude that for challenging problems (large, complex data with complex dependence structures) our new methods are particularly well suited.

ACKNOWLEDGMENTS

Part of this work was done while GF was at the Federal University of Viçosa (UFV). GC thanks the support of the following Brazilian funding agencies: FAPESP/MCTI/CGI.br (#2020/05182-3) and FAPES (#1026/2022).

REFERENCES

[1] Ehsan Mohammady Ardehaly and Aron Culotta. 2016. Domain Adaptation for Learning from Label Proportions Using Self-Training. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA). 3670–3676.

[2] Ehsan Mohammady Ardehaly and Aron Culotta. 2017. Co-training for demographic classification using deep learning from label proportions. In *2017 IEEE International Conference on Data Mining Workshops*. IEEE, 1017–1024.

[3] Denis Baručić and Jan Kybic. 2021. Fast learning from label proportions with small bags. *arXiv preprint arXiv:2110.03426* (2021).

[4] Jing Chai and Ivor W Tsang. 2021. Learning With Label Proportions by Incorporating Unmarked Data. *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[5] Zhensong Chen, Wei Chen, and Yong Shi. 2020. Ensemble learning with label proportions for bankruptcy prediction. *Expert Systems with Applications* 146 (2020), 113155.

[6] Zhensong Chen, Zhiquan Qi, Bo Wang, Limeng Cui, Fan Meng, and Yong Shi. 2017. Learning with label proportions based on nonparallel support vector machines. *Knowledge-Based Systems* 119 (2017), 126–141.

[7] Giovanni Comarela, Ramakrishnan Durairajan, Paul Barford, Dino Christenson, and Mark Crovella. 2018. Assessing Candidate Preference through Web Browsing History. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2018), 158–167. https://doi.org/10.1145/3219819.3219884

[8] A. P. Dawid. 1979. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society: Series B (Methodological)* 41, 1 (1979), 1–15. https://doi.org/10.1111/j.2517-6161.1979.tb01052.x

[9] Gabriel Dulac-Arnold, Neil Zeghidour, Marco Cuturi, Lucas Beyer, and Jean-Philippe Vert. 2019. Deep multi-class learning from label proportions. *arXiv preprint arXiv:1905.12909* (2019).

[10] Seth R. Flaxman, Yu-Xiang Wang, and Alexander J. Smola. 2015. Who Supported Obama in 2012? Ecological Inference through Distribution Regression. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 289–298. https://doi.org/10.1145/2783258.2783300

[11] Maxime Gasse and Alex Aussem. 2016. Identifying the irreducible disjoint factors of a multivariate probability distribution. In *Probabilistic Graphical Models*. Lugano, Switzerland, 183–194.

[12] Jerónimo Hernández-González. 2019. A framework for evaluation in learning from label proportions. *Progress in Artificial Intelligence* 8, 3 (2019), 359–373.

[13] Jerónimo Hernández-González, Inaki Inza, Lorena Crisol-Ortíz, María A Guembe, María J Iñarra, and Jose A Lozano. 2018. Fitting the data from embryo implantation prediction: Learning from label proportions. *Statistical methods in medical research* 27, 4 (2018), 1056–1066.

[14] Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

[15] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[16] Laura Elena Cué La Rosa and Dário Augusto Borges Oliveira. 2022. Learning from Label Proportions with Prototypical Contrastive Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 2153–2161.

[17] Yann LeCun, Corinna Cortes, and CJ Burges. 2010. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist* 2 (2010).

[18] Jiabin Liu, Zhiquan Qi, Bo Wang, YingJie Tian, and Yong Shi. 2022. SELF-LLP: Self-supervised learning from label proportions with self-ensemble. *Pattern Recognition* 129 (2022), 108767.

[19] Jiabin Liu, Bo Wang, Hanyuan Hang, Huadong Wang, Zhiquan Qi, Yingjie Tian, and Yong Shi. 2022. Llp-gan: a gan-based algorithm for learning from label proportions. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[20] Jiabin Liu, Bo Wang, Zhiquan Qi, YingJie Tian, and Yong Shi. 2019. Learning from Label Proportions with Generative Adversarial Networks. *Advances in Neural Information Processing Systems* 32 (2019).

[21] Jiabin Liu, Bo Wang, Xin Shen, Zhiquan Qi, and Yingjie Tian. 2021. Two-stage Training for Learning from Label Proportions. *arXiv preprint arXiv:2105.10635* (2021).

[22] Jay Nandy, Rishi Saket, Prateek Jain, Jatin Chauhan, Balaraman Ravindran, and Aravindan Raghuveer. 2022. Domain-Agnostic Contrastive Representations for Learning from Label Proportions. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 1542–1551.

[23] H. James Norton and George Divine. 2015. Simpson's paradox – and how to avoid it. *Significance* 12, 4 (2015), 40–43. https://doi.org/10.1111/j.1740-9713.2015.00844.x arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1740-9713.2015.00844.x

[24] Giorgio Patrini, Richard Nock, Paul Rivera, and Tiberio Caetano. 2014. (Almost) no label no cry. *Advances in Neural Information Processing Systems* 27 (2014), 190–198.

[25] Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[26] Rafael Poyiadzi, Raul Santos-Rodriguez, and Niall Twomey. 2018. Label propagation for learning with label proportions. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 1–6.

[27] Rafael Poyiadzi, Raul Santos-Rodriguez, and Niall Twomey. 2019. Active learning with label proportions. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3097–3101.

[28] Zhiquan Qi, Fan Meng, Yingjie Tian, Lingfeng Niu, Yong Shi, and Peng Zhang. 2018. Adaboost-LLP: A Boosting Method for Learning With Label Proportions. *IEEE Transactions on Neural Networks and Learning Systems* 29, 8 (2018), 3548–3559. https://doi.org/10.1109/TNNLS.2017.2727065

[29] Zhiquan Qi, Bo Wang, Fan Meng, and Lingfeng Niu. 2016. Learning with label proportions via NPSVM. *IEEE transactions on cybernetics* 47, 10 (2016), 3293–3305.

[30] Yaxing Qian, Qiang Tong, and Bo Wang. 2019. Multi-Class Learning from Label Proportions for Bank Customer Classification. *Procedia Computer Science* 162 (2019), 421–428.

[31] Yue Qiu, Mingjie Yan, and Zhensong Chen. 2021. Active learning from label proportions via pSVM. *Neurocomputing* 464 (2021), 227–241.

[32] Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. 2009. Estimating labels from label proportions. *Journal of Machine Learning Research* 10, 10 (2009).

[33] Sebastian Raschka. 2018. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *CoRR* abs/1811.12808 (2018). arXiv:1811.12808 http://arxiv.org/abs/1811.12808

[34] Stefan Rueping. 2010. SVM classifier estimation from group probabilities. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*. 911–918.

[35] Rishi Saket, Aravindan Raghuveer, and Balaraman Ravindran. 2022. On Combining Bags to Better Learn from Label Proportions. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 5913–5927.

[36] Clayton Scott and Jianxin Zhang. 2020. Learning from Label Proportions: A Mutual Contamination Framework. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 22256–22267. https://proceedings.neurips.cc/paper/2020/file/fcde14913c766cf307c75059e0e89af5-Paper.pdf

[37] Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. 2017. Model-powered conditional independence test. *Advances in neural information processing systems* 30 (2017).

[38] Yong Shi, Limeng Cui, Zhensong Chen, and Zhiquan Qi. 2019. Learning from label proportions with pinball loss. *International Journal of Machine Learning and Cybernetics* 10, 1 (2019), 187–205.

[39] Yong Shi, Jiabin Liu, and Zhiquan Qi. 2018. Inverse convolutional neural networks for learning from label proportions. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 643–646.

[40] Yong Shi, Jiabin Liu, Zhiquan Qi, and Bo Wang. 2018. Learning from label proportions on high-dimensional data. *Neural Networks* 103 (2018), 9–18.

[41] Yong Shi, Jiabin Liu, Bo Wang, Zhiquan Qi, and YingJie Tian. 2020. Deep learning from label proportions with labeled samples. *Neural Networks* 128 (2020), 73–81.

[42] Marco Stolpe and Katharina Morik. 2011. Learning from Label Proportions by Optimizing Cluster Model Selection. In *Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part III* (Athens, Greece) *(ECML PKDD'11)*. Springer-Verlag, Berlin, Heidelberg, 349–364.

[43] Kuen-Han Tsai and Hsuan-Tien Lin. 2020. Learning from label proportions with consistency regularization. In *Asian Conference on Machine Learning*. PMLR, 513–528.

[44] Yanshan Xiao, HuaiPei Wang, and Bo Liu. 2020. A new transfer learning-based method for label proportions problem. *Information Sciences* 541 (2020), 391–408.

[45] Felix Yu, Dong Liu, Sanjiv Kumar, Jebara Tony, and Shih-Fu Chang. 2013. \proptoSVM for Learning with Label Proportions. In *International Conference on Machine Learning*. PMLR, 504–512.

[46] Felix X Yu, Liangliang Cao, Michele Merler, Noel Codella, Tao Chen, John R Smith, and Shih-Fu Chang. 2014. Modeling attributes from category-attribute proportions. In *Proceedings of the 22nd ACM international conference on Multimedia*. 977–980.

[47] Felix X Yu, Krzysztof Choromanski, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. 2014. On learning from label proportions. *arXiv:1402.5902* (2014).

[48] Fan Zhang, Jiabin Liu, Bo Wang, Zhiquan Qi, and Yong Shi. 2019. A Fast Algorithm for Multi-Class Learning from Label Proportions. *Electronics* 8, 6 (2019), 609.

A SUPPLEMENTAL MATERIAL

A.1 Derivations

The set of allowable dependence relations in Table 1 is derived using inference rules. Here we show the two rules used to eliminate those dependence relations that are logically inconsistent.

In the derivations below, we use the following definition of conditional independence:

$$X \perp\!\!\!\perp Y \mid Z \triangleq p(x, y, z)p(z) = p(x, z)p(y, z)$$

as discussed in [11].

Here we derive the two rules that allow us to narrow the set of cases to just those in Table 1. We (re)derive these rules in order to identify any necessary conditions on the positivity of distributions.

(1) $X \perp\!\!\!\perp Y \mid Z,\ X \perp\!\!\!\perp Z \mid Y \rightarrow X \perp\!\!\!\perp Y,\ X \perp\!\!\!\perp Z$
Proof.
(a) $X \perp\!\!\!\perp Y \mid Z \rightarrow p(x, y, z)p(z) = p(x, z)p(y, z)$.
(b) $X \perp\!\!\!\perp Z \mid Y \rightarrow p(x, y, z)p(y) = p(x, y)p(y, z)$.
(c) (a) & (b) $\rightarrow$ for any $x, y, z,\ p(x, z)/p(z) = p(x, y)/p(y)$ if $p(y, z) \neq 0$.
(d) Hence $p(x \mid z) = p(x \mid y) = p(x)$.
(e) Hence $X \perp\!\!\!\perp Z, X \perp\!\!\!\perp Y$.
Thus we can conclude that the rule stated above follows, *unless* there are some values (y, z) where $p(y, z) = 0$ and, for those values it happens that $p(x \mid y) \neq p(x \mid z)$.

(2) $X \perp\!\!\!\perp Y \mid Z,\ X \not\perp\!\!\!\perp Z \mid Y \rightarrow X \not\perp\!\!\!\perp Z$
Proof.
(a) $X \perp\!\!\!\perp Y \mid Z \rightarrow p(x, y, z)p(z) = p(x, z)p(y, z)$.
(b) $X \not\perp\!\!\!\perp Z \mid Y \rightarrow \exists\, y_0, z_1, z_2$ s.t. $p(x \mid y_0, z_1) \neq p(x \mid y_0, z_2)$. By definition $p(y_0, z_1) \neq 0$ and $p(y_0, z_2) \neq 0$.
(c) Hence $\exists\ y_0, z_1, z_2$ s.t. $p(x, y_0, z_1)p(y_0, z_2) \neq p(x, y_0, z_2)p(y_0, z_1)$.
(d) (a) & (c) $\rightarrow$ $[p(x, z_1)/p(z_1)]p(y_0, z_1)p(y_0, z_2) \neq [p(x, z_2)/p(z_2)]p(y_0, z_2)p(y_0, z_1)$
(e) Hence $\exists\, z_1, z_2$ s.t. $p(x, z_1)/p(z_1) \neq p(x, z_2)/p(z_2)$. By definition $p(z_1) \neq 0$ and $p(z_2) \neq 0$.
(f) Hence $\exists\, z_1, z_2$ s.t. $p(x \mid z_1) \neq p(x \mid z_2)$.
(g) Hence $X \not\perp\!\!\!\perp Z$.
We conclude that no special conditions on the positivity of distributions are required for this rule.

A.2 Other LLP Variants

A.2.1 Colliders. The LLP variants presented in the 5th and 7th rows of Table 1 are examples of colliders, in which B is independent of either $\mathbf{X}$ (Label Collider) or Y (Feature Collider). However, given the third variable in each case, the independence disappears due to the explaining-away effect [25]. While we consider these cases interesting from a theoretical perspective, practical scenarios giving rise to these cases seem rare. Further, the associated dependence structure only arises as the result of a single DGM in each case. As a result, we leave detailed consideration of the implications of these variants for future work.

A.2.2 Cross-bag LLP. As illustrated by the DGMs (Table 1, 8th row), cross-bag cases occur when the correlation between $\mathbf{X}$ and Y is mediated by B. In other words, a correlation between $\mathbf{X}$ and Y exists across bags, but not within any given bag. This is an example of Simpson's paradox: no correlation between $\mathbf{X}$ and Y exists in the subgroups (bags), but when groups are combined, a correlation exists [23].

These considerations show that in Cross-bag LLP, it is important to use items coming from multiple bags as input to the learning step of model selection, as only in that way can the relationship between $\mathbf{X}$ and Y be observed.

A.3 Datasets

We used six classification datasets. For multi-class classification data, we consider some pairs of classes as a dataset. We made the Census-KDD dataset balanced with the same positive and negative instances. We also converted the CIFAR-10 images to greyscale in order to reduce the number of features. Table 2 presents the details and pointers to the data sources.

Table 2: Summary of datasets

Dataset	N	Features	Proportions of 1's	Source
Adult	48842	179	0.24	UCI
Default-Credit-Card	30000	23	0.22	UCI
Covertype (Classes 1 and 2)	495141	54	0.43	UCI
Census-KDD (Balanced)	37136	506	0.50	UCI
MNIST-Digits (Pairs of classes)	Ranges from 13138 to 14780	784	Ranges from 0.49 to 0.54	[17]
CIFAR-10 Grey (Pairs of classes)	12000	1024	0.50	[15]

Following [45], for each dataset, the features were scaled to $[-1, 1]$. Moreover, categorical features were one-hot encoded.

A.4 Bag generation process

Since we are using classification datasets in our experiments, we need to have a procedure to generate bags given a classification dataset. For each dataset, we generated four LLP datasets: a Naive LLP dataset (see Section 4.1), a Simple LLP dataset (see Section 4.2), an Intermediate LLP Dataset (see Section 4.3), and a Hard LLP dataset (see Section 4.4). We assume that $\mathbf{X}$ and Y are correlated in all datasets since they are classification datasets.

Our generation process is based on a clustering assignment. We used k-means to create a clustering assignment for the dataset. Then, given a base dataset, we just use the clusters as bags to generate an Intermediate dataset.

Let $\mathbf{p}^t$ be the vector of proportions from the Intermediate dataset, and $\mathbf{B}^t$ the vector of bag sizes from the Intermediate dataset ($\mathbf{p}^t_i$, $\mathbf{B}^t_i$ denotes the proportion and size of the i-th bag of the Intermediate dataset respectively). Our goal is to have similar bag sizes across all variants and similar proportions in Simple, Intermediate, and Hard variants.

For the Naive variant, we randomly assigned bags to items. For each bag i, the probability of an item being assigned to this bag is proportional to $\mathbf{B}^t_i$. Then, we have a correspondence between the clustering assignment sizes and bag sizes for the Naive LLP datasets.

For the Simple variant, each item has a probability of being assigned to a certain bag based on its label, i.e., these probabilities are different from when the item label is positive or negative. These probabilities are computed in a way such the dataset will have approximately proportions $\mathbf{p}^t$ and bag sizes $\mathbf{B}^t$.

For the Hard variant, each item has a probability of being assigned to a certain bag based on its label and its cluster id, i.e., these probabilities are different for each label and cluster combination.

These probabilities are also computed in a way such the dataset will have approximately proportions $\mathbf{p}^t$ and bag sizes $\mathbf{B}^t$.

More precise algorithms to generate LLP datasets from base datasets and formal tests to verify their respective variant will be addressed in future work.