LEARNING FROM LABEL PROPORTIONS WITH CONSISTENCY REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The problem of learning from label proportions (LLP) involves training classifiers with weak labels on bags of instances, rather than strong labels on individual instances. The weak labels only contain the label proportions of each bag. The LLP problem is important for many practical applications that only allow label proportions to be collected because of data privacy or annotation costs, and has recently received lots of research attention. Most existing works focus on extending supervised learning models to solve the LLP problem, but the weak learning nature makes it hard to further improve LLP performance with a supervised angle. In this paper, we take a different angle from semi-supervised learning. In particular, we propose a novel model inspired by consistency regularization, a popular concept in semi-supervised learning that encourages the model to produce a decision boundary that better describes the data manifold. With the introduction of consistency regularization, we further extend our study to non-uniform bag-generation and validation-based parameter-selection procedures that better match practical needs. Experiments not only justify that LLP with consistency regularization achieves superior performance, but also demonstrate the practical usability of the proposed procedures.

1 INTRODUCTION

In traditional supervised learning, a classifier is trained on a dataset where each data point is associated with a class label. However, label annotation is expensive or difficult to obtain. Sometimes, only label proportions about groups of data points are provided to the classifier. This problem setting is known as learning from label proportions (LLP). In LLP, each *bag* is associated with a proportion of class labels. A classifier is then trained on these bags and proportions in order to predict class labels for unseen data points. Recently, LLP has attracted much attention among researchers because its problem setting occurs in many real-life scenarios. For example, the census data, medical databases, and US presidential election results are all provided in the form of groups due to privacy issues (Patrini et al., 2014; Sun et al., 2017). Other LLP applications include fraud detection (Rueping, 2010), object recognition (Kuck & de Freitas, 2012), video event detection (Lai et al., 2014), and ice-water classification (Li & Taylor, 2015).

The challenge in LLP is to train models without direct instance-level label supervision. To overcome this issue, prior work seek to estimate either the individual label (Yu et al., 2013; Dulac-Arnold et al., 2019) or the mean of each bag (Quadrianto et al., 2009; Rueping, 2010; Patrini et al., 2014) by the label proportions. However, the methodology behind developing these models do not portray LLP situations that occur in real life. First, these models can be improved by considering methods that can better leverage label scarcity. Second, these models assume that bags of data are randomly generated, which is not the case for many applications. Consider the population census where the data are collected on region, age, or occupation with varying group sizes. Third, training these models requires a validation set with labeled data. It would be more ideal if model selection only relied on label proportions.

To resolve the above problems, our main contributions are listed as follows:

- We first apply a semi-supervised learning technique, consistency regularization, to the multi-class LLP problem. Consistency regularization considers an auxiliary loss term to

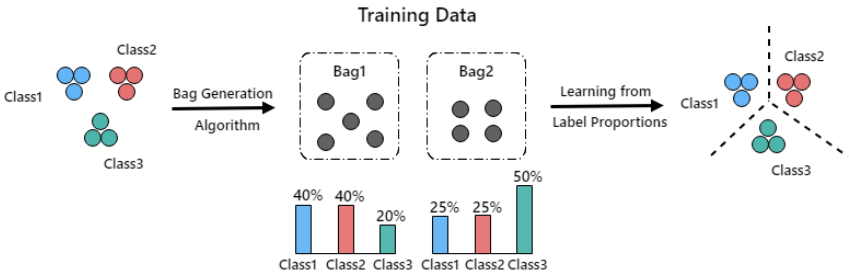

Figure 1: Learning from label proportions.

enforce network predictions to be consistent when its input is perturbed. By exploiting the unlabeled instances, our method captures the latent structure of data and obtains the SOTA performance on three benchmark datasets.

- We develop a new bag generation algorithm – the K-means bag generation, where training data are grouped by attribute similarity. Using this setup can help train models that are more applicable to actual LLP scenarios.
- We show that it is possible to select models with a validation set consisting of only bags and associated label proportions. The experiments demonstrate correlation between bag-level validation error and instance-level test error. This potentially reduces the need of a validation set with instance-level labels.

2 PRELIMINARY

2.1 LEARNING FROM LABEL PROPORTIONS

We consider a multi-class classification problem in this paper. Let $\boldsymbol{x}_i \in \mathbb{R}^D$ be a feature vector of i-th example, $y_i \in \{1, \ldots, L\}$ be a class label of i-th example, and $\boldsymbol{e}^{(i)}$ be a standard basis vector $[0, \ldots, 1, \ldots, 0]$ with 1 at i-th position. Traditionally, we have a training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{n=1}^N$ with each instance i.i.d. drawn from an unknown probability distribution. However, in the setting of learning from label proportions (LLP), each individual label y_i is hidden from the training data. On the other hand, the training data are aggregated by an unknown generation algorithm and come in the form of M bags $B_1, \ldots, B_M$. The m-th bag contains a set of instances $\mathbb{X}_m$ and a proportion label $\boldsymbol{p}_m$, where

$$\boldsymbol{p}_m = \frac{1}{|\mathbb{X}_m|} \sum_{i:\boldsymbol{x}_i \in \mathbb{X}_m} \boldsymbol{e}^{(y_i)}, \quad \bigcup_{m=1}^{M} \mathbb{X}_m = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}.$$

We do not require each subset to be disjoint. Also, each bag may have different size. The task of the LLP is to learn an individual-level classifier $f_\theta : \mathbb{R}^D \to \mathbb{R}^L$ to predict the unknown label y for a new instance $\boldsymbol{x}$. Figure 1 illustrates the setting of learning from label proportions in the multi-class classification.

2.2 PROPORTION LOSS

The feasibility of the binary LLP setting has been theoretically justified by Yu et al. (2014). Specifically, Yu et al. (2014) propose the framework of *Empirical Proportion Risk Minimization* (EPRM) proving that the LLP problem is PAC-learnable by assuming that bags are i.i.d sampled from an unknown probability distribution. The EPRM framework provides a generalization bound of label proportions and guarantees to probably learn an approximately correct proportion predictor as the number of label proportions increases. Furthermore, the authors prove that the instance label error can be bounded by the bag proportion error. That is, a decent bag proportion predictor guarantees a decent instance label predictor.

Based on the profound theoretical analysis, a vast of LLP approaches learn an instance-level classifier by directly minimizing the proportion loss without acquiring the individual labels. To be

more precise, given a bag $B = (\mathbb{X}, \boldsymbol{p})$, an instance-level classifier f_θ and a divergence function $d : \mathbb{R}^K \times \mathbb{R}^K \rightarrow \mathbb{R}$, the proportion loss penalizes the difference between the real label proportions $\boldsymbol{p}_m$ and the estimated proportions $\hat{\boldsymbol{p}} = \frac{1}{|\mathbb{X}|} \sum_{\boldsymbol{x} \in \mathbb{X}} f_\theta(\boldsymbol{x})$, which is an average of the instance predictions within a bag. Thus, the proportion loss $\mathcal{L}_{prop}$ can be defined as follows:

$$\mathcal{L}_{prop}(\theta) = d(\boldsymbol{p}, \hat{\boldsymbol{p}}).$$

Typically, the common used divergence function is L^1 or L^2 function in prior work (Musicant et al., 2007; Yu et al., 2013). Ardehaly & Culotta (2017) and Dulac-Arnold et al. (2019), on the other hand, consider the cross-entropy function for the multi-class LLP problem.

2.3 CONSISTENCY REGULARIZATION

Since collecting labeled data is expensive and time-consuming, the semi-supervised learning approaches aim to leverage a large amount of unlabeled data to alleviate the needs for labeled data. There are many semi-supervised learning methods, such as pseudo-labeling (Lee, 2013), generative approaches (Kingma et al., 2014), and consistency-based methods (Laine & Aila, 2016; Miyato et al., 2018; Tarvainen & Valpola, 2017). Consistency-based approaches encourage the network to produce consistent output probabilities between unlabeled data and the perturbed examples. These methods rely on the smoothness assumption (Chapelle et al., 2009): if two data points x_i and x_j are close, then so should be the corresponding output distributions y_i and y_j. Then, the consistency-based approaches can enforce the decision boundary to traverse through the low-density region. More precisely, given a perturbed input $\hat{\boldsymbol{x}}$ taken from the input $\boldsymbol{x}$, consistency regularization penalizes the distinction of model predictions between $f_\theta(\boldsymbol{x})$ and $f_\theta(\hat{\boldsymbol{x}})$ by a divergence function $d : \mathbb{R}^K \times \mathbb{R}^K \rightarrow \mathbb{R}$. The consistency loss can be written as follows:

$$\mathcal{L}_{cons}(\theta) = d(f_\theta(\boldsymbol{x}), f_\theta(\hat{\boldsymbol{x}})).$$

Modern consistency-based methods are different in the way that perturbed examples are generated for the unlabeled data. Laine & Aila (2016) introduce the Π-Model using the additive Gaussian noise for perturbed examples and choose the L^2 error as the divergence function. However, a drawback to Π-Model is that the consistency target obtained from the stochastic network is unstable since the network changes rapidly during training. To address this problem, Temporal Ensembling (Laine & Aila, 2016) takes the exponential moving average of the network predictions as the consistency target. Mean Teacher (Tarvainen & Valpola, 2017), on the other hand, proposes averaging the model parameter values instead of predictions. Overall, the Mean Teacher approach significantly improves the quality of target predictions and the empirical results on semi-supervised benchmarks.

Instead of applying the stochastic perturbations to the inputs, Virtual Adversarial Training or VAT (Miyato et al., 2018) computes the adversarial examples $\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{r}_{adv}$, where

$$\boldsymbol{r}_{adv} = \underset{\boldsymbol{r}:||\boldsymbol{r}||_2 \leq \epsilon}{\arg\max}\, D_{\mathrm{KL}}(f_\theta(\boldsymbol{x}) \| f_\theta(\boldsymbol{x} + \boldsymbol{r})).$$

That is, the VAT approach attempts to generate a perturbation which most likely causes the model to misclassify the input in an adversarial direction. In comparison to the stochastic perturbation, the VAT approach has demonstrated the efficiency and effectiveness in the semi-supervised learning problem. Other consistency-based works, including Interpolation Consistency Training (Verma et al., 2019) and MixMatch (Berthelot et al., 2019), utilize the methodology of data augmentation to generate the perturbed examples for unlabeled instances.

3 LLP WITH CONSISTENCY REGULARIZATION

With regards to label scarcity, the LLP scenario is similar to the semi-supervised learning problem. In the semi-supervised learning setting, only a small portion of training examples is labeled. On the other hand, in the LLP scenario, we are given the label proportions without individual labels. The similarity between these settings is that a large amount of training examples is unlabeled. To address this challenge, semi-supervised approaches seek to exploit the unlabeled examples to further capture the latent structure of data.

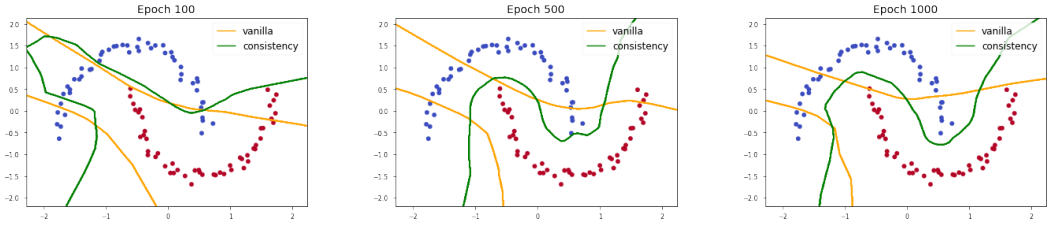

Figure 2: In the toy example, we generate 5 bags, each of which contains 20 examples uniformly sampled from the "two moons" dataset without replacement. The vanilla approach, which simply optimizes the proportion loss, suffers from poor performance as the label information is insufficient. Conversely, the "two moons" can be effectively separated into two clusters by LLP with consistency regularization. Our method captures the underlying structure of data by enforcing the network producing consistent outputs for perturbed examples.

Motivated by these semi-supervised approaches, we combine the idea of leveraging the unlabeled data into the LLP problem. We make the same smoothness assumption and introduce a new concept incorporating consistency regularization with LLP. In particular, we consider the typical cross-entropy loss between real label proportions and estimated label proportions. We define the expected proportion cost J_{prop} as follows:

$$\mathcal{J}_{prop} = \mathbb{E}_{\mathbb{X},\boldsymbol{p}}\left[-\sum_{i=1}^{L} \boldsymbol{p}_i \log \frac{1}{|\mathbb{X}|}\sum_{\boldsymbol{x}\in\mathbb{X}} f_\theta(\boldsymbol{x})_i\right].$$

To encourage learning a decision boundary that better reflects the data manifold, we add an auxiliary consistency loss by leveraging the unlabeled data. More formally, the expected consistency cost $\mathcal{J}_{cons}$ can be written as follows:

$$\mathcal{J}_{cons} = \mathbb{E}_{\mathbb{X},\boldsymbol{p}}\left[\frac{1}{|\mathbb{X}|}\sum_{\boldsymbol{x}\in\mathbb{X}} d(f_\theta(\boldsymbol{x}), f_\theta(\hat{\boldsymbol{x}}))\right],$$

where d is a divergence function, and $\hat{\boldsymbol{x}}$ is a perturbed input of $\boldsymbol{x}$. We can use any consistency-based approach to generate the perturbed examples and to compute the consistency cost. The combinatorial cost $\mathcal{J}$ for LLP is computed as follows:

$$\mathcal{J} = \mathcal{J}_{prop} + \alpha\mathcal{J}_{cons},$$

where α is a coefficient hyperparameter. Following Miyato et al. (2018), we use KL divergence as the consistency cost and adopt VAT to generate the perturbed inputs in most of our experiments.

To understand the intuition behind consistency regularization, we use the stochastic Gaussian noise and adopt the L^2 error as the consistency cost in the toy example. Figure 2 illustrates how the LLP with consistency regularization produces the decision boundary passing through the low-density region by capturing the data manifold. On the other hand, the vanilla approach, which simply optimizes the proportion loss, gets easily stuck at a degenerate solution due to the lack of label information. This shows the advantage of applying consistency regularization to LLP.

4 Experiments

4.1 Datasets

To evaluate the effectiveness of our proposed method, we conduct experiments on three benchmark datasets, including SVHN (Netzer et al., 2011), CIFAR10, and CIFAR100 (Krizhevsky & Hinton, 2009). The SVHN dataset consists of 32x32 RGB digit images with 73,257 examples for training, 26,032 examples for testing, and 531,131 extra training examples that are not used in our experiments. The CIFAR10 and CIFAR100 datasets both consist of 50,000 training examples and 10,000 test examples. Each example is a 32x32 colored natural image, drawn from 10 classes and 100 classes respectively.

4.2 EXPERIMENT SETUP

Implementation details For all experiments in this section, we adopt a Wide Residual Network with depth 28 and width 2 (WRN-28-2) following the standard specification in the paper (Zagoruyko & Komodakis, 2016). We use the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.0003. Additionally, we train models for a maximum of 400 epochs with a learning rate scheduler where the learning rate is reduced by 0.2 once the number of epochs reaches 320. To evaluate the performance of a model in LLP, we split the training data by a bag generation algorithm described in section 4.3 and 4.4. Once completing the bag generation, we compute the corresponding label proportions for each bag by averaging the class labels. To avoid over-fitting, we follow the common practice of data augmentation (He et al., 2016; Lin et al., 2013) padding an image by 4 pixels on each side, taking a random 32x32 crop and randomly flipping the image horizontally with the probability of 0.5 for all benchmarks.

Model selection For a fair comparison, we randomly sample 90% of bags for training and reserve the rest for validation. In this case, there are no individual labels available in the validation set. Therefore, we tune the hyperparameters based on the bag-level L^1 error (hard L^1 error) on the validation set. To be more specific, the bag-level L^1 error for a bag $B = (\mathbb{X}, \boldsymbol{p})$ is defined by

$$Err = ||\boldsymbol{p} - \hat{\boldsymbol{p}}||_1, \quad \hat{\boldsymbol{p}} = \frac{1}{|\mathbb{X}|} \sum_{\boldsymbol{x} \in \mathbb{X}} \boldsymbol{e}^{(i^*)},$$

where $i^* = \arg\max_i f(\boldsymbol{x})_i$ and $\boldsymbol{e}^{(i^*)}$ is a one-hot encoding of the prediction. Lastly, we report the test instance accuracy averaged over the last 10 epochs.

Hyperparameters We compare our method, learning from label proportions with consistency regularization (LLP-CR), to the ROT loss (Dulac-Arnold et al., 2019) and the vanilla approach, which directly minimizes the proportion loss. For the ROT method, we conduct experiments with a hyperparameter of $\alpha \in \{0.1, 0.4, 0.7, 0.9\}$. Following Oliver et al. (2018), we adopt the VAT approach to generate adversarial examples with a perturbation weight ϵ of 1 and 6 for SVHN and CIFAR10 (or CIFAR100) respectively. Furthermore, we measure the consistency loss with the KL divergence and a consistency weight of $\alpha \in \{0.5, 0.1, 0.05, 0.01\}$. In case the model gets stuck in a corrupt situation in the early stage, we adopt the exponential ramp-up function (Laine & Aila, 2016) to gradually increase the consistency weight.

4.3 UNIFORM BAG GENERATION

For convenience, most LLP works validate their proposed methods with the uniform bag generation where the training data are partitioned into bags of the same size. Precisely, we uniformly sample n instances without replacement from the training set, where $n \in \{16, 32, 64, 128, 256\}$ denotes the bag size. We drop the last incomplete bag if the number of training data is indivisible by the bag size.

As shown in Table 1, the LLP with consistency regularization improves the performance on CIFAR10 and CIFAR100 for the LLP scenario with a uniform bag generation in most cases. As for SVHN, since the test accuracy is close to the fully-supervised performance when the bag size is small, there is no clear difference among the three methods.

4.4 K-MEANS BAG GENERATION

In the real-life LLP scenario, data are usually grouped by attribute similarity instead of uniformly sampled at random. For example, the presidential election results (Sun et al., 2017) are gathered by states consisting of varying numbers of votes. To better simulate the realistic situations, we formulate a new bag generation scheme - the K-means bag generation, where examples are aggregated by the correlation. Although those bags generated from the K-means bag generation are dependent on each other, violating the i.i.d assumption, this setting is both challenging and worth-studying.

Since we perform experiments on image datasets, crafting bags based on the selected RGB pixels is meaningless. To group the images by its objects, shapes, or other underlying relations, we adopt the unsupervised reduction technique, PCA algorithm, to project data into a low-dimensional representation space by capturing the latent structure. We use the mini-batch K-means algorithm to

Table 1: Test accuracy with the uniform bag generation.

		Bag Size				
Dataset	Method	16	32	64	128	256
SVHN	vanilla	95.28	95.20	94.41	88.93	12.64
	ROT	95.35	94.84	93.74	**92.29**	**13.14**
	LLP-CR	**95.66**	**95.73**	**94.60**	91.24	11.18
CIFAR10	vanilla	88.77	85.02	70.68	47.48	38.69
	ROT	86.97	77.01	62.93	48.95	40.16
	LLP-CR	**89.30**	**85.41**	**72.49**	**50.78**	**41.62**
CIFAR100	vanilla	58.58	48.09	20.66	5.82	2.82
	ROT	54.16	47.75	**29.38**	7.95	2.63
	LLP-CR	**59.47**	**48.98**	22.84	**9.40**	**3.29**

Table 2: Test accuracy with the K-means bag generation on SVHN.

		K				
Dataset	Method	4576	2288	1144	572	286
SVHN	vanilla	92.07	91.16	92.00	78.70	**47.16**
	LLP-CR	**93.11**	**91.69**	**93.21**	**82.05**	46.38

aggregate similar images based on the feature correlation. We conduct experiments with clusters of $K \in \{3120, 1560, 780, 390, 195\}$ on the CIFAR10 and CIFAR100 dataset. On the other hand, we experiment with $K \in \{4576, 2288, 1144, 572, 286\}$ clusters on the SVHN dataset.

In this section, we do not compare our proposed method to the ROT loss, which needs to iteratively estimate individual labels for each bag. The procedure of the ROT algorithm is time-consuming and cannot be accelerated if bags are of varying sizes. Besides, for the K-means bag generation, there may be some large bags when the value of K is small. Because of the limited computational resource, we take a subsample in each bag if the bag size is larger than the threshold of 256.

The experimental results of the K-means bag generation are shown in Table 2 and Table 3. Although this scenario violates the i.i.d assumption, the results show that it is feasible to learn an instance-level classifier by minimizing the proportion loss. Also, our approach significantly brings benefits for the k-means bag generation scenario on benchmarks. Interestingly, the performance of a model is not well-correlated with the value of K. One possible reason is that we might drop informative bags as we randomly split bags into validation and training.

4.5 Validation metrics

Many modern machine learning models require a wide range of hyperparameter selections about the architecture, optimizer and regularization. However, for the realistic LLP scenario, we have no access to labeled instances during training. It is crucial to tune hyperparameters based on the bag-level validation error. To evaluate the performance at the bag level, we consider four validation metrics: soft L^1 error, hard L^1 error, soft KL divergence, and hard KL divergence. Their definitions are given as follows. First, we define the output probabilities of an instance as the soft prediction and its one-hot encoding as the hard prediction. For each bag, we then compute the estimated label proportions by averaging these soft or hard predictions. Finally, we use the L^1 error or KL divergence to measure the bag-level prediction error.

To investigate the relationship between the instance-level test error and the bag-level validation error, we compute the Pearson correlation coefficient between them on models trained for 400 epochs. The results are shown in Table 4. Surprisingly, we find that the hard L^1 error has a strong positive correlation to test error rate on all benchmarks. This implies that it is feasible to select hyperparam-

Table 3: Test accuracy with the K-means bag generation on CIFAR10 and CIFAR100.

		K				
Dataset	Method	3120	1560	780	390	195
CIFAR10	vanilla	73.93	66.54	44.12	49.85	**39.86**
	LLP-CR	**77.43**	**68.01**	**51.04**	**50.22**	38.27
CIFAR100	vanilla	**38.65**	**22.16**	16.07	**15.47**	7.82
	LLP-CR	37.98	21.90	**15.61**	15.31	**8.13**

Table 4: The Pearson correlation coefficient between the test error rate and the following validation metrics on benchmarks.

	Uniform			K-means		
	SVHN	CIFAR10	CIFAR100	SVHN	CIFAR10	CIFAR100
Hard L^1	0.97	0.81	0.81	0.99	0.75	0.67
Soft L^1	0.83	0.33	-0.50	0.90	0.61	0.26
Hard KL	0.69	-0.18	0.64	0.81	0.10	0.40
Soft KL	0.69	0.89	-0.16	0.71	0.62	0.57

eters with only label proportions in realistic LLP scenarios. Interestingly, our finding is coherent to Yu et al. (2013). Although their and our works both adopt the hard L^1 error for model selection, we focus on the multi-class LLP scenario instead of the binary classification problem they considered. Therefore, we suggest future multi-class LLP works could adopt the hard L^1 metric for model selection.[1]

5 Related work

Kuck & de Freitas (2012) first introduce the LLP scenario and formulate the probabilistic model with the MCMC algorithm to generate consistent label proportions. Several following works (Chen et al., 2006; Musicant et al., 2007) extend the LLP setting to a variety of standard supervised learning algorithms. Without directly inferring instance labels, Quadrianto et al. (2009) propose a Mean Map algorithm with exponential-family parametric models. The algorithm uses empirical mean operators of each bag to solve a convex optimization problem. However, the success of the Mean Map algorithm is based on a strong assumption that the class-conditional distribution of data is independent of bags. To loosen the restriction, Patrini et al. (2014) propose a Laplacian Mean Map algorithm imposing an additional Laplacian regularization. Nevertheless, these Mean Map algorithms suffer from a fundamental drawback: they require the classifier to be a linear model.

Several works tackle the LLP problem from Bayesian perspectives. For example, Fan et al. (2014) propose an RBM-based generative model to estimate the group-conditional likelihood of data. Hernández-González et al. (2013), on the other hand, develop a Bayesian classifier with an EM algorithm. Recently, Sun et al. (2017) propose a graphical model using counting potential to predict instance labels for the US presidential election. Furthermore, other works (Chen et al., 2009; Stolpe & Morik, 2011) adopt a k-means approach to cluster training data by label proportions. While some works (Fan et al., 2014; Sun et al., 2017) claim that they are suitable for large-scale settings, both Bayesian methods and clustering-based algorithms are rather inefficient and computationally expensive when applied to large image datasets.

Another line of work adopts a large-margin framework for the problem of LLP. Stolpe & Morik (2011) propose a variant of support vector regression using the inverse calibration method to estimate the class-conditional probability for bags. On the other hand, Yu et al. (2013) propose a

[1]Nevertheless, we do not suggest using our validation metric for early stopping since the correlation is computed after the model converges.

procedure that alternates between assigning a label to each instance, also known as *pseudo-labeling* in the literature, and fitting an SVM classifier. Motivated by this idea, a number of works (Wang et al., 2015; Qi et al., 2016; Chen et al., 2017) infer individual labels and updated model parameters alternately. One major drawback of SVM-based approaches is that they are tailored for binary classification; they cannot extend to the multi-class classification setting efficiently.

As deep learning has garnered huge success in a number of areas, such as natural language processing, speech recognition, and computer vision, many works leverage the power of neural networks for the LLP problem. Ardehaly & Culotta (2017) are the first to apply deep models to the multi-class LLP setting. Also, Bortsova et al. (2018) propose a deep LLP method learning the extent of emphysema from the proportions of disease tissues. Concurrent to our work, Dulac-Arnold et al. (2019) also considers the multi-class LLP setting with bag-level cross-entropy loss. They introduce a ROT loss that combines two goals: jointly maximizing the probability of instance predictions and minimizing the bag proportion loss.

6 CONCLUSION

In this paper, we first apply a novel semi-supervised learning technique, consistency regularization, to the multi-class LLP problem. Our proposed approach leverages the unlabeled data to learn a decision boundary that better depicts the data manifold. The empirical results validate that our approach obtains better performance than that achieved by existing LLP works. Furthermore, we introduce a non-uniform bag scenario - the K-means bag generation, where training instances are clustered by attribute relationships. This setting simulates more practical LLP situations than the uniform bag generation setting, which is often used in previous works. Lastly, we introduce a bag-level validation metrics, hard L^1 error, for model selection with only label proportions. We empirically show that the bag-level hard L^1 error has a strong correlation to the test classification error. For real-world applicability, we suggest that multi-class LLP methods relying on hyper-parameter tuning could evaluate their methodology based on the bag-level hard L^1 error. We hope that future LLP work can further explore the ideas presented in this paper.

REFERENCES

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

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.

Gerda Bortsova, Florian Dubost, Silas Ørting, Ioannis Katramados, Laurens Hogeweg, Laura Thomsen, Mathilde Wille, and Marleen de Bruijne. Deep learning from label proportions for emphysema quantification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 768–776. Springer, 2018.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

Bee-Chung Chen, Lei Chen, Raghu Ramakrishnan, and David R Musicant. Learning from aggregate views. In *22nd International Conference on Data Engineering (ICDE'06)*, pp. 3–3. IEEE, 2006.

Shuo Chen, Bin Liu, Mingjie Qian, and Changshui Zhang. Kernel k-means based framework for aggregate outputs classification. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 356–361. IEEE, 2009.

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

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

Kai Fan, Hongyi Zhang, Songbai Yan, Liwei Wang, Wensheng Zhang, and Jufu Feng. Learning a generative classifier from label proportions. *Neurocomputing*, 139:47–55, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

Jerónimo Hernández-González, Iñaki Inza, and Jose A Lozano. Learning bayesian network classifiers from label proportions. *Pattern Recognition*, 46(12):3425–3440, 2013.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pp. 3581–3589, 2014.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Hendrik Kuck and Nando de Freitas. Learning about individuals from group statistics. *arXiv preprint arXiv:1207.1393*, 2012.

Kuan-Ting Lai, Felix X Yu, Ming-Syan Chen, and Shih-Fu Chang. Video event detection by inferring temporal instance labels. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 2243–2250, 2014.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, pp. 2, 2013.

Fan Li and Graham Taylor. Alter-cnn: An approach to learning from label proportions with application to ice-water classification. In *Neural Information Processing Systems Workshops (NIPSW) on Learning and privacy with incomplete data and weak supervision*, 2015.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

Takeru Miyato, Shin-ichi Maeda, Shin Ishii, and Masanori Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

David R Musicant, Janara M Christensen, and Jamie F Olson. Supervised learning by training on aggregate outputs. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pp. 252–261. IEEE, 2007.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Ng. Reading digits in natural images with unsupervised feature learning. *Advances in neural information processing systems*, 01 2011.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, pp. 3235–3246, 2018.

Giorgio Patrini, Richard Nock, Paul Rivera, and Tiberio Caetano. (almost) no label no cry. In *Advances in Neural Information Processing Systems*, pp. 190–198, 2014.

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

Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10(Oct):2349–2374, 2009.

Stefan Rueping. Svm classifier estimation from group probabilities. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 911–918, 2010.

Marco Stolpe and Katharina Morik. Learning from label proportions by optimizing cluster model selection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 349–364. Springer, 2011.

Tao Sun, Dan Sheldon, and Brendan OConnor. A probabilistic approach for learning with label proportions applied to the us presidential election. In *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 445–454. IEEE, 2017.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pp. 1195–1204, 2017.

Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *arXiv preprint arXiv:1903.03825*, 2019.

Bo Wang, Zhensong Chen, and Zhiquan Qi. Linear twin svm for learning from label proportions. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pp. 56–59. IEEE, 2015.

Felix X Yu, Dong Liu, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. $\propto$svm for learning with label proportions. *arXiv preprint arXiv:1306.0886*, 2013.

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

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *Procedings of the British Machine Vision Conference 2016*, 2016.