Learnability of the Superset Label Learning Problem

Li-Ping Liu LIULI@EECS.OREGONSTATE.EDU
Thomas G. Dietterich TGD@EECS.OREGONSTATE.EDU
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, Oregon 97331, USA

Abstract

In the Superset Label Learning (SLL) problem, weak supervision is provided in the form of a *superset* of labels that contains the true label. If the classifier predicts a label outside of the superset, it commits a *superset error*. Most existing SLL algorithms learn a multiclass classifier by minimizing the superset error. However, only limited theoretical analysis has been dedicated to this approach. In this paper, we analyze Empirical Risk Minimizing learners that use the superset error as the empirical risk measure. SLL data can arise either in the form of independent instances or as multiple-instance bags. For both scenarios, we give the conditions for ERM learnability and sample complexity for the realizable case.

1. Introduction

In multiclass supervised learning, the task is to learn a classifier that maps an object to one of several candidate classes. When each training example is labeled with one label, many successful multiclass learning methods can solve this problem (Aly, 2005; Mukherjee and Schapire, 2013). In some applications, however, we cannot obtain training examples of this kind. Instead, for each instance we are given a *set* of possible labels. The set is guaranteed to contain the true label as well as one or more *distractor labels*.

Despite these distractor labels, we still wish learn a multiclass classifier that has low error when measured according to the traditional 0/1 misclassification loss. This learning problem has been given several names, including the "multiple label problem", the "partial label problem" and the "superset label learning problem" (Jin & Ghahramani, 2002; Nguyen & Caruana, 2008; Cour et al., 2011; Liu & Dietterich, 2012). In this paper, we adopt the last of these.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

Several learning algorithms for the superset label learning problem have shown good experimental results. Most of these algorithms seek an hypothesis that explicitly minimizes superset errors on the training instances (possibly including a regularization penalty). An exception is the ECOC-based algorithm proposed by Zhang (2014). Though it does not minimize superset errors explicitly, this algorithm generally predicts a class label of an training instance from its superset and thus are also minimizing superset errors. In this paper, we define the superset error as the empirical risk in the SLL problem, and we analyze the performance of learners that minimize the empirical risk (ERM learners). We only investigate the realizable case where the true multiclass classifier is in the hypothesis space.

The key to SLL learnability is that any hypothesis with nonzero classification error must have a significant chance of making superset errors. This in turn depends on the size and distribution of the label supersets. Small supersets, for example, are more informative than large ones. Precisely speaking, a sufficient condition for learnability is that the classification error can be bounded by the superset error.

The SLL problem arises in two settings. In the first setting, which we call SSL-I, instances are independent of each other, and the superset is selected independently for each instance. The second setting, which we call SLL-B, arises from the multi-instance multi-label learning (MIML) problem (Zhou & Zhang, 2006; Zhou et al., 2012), where the training data are given in the form of MIML bags.

For SLL-I, Cour, Sapp, and Taskar (2011) proposed the concept of *ambiguity degree*. This bounds the probability that a specific distractor label appears in the superset of a given instance. When the ambiguity degree is less than 1, Cour, et al. give a relationship between the classification error and the superset error. With the same condition, we show that the sample complexity of SLL-I with ambiguity degree zero matches the complexity of multiclass classification.

For SLL-B, the training data have the form of independent MIML bags. Each MIML bag consists of a bag of instances

and a set of labels (the "bag label"). Each instance has exactly one (unknown) label, and the bag label is exactly the union of the labels of all of the instances. (See Zhou et al. (2012) for discussion of other ways in which MIML bags can be generated.) The learner only observes the bag label and not the labels of the individual instances.

It is interesting to note that several previous papers test their algorithms on synthetic data corresponding to the SSL-I scenario, but then apply them to a real application corresponding to the SSL-B setting.

To show learnability, we convert the SLL-B problem to a binary classification problem with the general condition that the classification error can be bounded by the superset error times a multiplicative constant. We then provide a concrete condition for learnability: for any pair of class labels, they must not always co-occur on a bag label. That is, there must be non-zero probability of observing a bag that contains an instance of only one of the two labels. Given enough training data, we can verify with high confidence whether this condition holds.

The success of weakly supervised learning depends on the degree of correlation between the supervision information and the classification error. We show that superset learning exhibits a strong correlation between these because a superset error is always caused by a classification error. Our study of the SLL problem can be seen as a first step toward the analysis of more general weakly-supervised learning problems.

2. Related Work

Different superset label learning algorithms have been proposed by Jin and Ghahramani (2002); Nguyen and Caruana (2008); Cour, Sapp, and Taskar (2011); Liu and Dietterich (2012); and Zhang (2014). All these algorithms employ some loss to minimize superset errors on the training set. Cour, Sapp, and Taskar (2011) conducted some theoretical analysis of the problem. They proposed the concept of "ambiguity degree" and established a relationship between superset error and classification errors. They also gave a generalization bound for their algorithm. In their analysis, they assume instances are independent of each other.

Sample complexity analysis of multiclass learning provides the basis of our analysis of SLL problem with independent instances. The Natarajan dimension (Natarajan, 1989) is an important instrument for characterizing the capacity of multiclass hypothesis spaces. Ben-David et al. (1995) and (Daniely et al., 2011) give sample complexity bounds in terms of this dimension.

The MIML framework was proposed by Zhou & Zhang (2006), and the instance annotation problem is raised by Briggs et al. (2012). Though the Briggs, et al. algorithm explicitly uses bag information, it is still covered by our analysis of the SLL-B problem. There is some theoretical analysis of multi-instance learning (Blum & Kalai, 1998; Long & Tan, 1998; Sabato & Tishby, 2009), but we only know of one paper on the learnability of the MIML problem (Wang & Zhou, 2012). In that work, no assumption is made for the distribution of instances within a bag, but the labels on a bag must satisfy some correlation conditions. In our setting, we assume the distribution of the labels of instances in a bag is arbitrary, but that these instances are independent of each other given their labels.

3. Superset Label Learning Problem with Independent Instances (SLL-I)

Let $\mathcal{X}$ be the instance space and $\mathcal{Y} = \{1, 2, \ldots, L\}$ be the finite label space. The superset space $\mathcal{S}$ is the powerset of $\mathcal{Y}$ without the empty set: $\mathcal{S} = 2^{\mathcal{Y}} - \{\emptyset\}$. A "complete" instance $(x, y, s) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{S}$ is composed of its features x, its true label y, and the label superset s. We decompose $\mathcal{D}$ into the standard multiclass distribution $\mathcal{D}^{xy}$ defined on $\mathcal{X} \times \mathcal{Y}$ and the label set conditional distribution $\mathcal{D}^s(x, y)$ defined over $\mathcal{S}$ given (x, y). We assume that $Pr_{s\sim\mathcal{D}^s(x,y)}(y \in s) = 1$, that is, the true label is always in the label superset. Other labels in the superset will be called *distractor labels*. Let $\mu(\cdot)$ denote the probability measure of a set; its subscript indicates the distribution. Denote a sample of instances by $\mathbf{z} = \left\{(x_i, y_i, s_i)\right\}_{i=1}^{n}$, where each instance is sampled from $\mathcal{D}$ independently. The size of the sample is always n. Although the true label is included in the training set in our notation, it is not visible to the learner. Let $\mathbb{I}(\cdot)$ denote the indicator function, which has the value 1 when its argument is true and 0 otherwise.

The hypothesis space is denoted by $\mathcal{H}$, and each $h \in \mathcal{H} : \mathcal{X} \to \mathcal{Y}$ is a multiclass classifier. The *expected classification error* of h is defined as

$$Err_{\mathcal{D}}(h) = E_{(x,y,s)\sim\mathcal{D}}\mathbb{I}(h(x) \neq y). \tag{1}$$

We use H_ϵ to denote the set of hypotheses with error at least ϵ, $H_\epsilon = \{h \in \mathcal{H} : Err_{\mathcal{D}}(h) \geq \epsilon\}$. The superset error is defined as the event that the predicted label is not in the superset: $h(x) \notin s$. The *expected superset error* and the *average superset error* on set $\mathbf{z}$ are defined as

$$Err^s_{\mathcal{D}}(h) = E_{(x,y,s)\sim\mathcal{D}}\mathbb{I}(h(x) \notin s) \tag{2}$$

$$Err^s_{\mathbf{z}}(h) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(h(x_i) \notin s_i) \tag{3}$$

It is easy to see that the expected superset error is always no greater than the expected classification error.

For conciseness we often omit the word "expected" or "average" when referring these errors defined above. The

meaning should be clear from how the error is calculated.

An Empirical Risk Minimizing (ERM) learner $\mathcal{A}$ for $\mathcal{H}$ is a function, $\mathcal{A} : \cup_{n=0}^{\infty}(\mathcal{X} \times \mathcal{S})^n \mapsto \mathcal{H}$. We define the *empirical risk* as the average superset error on the training set. The ERM learner for hypothesis space $\mathcal{H}$ always returns an hypothesis $h \in \mathcal{H}$ with minimum superset error for training set $\mathbf{z}$.

$$\mathcal{A}(\mathbf{z}) = \arg\min_{h \in \mathcal{H}} Err^s_{\mathbf{z}}(h)$$

Since the learning algorithm $\mathcal{A}$ can only observe the superset label, this definition of ERM is different from that of multiclass classification. In the realizable case, there exists $h_0 \in \mathcal{H}$ such that $Err_{\mathcal{D}}(h_0) = 0$.

3.1. Small ambiguity degree condition

On a training set with label supersets, an hypothesis will not be rejected by an ERM learner as long as its predictions are contained in the superset labels of the training instances. If a distractor label always co-occurs with one of the true labels under distribution $\mathcal{D}^s$, there will be no information to discriminate the true label from the distractor. On the contrary, if for any instance all labels except the true label have non-zero probability of being missing from the superset, then the learner always has some probability of rejecting an hypothesis if it predicts this instance incorrectly. The *ambiguity degree*, proposed by Cour et al. (2011), is defined as

$$\gamma = \sup_{\substack{(x,y)\in\mathcal{X}\times\mathcal{Y},\ \ell\in\mathcal{Y}:\\ p(x,y)>0,\ \ell\neq y}} Pr_{s\sim\mathcal{D}^s(x,y)}(\ell \in s). \tag{4}$$

This is the maximum probability that some particular distractor label ℓ co-occurs with the true label y. If $\gamma = 0$, then with probability one there are no distractors. If $\gamma = 1$, then there exists at least one pair y and ℓ that always co-occur. If a problem exhibits ambiguity degree γ, then a classification error made on any instance will be detected (i.e., lie outside the bag label) with probability at least $1 - \gamma$:

$$Pr(h(x) \notin s | h(x) \neq y, x, y) \geq 1 - \gamma.$$

If an SSL-I problem exhibits $\gamma < 1$, then we say that it satisfies the *small ambiguity degree condition*. We prove that this is sufficient for ERM learnability of the SSL-I problem.

Theorem 3.1 *Suppose an SLL-I problem has ambiguity degree $\gamma, 0 \leq \gamma < 1$. Let $\theta = \log\frac{2}{1+\gamma}$, and suppose the Natarajan dimension of the hypothesis space $\mathcal{H}$ is $d_{\mathcal{H}}$. Define*

$$n_0(\mathcal{H}, \epsilon, \delta) = \frac{4}{\theta\epsilon}\left(d_{\mathcal{H}}\left(\log(4d_{\mathcal{H}}) + 2\log L + \log\frac{1}{\theta\epsilon}\right) + \log\frac{1}{\delta} + 1\right),$$

Then when $n > n_0(\mathcal{H}, \epsilon, \delta)$, $Err_{\mathcal{D}}(\mathcal{A}(\mathbf{z})) < \epsilon$ with probability $1 - \delta$.

We follow the method of proving learnability of binary classification to prove this theorem (Anthony & Biggs, 1997). Let $R_{n,\epsilon}$ be the set of all n-samples for which there exists an ϵ-bad hypothesis h with zero empirical risk:

$$R_{n,\epsilon} = \{\mathbf{z} \in (\mathcal{X}\times\mathcal{Y}\times\mathcal{S})^n : \exists h \in H_\epsilon, Err^s_{\mathbf{z}}(h) = 0\}.$$

Essentially we need to show that $Pr(R_{n,\epsilon}) \leq \delta$.

The proof is composed of the following two lemmas.

Lemma 3.2 *We introduce a testing set $\mathbf{z}'$ of size n with each instance drawn independently from distribution $\mathcal{D}$, and define the set $S_{n,\epsilon}$ to be the event that there exists a hypothesis in H_ϵ that makes no superset errors on training set $\mathbf{z}$ but makes at least $\frac{\epsilon}{2}$ classification errors on testing set $\mathbf{z}'$.*

$$S_{n,\epsilon} = \Big\{(\mathbf{z}, \mathbf{z}') \in (\mathcal{X}\times\mathcal{Y}\times\mathcal{S})^{2n} : \exists h \in H_\epsilon, Err^s_{\mathbf{z}}(h) = 0, Err_{\mathbf{z}'}(h) \geq \frac{\epsilon}{2}\Big\}.$$

Then $Pr\big((\mathbf{z}, \mathbf{z}') \in S_{n,\epsilon}\big) \geq \frac{1}{2}Pr(\mathbf{z} \in R_{n,\epsilon})$ when $n > \frac{8}{\epsilon}$.

Proof This lemma is used in many learnability proofs. Here we only give a proof sketch. $S_{n,\epsilon}$ is a subevent of $R_{n,\epsilon}$. We can apply the chain rule of probability to write $Pr\big((\mathbf{z}, \mathbf{z}') \in S_{n,\epsilon}\big) = Pr\big((\mathbf{z}, \mathbf{z}') \in S_{n,\epsilon} | \mathbf{z} \in R_{n,\epsilon}\big) Pr\big(\mathbf{z} \in R_{n,\epsilon}\big)$. Let $H(\mathbf{z}) = \{h \in \mathcal{H} : Err^s_{\mathbf{z}}(h) = 0\}$ be the set of hypotheses with zero empirical risk on the training sample. Then we have

$$\begin{aligned} & Pr\left((\mathbf{z}, \mathbf{z}') \in S_{n,\epsilon} | \mathbf{z} \in R_{n,\epsilon}\right) \\ = \ & Pr\left(\left\{\exists h \in H_\epsilon \cap H(\mathbf{z}), Err_{\mathbf{z}'}(h) \geq \frac{\epsilon}{2}\right\} \middle| \mathbf{z} \in R_{n,\epsilon}\right) \\ \geq \ & Pr\left(\left\{h \in H_\epsilon \cap H(\mathbf{z}), Err_{\mathbf{z}'}(h) \geq \frac{\epsilon}{2}\right\} \middle| \mathbf{z} \in R_{n,\epsilon}\right) \end{aligned}$$

In the last line, h is a particular hypothesis in the intersection. Since h has error at least ϵ, we can bound the probability via the Chernoff bound. When $n > \frac{8}{\epsilon}$, we obtain $Pr\big((\mathbf{z}, \mathbf{z}') \in S_{n,\epsilon} \mid \mathbf{z} \in R_{n,\epsilon}\big) > \frac{1}{2}$ which completes the proof. ∎

With Lemma 3.2, we can bound the probability of $R_{n,\epsilon}$ by bounding the probability of $S_{n,\epsilon}$. We will do this using the technique of swapping training/testing instance pairs, which is used in various proofs of learnability.

In the following proof, the sets $\mathbf{z}$ and $\mathbf{z}'$ are expanded to $(\mathbf{x}, \mathbf{y}, \mathbf{s})$ and $(\mathbf{x}', \mathbf{y}', \mathbf{s}')$ respectively when necessary. Let the training and testing instances form n training/testing pairs by arbitrary pairing. The two instances in each pair are respectively from the training set and the testing set,

and these two instances are both indexed by the pair index, which is indicated by a subscript. Define a group G of swaps with size $|G| = 2^n$. A swap $\sigma \in G$ has an index set $J_\sigma \subseteq \{1, \ldots, n\}$, and it swaps the training and testing instances in the pairs indexed by J_σ. We write σ as a superscript to indicate the result of applying σ to the training and testing sets, that is, $\sigma(\mathbf{z}, \mathbf{z}') = (\mathbf{z}^\sigma, \mathbf{z}'^\sigma)$.

Lemma 3.3 *If the hypothesis space $\mathcal{H}$ has Natarajan dimension $d_\mathcal{H}$ and $\gamma < 1$, then*

$$Pr(S_{n,\epsilon}) \leq (2n)^{d_\mathcal{H}} L^{2d_\mathcal{H}} \exp\left(-\frac{n\theta\epsilon}{2}\right)$$

Proof Since the swap does not change the measure of $(\mathbf{z}, \mathbf{z}')$,

$$\begin{aligned} & 2^n Pr\left((\mathbf{z}, \mathbf{z}') \in S_{n,\epsilon}\right) \\ = & \sum_{\sigma \in G} E\left[Pr\left((\mathbf{z}, \mathbf{z}') \in S_{n,\epsilon} | \mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}'\right)\right] \\ = & \sum_{\sigma \in G} E\left[Pr\left(\sigma(\mathbf{z}, \mathbf{z}') \in S_{n,\epsilon} | \mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}'\right)\right] \\ = & E\left[\sum_{\sigma \in G} Pr\left(\sigma(\mathbf{z}, \mathbf{z}') \in S_{n,\epsilon} | \mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}'\right)\right] \quad (5) \end{aligned}$$

The expectations are taken with respect to $(\mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}')$. The probability in the expectation comes from the randomness of $(\mathbf{s}, \mathbf{s}')$ given $(\mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}')$.

Let $\mathcal{H}|(\mathbf{x}, \mathbf{x}')$ be the set of hypothesis making different classifications for $(\mathbf{x}, \mathbf{x}')$. Define set $S^h_{n,\epsilon}$ for each hypothesis $h \in \mathcal{H}$ as

$$S^h_{n,\epsilon} = \left\{(\mathbf{z}, \mathbf{z}') : Err^s_\mathbf{z}(h) = 0, Err_{\mathbf{z}'}(h) \geq \frac{\epsilon}{2}\right\}$$

By the union bound, we have

$$\begin{aligned} & \sum_{\sigma \in G} Pr\left(\sigma(\mathbf{z}, \mathbf{z}') \in S_{n,\epsilon} | \mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}'\right) \\ \leq & \sum_{h \in \mathcal{H}|(\mathbf{x}, \mathbf{x}')} \sum_{\sigma \in G} Pr\left(\sigma(\mathbf{z}, \mathbf{z}') \in S^h_{n,\epsilon} \big| \mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}'\right) \quad (6) \end{aligned}$$

By Natarajan (1989), $\left|\mathcal{H}|(\mathbf{z}, \mathbf{z}')\right| \leq (2n)^{d_\mathcal{H}} L^{2d_\mathcal{H}}$. The only work left is to bound $\sum_{\sigma \in G} Pr\big(\sigma(\mathbf{z}, \mathbf{z}') \in S^h_{n,\epsilon} | \mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}'\big)$, and this part is our contribution.

Here is our strategy. We first fix $(\mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}')$ and σ and bound $Pr\big(\sigma(\mathbf{z}, \mathbf{z}') \in S^h_{n,\epsilon} \mid \mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}'\big)$ with the ambiguity degree assumption. Then we find an upper bound of the summation over σ. Start by expanding the condition in

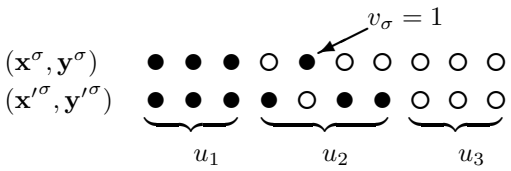

Figure 1. A situation of $(\mathbf{x}^\sigma, \mathbf{y}^\sigma, \mathbf{x}'^\sigma, \mathbf{y}'^\sigma)$. Black dots represent instances misclassified by h and circles represent instances correctly classified by h.

$S^h_{n,\epsilon}$,

$$\begin{aligned} & Pr\big(\sigma(\mathbf{z}, \mathbf{z}') \in S^h_{n,\epsilon} \mid \mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}'\big) \\ = & \mathbb{I}\big(Err_{\mathbf{z}'^\sigma}(h) \geq \frac{\epsilon}{2}\big) \cdot \\ & \qquad Pr\big(h(\mathbf{x}^\sigma_i) \in \mathbf{s}^\sigma_i, 1 \leq i \leq n | \mathbf{x}^\sigma, \mathbf{y}^\sigma\big) \\ = & \mathbb{I}\big(Err_{\mathbf{z}'^\sigma}(h) \geq \frac{\epsilon}{2}\big) \prod_{i=1}^n Pr\big(h(\mathbf{x}^\sigma_i) \in \mathbf{s}^\sigma_i | \mathbf{x}^\sigma_i, \mathbf{y}^\sigma_i\big). \end{aligned}$$

For a pair of training/testing sets $(\mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}')$, let u_1, u_2 and u_3 represent the number of pairs for which h classifies both incorrectly, one incorrectly, and both correctly. Let v_σ, $0 \leq v_\sigma \leq u_2$, be the number of wrongly-predicted instances swapped into the training set $(\mathbf{x}^\sigma, \mathbf{y}^\sigma)$. One such situation is shown in Figure 1. There are $u_1 + u_2 - v_\sigma$ wrongly-predicted instances in the testing set. The error condition $Err_{\mathbf{z}'^\sigma}(h) \geq \frac{\epsilon}{2}$ is equivalent to $u_1 + u_2 - v_\sigma \geq \frac{\epsilon}{2}n$, which always indicates $u_1 + u_2 \geq \frac{\epsilon}{2}n$. So we have $\mathbb{I}\big(Err_{\mathbf{z}'^\sigma}(h) \geq \frac{\epsilon}{2}\big) \leq \mathbb{I}\big(u_1 + u_2 \geq \frac{\epsilon}{2}n\big)$.

There are $u_1 + v_\sigma$ wrongly-predicted instances in the training set. Since the true label is in the superset with probability one, while the wrong label appears in the superset with probability no greater than γ by (4), we have $\prod_{i=1}^n Pr\big(h(\mathbf{x}^\sigma_i) \in \mathbf{s}^\sigma_i | \mathbf{x}^\sigma_i\big) \leq \gamma^{u_1 + v_\sigma}$.

Now for a single swap σ, we have the bound

$$\begin{aligned} & Pr\big(\sigma(\mathbf{z}, \mathbf{z}') \in S^h_{n,\epsilon} \mid \mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}'\big) \\ & \qquad \leq \mathbb{I}\big(u_1 + u_2 \geq \frac{\epsilon}{2}n\big)\gamma^{u_1 + v_\sigma} \quad (7) \end{aligned}$$

Let us sum up (7) over σ. Any swap σ can freely switch instances in $u_1 + u_3$ without changing the bound in (7), and choose from the u_2 pairs v_σ to switch. For each value $0 \leq j \leq u_2$, there are $2^{u_1 + u_3}\binom{u_2}{j}$ swaps that have $v_\sigma = j$.

Therefore,

$$\begin{aligned}
&\sum_{\sigma\in G}\mathbb{I}\big(u_1+u_2\geq\frac{\epsilon}{2}n\big)\gamma^{u_1+v_\sigma}\\
\leq\ &\mathbb{I}\big(u_1+u_2\geq\frac{\epsilon}{2}n\big)\,2^{u_1+u_3}\sum_{j=0}^{u_2}\binom{u_2}{j}\gamma^{u_1+j}\\
=\ &\mathbb{I}\big(u_1+u_2\geq\frac{\epsilon}{2}n\big)\,2^{n-u_2}\gamma^{u_1}(1+\gamma)^{u_2}\\
=\ &\mathbb{I}\big(u_1+u_2\geq\frac{\epsilon}{2}n\big)\,2^{n}\gamma^{u_1}\left(\frac{1+\gamma}{2}\right)^{u_2}
\end{aligned}$$

When $(\mathbf{x},\mathbf{y},\mathbf{x}',\mathbf{y}')$ and h make $u_1=0$ and $u_2=\frac{\epsilon}{2}n$, the right side reaches its maximum $2^n\left(\frac{1+\gamma}{2}\right)^{n\epsilon/2}$, which is $2^n e^{-n\theta\epsilon/2}$ with the definition of θ in Theorem 3.1. Applying this to (6) and (5), completes the proof. ∎

Proof of Theorem 3.1. By combining the results of the two lemmas, we have $P(R_{n,\epsilon})\leq 2^{(d_\mathcal{H}+1)}n^{d_\mathcal{H}}L^{2d_\mathcal{H}}\exp(-\frac{n\theta\epsilon}{2})$. Bound this with δ on a log scale to obtain

$$(d_\mathcal{H}+1)\log 2+d_\mathcal{H}\log n+2d_\mathcal{H}\log L-\frac{\theta\epsilon n}{2}\leq\log\delta$$

By bounding $\log n$ with $(\log(\frac{4d_\mathcal{H}}{\theta\epsilon})-1)+\frac{\theta\epsilon}{4d_\mathcal{H}}n$, we get a linear form for n. Then we solve for n to obtain the result. ∎

Remark Theorem 3.1 includes the multiclass problem as a special case. When $\gamma=0$ and $\theta=\log 2$, we get the same sample complexity as the multiclass classification problem. (See Theorem 6 in Daniely et al. (2011).)

The small ambiguity degree condition is strong. A good aspect of our result is that we only require that $\mathcal{H}$ has finite Natarajan dimension to ensure that an ERM learning finds a good hypothesis. However, the result is not general enough. Here is an example where the small ambiguity degree condition is not satisfied but the problem is still learnable. Consider a case where the true label is ℓ_1, but the label superset always contains a distractor label ℓ_2; otherwise the superset contains only the true label. Such a distribution of superset labels certainly does not satisfy the small ambiguity condition, but if no hypothesis in $\mathcal{H}$ classifies ℓ_1 as ℓ_2, then the problem is still learnable by an ERM learner. Though the example is not very realistic, it suggests that there should be a more general condition for ERM learnability.

3.2. A general condition for learnability of SLL-I

We can obtain a more general condition for ERM learnability by constructing a binary classification task from the superset label learning task. Two key points need special attention in the construction: how to relate the classification errors of the two tasks, and how to specify the hypothesis space of the binary classification task and obtain its VC-dimension.

We first construct a binary classification problem from the SLL-I problem. Given an instance (x,s), the binary classifier needs to predict whether the label of x is outside of s, that is, predict the value of $\mathbb{I}(y_x\notin s)$. If (x,s) is from the distribution $\mathcal{D}$, then "0" is always the correct prediction. Let $\mathcal{F}_\mathcal{H}$ be the space of these binary classifiers. $\mathcal{F}_\mathcal{H}$ is induced from $\mathcal{H}$ as follows:

$$\mathcal{F}_\mathcal{H}=\{f_h:\ f_h(x,s)=\mathbb{I}(h(x)\notin s),h\in\mathcal{H}\}.$$

Though the inducing relation is a surjection from $\mathcal{H}$ to $\mathcal{F}_\mathcal{H}$, we assume the subscript h in the notation f_h can be any $h\in\mathcal{H}$ inducing f_h. The binary classification error of f_h is the superset error of h in the SLL-I problem.

Denote the ERM learner of the binary classification problem by $\mathcal{A}^s$. The learner $\mathcal{A}^s$ actually calls $\mathcal{A}$, which returns an hypothesis $h^\star$ and then $\mathcal{A}^s$ returns the induced hypothesis $f_{h^\star}$. In the realizable case, both $h^\star$ and $f_{h^\star}$ have zero training error on their respective classification tasks.

A sufficient condition for learnability of an SLL-I problem is that any hypothesis with non-zero classification error will cause a superset error with relatively large probability Let η be defined by

$$\eta=\inf_{h\in\mathcal{H}:Err_\mathcal{D}(h)>0}\frac{Err^s_\mathcal{D}(h)}{Err_\mathcal{D}(h)}. \tag{8}$$

Define the *tied error condition* as $\eta>0$. Then we can bound the multiclass classification error by bounding the superset error when this condition holds.

We need to find the VC-dimension of $\mathcal{F}_\mathcal{H}$ first. It can be bounded by the Natarajan dimension of $\mathcal{H}$.

Lemma 3.4 *Denote the VC-dimension of $\mathcal{F}_\mathcal{H}$ as $d_\mathcal{F}$ and the Natarajan dimension of $\mathcal{H}$ is $d_\mathcal{H}$, then*

$$d_\mathcal{F}\ <\ 4\,d_\mathcal{H}(\log d_\mathcal{H}+2\log L)$$

Proof There is a set of $d_\mathcal{F}$ instances that can be shattered by $\mathcal{F}_\mathcal{H}$—that is, there are functions in $\mathcal{F}_\mathcal{H}$ that implement each of the $2^{d_\mathcal{F}}$ different ways of classifying these $d_\mathcal{F}$ instances. Any two different binary classifications must be the result of two different label predictions for these $d_\mathcal{F}$ instances. According to Natarajan's (1989) original result on Natarajan dimension, there are at most $d_\mathcal{F}^{d_\mathcal{H}}L^{2d_\mathcal{H}}$ different multiclass classifications on these instances. Therefore,

$$2^{d_\mathcal{F}}\leq d_\mathcal{F}^{d_\mathcal{H}}L^{2d_\mathcal{H}}.$$

Taking the logarithm of both sides gives us

$$d_\mathcal{F}\log 2\ \leq\ d_\mathcal{H}\log d_\mathcal{F}+2d_\mathcal{H}\log L.$$

By bounding $\log d_{\mathcal{F}}$ above by $\log d_{\mathcal{H}} + \frac{d_{\mathcal{F}}}{e d_{\mathcal{H}}}$, we get

$$d_{\mathcal{F}} \log 2 \quad \leq \quad d_{\mathcal{H}} \log d_{\mathcal{H}} + \frac{d_{\mathcal{F}}}{e} + 2 d_{\mathcal{H}} \log L.$$

By observing that $(\log 2 - e^{-1})^{-1} < 4$, we obtain the result. ∎

Now we can bound the multiclass classification error by bounding the superset error.

Theorem 3.5 *Assume $\eta > 0$ and assume $\mathcal{H}$ has a finite Natarajan dimension $d_{\mathcal{H}}$, then $\mathcal{A}$ returns an hypothesis with error less than ϵ with probability at least $1 - \delta$ when the training set has size $n > n_1(\mathcal{H}, \delta, \epsilon)$, which is defined as*

$$n_1(\mathcal{H}, \delta, \epsilon) = \frac{4}{\eta\epsilon}\left(4 d_{\mathcal{H}}(\log d_{\mathcal{H}} + 2 \log L) \log\left(\frac{12}{\eta\epsilon}\right) + \log\left(\frac{2}{\delta}\right)\right).$$

Proof Learner $\mathcal{A}$ returns hypothesis $h^\star \in \mathcal{H}$ with $Err^s_{\mathbf{D}}(h^\star) = 0$. The corresponding binary hypothesis $f_{h^\star} \in \mathcal{F}_{\mathcal{H}}$ makes no superset error on the training data.

By Lemma (3.4), $n_1(\mathcal{H}, \delta, \epsilon) \geq n_1^s(\mathcal{F}_{\mathcal{H}}, \delta, \eta\epsilon)$, where

$$n_1^s(\mathcal{H}, \delta, \eta\epsilon) = \frac{4}{\eta\epsilon}\left(d_{\mathcal{F}} \log\left(\frac{12}{\eta\epsilon}\right) + \log\left(\frac{2}{\delta}\right)\right).$$

When $n > n_1(\mathcal{H}, \delta, \epsilon) \geq n_1^s(\mathcal{F}_{\mathcal{H}}, \delta, \eta\epsilon)$, by Theorem 8.4.1 in Anthony & Biggs (1997), $Err^s_{\mathcal{D}}(f_{h^\star}) < \eta\epsilon$ with probability at least $1 - \delta$. Therefore $Err_{\mathcal{D}}(h^\star) < \epsilon$ with probability at least $1 - \delta$. ∎

The necessary condition for the learnability of the SLL problem is that no hypothesis have non-zero multiclass classification error but have zero superset error. Otherwise, no training data can reject such an incorrect hypothesis.

Theorem 3.6 *If there exists an hypothesis $h \in \mathcal{H}$ such that $Err_{\mathcal{D}}(h) > 0$ and $Err^s_{\mathcal{D}}(h) = 0$, then the SLL-I problem is not learnable by an ERM learner.*

The gap between the general sufficient condition and the necessary condition is small.

Let's go back and check the small ambiguity degree condition against the tied error condition. The condition $\gamma < 1$ indicates $Pr(h(x) \notin s | x, y) \geq (1 - \gamma) Pr(h(x) \neq y | x, y)$. By taking the expectation of both sides, we obtain the tied error condition $\eta \geq 1 - \gamma > 0$. This also shows that the tied error condition is more general. The tied error condition is less practical, but it can be used as a guideline to find more sufficient conditions.

4. Superset Label Learning Problem with Bagged Training Data (SLL-B)

The second type of superset label problem arises in multi-instance multilabel learning (Zhou & Zhang, 2006). The data are given in the form of i.i.d. bags. Each bag contains multiple instances, which are generally not independent of each other. The bag label set consists of the labels of these instances. In this form of superset label learning problem, the bag label set provides the label superset for each instance in the bag. An ERM learning algorithm for SSL-I can naturally be applied here regardless of the dependency between instances. In this section, we will show learnability of this SSI-B problem.

We assume each bag contains r instances, where r is fixed as a constant. The space of labeled instances is still $\mathcal{X} \times \mathcal{Y}$. The space of sets of bag labels is also $\mathcal{S}$. The space of bags is $\mathcal{B} = (\mathcal{X} \times \mathcal{Y})^r \times \mathcal{S}$. Denote a bag of instances as $B = (X, Y, S)$, where $(X, Y) \in (\mathcal{X} \times \mathcal{Y})^r$ and $S \in \mathcal{S}$. Note that although (X, Y) is written as a vector of instances for notational convenience, the order of these instances is not essential to any conclusion in this work. We assume that each bag $B = (X, Y, S)$ is sampled from the distribution $\mathcal{D}^B$. The label set S consists of labels in Y in the MIML setting, but the following analysis still applies when S contains extra labels. The learner can only observe (X, S) during training, whereas the learned hypothesis is tested on (X, Y) during testing. The boldface $\mathbf{B}$ always denotes a set with m independent bags drawn from the distribution $\mathcal{D}^B$.

The hypothesis space is still denoted by $\mathcal{H}$. The expected classification error of hypothesis $h \in \mathcal{H}$ on bagged data is defined as

$$Err_{\mathcal{D}^B}(h) \quad = \quad E_{B \sim \mathcal{D}^B}\left[\frac{1}{r}\sum_{i=1}^{r} \mathbb{I}(h(X_i) \neq Y_i)\right] \quad (9)$$

The expected superset error and the average superset error on the set $\mathbf{B}$ are defined as

$$Err^s_{\mathcal{D}^B}(h) \quad = \quad E_{B \sim \mathcal{D}^B}\left[\frac{1}{r}\sum_{i=1}^{r} \mathbb{I}(h(X_i) \notin S_i)\right] \quad (10)$$

$$Err^s_{\mathbf{B}}(h) \quad = \quad \frac{1}{mr}\sum_{B \in \mathbf{B}}\sum_{i=1}^{r} \mathbb{I}(h(X_i) \notin S_i). \quad (11)$$

The ERM learner $\mathcal{A}$ for hypothesis space $\mathcal{H}$ returns the hypothesis with minimum average superset error.

4.1. The general condition for learnability of SLL-B

Using a technique similar to that employed in the last section, we convert the SLL-B problem into a binary classification problem over bags. By bounding the error on the

binary classification task, we can bound the multiclass classification error of the hypothesis returned by $\mathcal{A}$. Let $\mathcal{G}_{\mathcal{H}}$ be the hypothesis space for binary classification of bags. $\mathcal{G}_{\mathcal{H}}$ is induced from $\mathcal{H}$ as follows:

$$\mathcal{G}_{\mathcal{H}} = \{g_h : g_h(X,S) = \max_{1\le i\le r} \mathbb{I}(h(X_i) \notin S_i), h \in \mathcal{H}\}.$$

Every hypothesis $g_h \in \mathcal{G}_{\mathcal{H}}$ is a binary classifier for bags that predicts whether any instance in the bag has its label outside of the bag label set. Since 0 is the correct classification for every bag from the distribution $\mathcal{D}^B$, we define the expected and empirical *bag error* of g_h as

$$Err^B_{\mathcal{D}^B}(g_h) = E_{B\sim\mathcal{D}^B} \max_{1\le i\le r} \mathbb{I}(h(X_i) \notin S_i) \quad (12)$$

$$Err^B_{\mathbf{B}}(g_h) = \frac{1}{m}\sum_{B\in\mathbf{B}} \max_{1\le i\le r} \mathbb{I}(h(X_i) \notin S_i). \quad (13)$$

It is easy to check the following relation between $Err^B_{\mathcal{D}^B}(g_h)$ and $Err^s_{\mathcal{D}^B}(h)$.

$$Err^s_{\mathcal{D}^B}(h) \le Err^B_{\mathcal{D}^B}(g_h) \le rErr^s_{\mathcal{D}^B}(h). \quad (14)$$

Denote by $\mathcal{A}^B$ the ERM learner for this binary classification problem. In the realizable case, if the hypothesis $g_{h^\star}$ is returned by $\mathcal{A}^B$, then $g_{h^\star}$ has no binary classification error on the training bags and $h^\star$ makes no superset error on any instance in training data.

To bound the error for the binary classification problem, we need to bound the VC-dimension of $\mathcal{G}_{\mathcal{H}}$.

Lemma 4.1 *Denote the VC-dimension of $\mathcal{G}_{\mathcal{H}}$ by $d_{\mathcal{G}}$ and the Natarajan dimension of $\mathcal{H}$ by $d_{\mathcal{H}}$, then*

$$d_{\mathcal{G}} < 4\, d_{\mathcal{H}}(\log d_{\mathcal{H}} + \log r + 2\log L).$$

The proof of this lemma is almost the same as Lemma 3.4. The only difference is that different classifications of $d_{\mathcal{G}}$ bags are caused by different classifications of $rd_{\mathcal{G}}$ instances.

The tied error condition for the SLL-B problem is $\lambda > 0$,

$$\lambda = \inf_{h\in\mathcal{H}: Err_{\mathcal{D}}(h)>0} \frac{Err^s_{\mathcal{D}^B}(h)}{Err_{\mathcal{D}^B}(h)}.$$

With the tied error condition, we can give the sample complexity of learning a multiclass classifier with multi-instance bags.

Theorem 4.2 *Suppose the tied error condition holds, $\lambda > 0$, and assume $\mathcal{H}$ has finite Natarajan dimension $d_{\mathcal{H}}$, then $\mathcal{A}(\mathbf{B})$ returns an hypothesis with error less than ϵ with probability at least $1-\delta$ when the training set $\mathbf{B}$ has size m and $m > m_0(\mathcal{H},\delta,\epsilon)$,*

$$m_0(\mathcal{H},\delta,\epsilon) = \frac{4}{\lambda\epsilon}\left(4d_{\mathcal{H}}\log(d_{\mathcal{H}}rL^2)\log\left(\frac{12}{\lambda\epsilon}\right) + \log\left(\frac{2}{\delta}\right)\right)$$

Proof With the relation in (14), we have $Err_{\mathcal{D}^B}(h) \le \frac{1}{\lambda}Err^B_{\mathcal{D}^B}(g_h)$. The hypothesis $h^\star$ returned by $\mathcal{A}$ has zero superset error on the training bags, thus $g_{h^\star}$ has zero bag error. When $m > m_0$, we have $Err^B_{\mathcal{D}^B}(g_h) < \lambda\epsilon$ with probability at least $1-\delta$ by Theorem 8.4.1 in Anthony & Biggs (1997). Hence, we have $Err_{\mathcal{D}^B}(h) < \epsilon$ with probability at least $1-\delta$. ∎

4.2. The no co-occurring label condition

We now provide a more concrete sufficient condition for learnability with a principle similar to the ambiguity degree. It generally states that any two labels do not always co-occur in bag label sets.

First we make an assumption about the data distribution.

Assumption 4.3 (Conditional independence assumption)

$$Pr(X|Y) = \prod_{i=1}^{r} Pr(X_i|Y_i)$$

This assumption states that the covariates of an instance are determined by its label only. With this assumption, a bag is sampled in the following way. Instance labels Y of the bag are first drawn from the marginal distribution $\mathcal{D}(Y_i)$, then for each $Y_i, 1 \le i \le r$, X_i is sampled from distribution $\mathcal{D}^x(Y_i)$ independently.

Remark We can compare our assumption of bag distribution with assumptions in previous work. Blum & Kalai (1998) assume that all instances in a bag are independently sampled from the instance distribution. As stated by Sabato & Tishby (2009), this assumption is too strict for many applications. In their work as well as in Wang & Zhou (2012), the instances in a bag are assumed to be dependent, and they point out that a further assumption is needed to get a low error classifier. In our assumption above, we also assume dependency among instances in the same bag. The dependency only comes from the instance labels. This assumption is roughly consistent with human descriptions of the world. For example, in the description "a tall person stands by a red vehicle", the two labels "person" and "vehicle" capture most of the correlation, while the detailed descriptions of the person and the vehicle are typically much less correlated.

The distribution D^B of bags can be decomposed into $\mathcal{D}^Y$ and $\mathcal{D}^x(\ell), \ell \in \mathcal{Y}$. Here $\mathcal{D}^Y$ is a distribution over $\mathcal{Y}^r$. For each class $\ell \in \mathcal{Y}$, $\mathcal{D}^x(\ell)$ is a distribution over the instance space $\mathcal{X}$. As a whole, the distribution of X is denoted as $\mathcal{D}^X(Y)$.

With Assumption 4.3, we propose the *no co-occurring label condition* in the following theorem.

Theorem 4.4 *Define*

$$\alpha = \inf_{(\ell,\ell')\in I} E_{(X,Y,S)\sim\mathcal{D}^B}[\mathbb{I}(\ell \in S)\mathbb{I}(\ell' \notin S)]$$

where I is the index set $\{(\ell,\ell') : \ell,\ell' \in \mathcal{Y}, \ell \neq \ell'\}$. Suppose Assumption 4.3 holds and $\alpha > 0$, then

$$\inf_{h\in\mathcal{H}:\, Err_{\mathcal{D}^B}(h)>0} \frac{Err^s_{\mathcal{D}^B}(h)}{Err_{\mathcal{D}^B}(h)} \geq \frac{\alpha}{r}.$$

Proof Denote the row-normalized confusion matrix of each hypothesis $h \in \mathcal{H}$ as $U_h \in [0,1]^{L\times L}$.

$$U_h(\ell,\ell') = Pr_{x\sim\mathcal{D}^x(\ell)}(h(x) = \ell'),\ \ 1 \leq \ell,\ell' \leq L$$

The entry (ℓ,ℓ') of U_h means a random instance from class ℓ is classified as class ℓ' by h with probability $U_h(\ell,\ell')$. Each row of U_h sums to 1. Denote $k(Y,\ell)$ as the number of occurrences of label ℓ in Y.

Then the error of h can be expressed as

$$\begin{aligned} Err_{\mathcal{D}^B}(h) &= \frac{1}{r}E_Y\Big[\sum_{i=1}^{r}\sum_{\ell'\neq Y_i} U_h(Y_i,\ell') \mid Y\Big] \\ &= \frac{1}{r}E_Y\Big[\sum_{(\ell,\ell')\in I} k(Y,\ell)U_h(\ell,\ell') \mid Y\Big] \\ &\leq \frac{1}{r}\sum_{(\ell,\ell')\in I} rU_h(\ell,\ell') \\ &= \sum_{(\ell,\ell')\in I} U_h(\ell,\ell'). \end{aligned}$$

The expected superset error of h can be computed as

$$\begin{aligned} & Err^s_{\mathcal{D}^B}(h) \\ =\ & \frac{1}{r}E_Y\Big[\sum_{i=1}^{r}\sum_{\ell'\neq Y_i} \mathbb{I}(\ell' \notin S_i)U_h(Y_i,\ell') \mid Y\Big] \\ =\ & \frac{1}{r}E_Y\Big[\sum_{(\ell,\ell')\in I} k(Y,\ell)\mathbb{I}(\ell' \notin S_i)U_h(\ell,\ell') \mid Y\Big] \\ \geq\ & \frac{1}{r}\sum_{(\ell,\ell')\in I} \alpha U_h(\ell,\ell'). \end{aligned}$$

These two inequalities hold for any $h \in \mathcal{H}$, so the theorem is proved. ∎

With the no co-occurring label condition that $\alpha > 0$, the remaining learnability requirement is that the hypothesis space $\mathcal{H}$ have finite Natarajan dimension. A merit of the condition of Theorem 4.4 is that it can be checked on the training data with high confidence. Suppose we have a training data $\mathbf{B}$. Then the empirical estimate of α is

$$\alpha_{\mathbf{B}} = \frac{1}{m}\min_{(\ell,\ell')\in I} \sum_{(X,Y,S)\in\mathbf{B}} \mathbb{I}(\ell \in S)\mathbb{I}(\ell' \notin S),$$

If $\alpha_{\mathbf{B}} > 0$, then by the Chernoff bound, we can obtain a lower bound on $\alpha > 0$ with high confidence. Conversely, if $\alpha_{\mathbf{B}} = 0$ for a large training set, then it is quite possible that α is very small or even zero.

5. Conclusion and Future Work

In this paper, we analyzed the learnability of an ERM learner on two superset label learning problems: SLL-I (for independent instances) and SLL-B (for bagged instances). Both problems can be learned by the same learner regardless the (in)dependency among the instances.

For both problems, the key to ERM learnability is that the expected classification error of any hypothesis in the space can be bounded by the superset error. If the tied error condition holds, then we can construct a binary classification problem from the SLL problem and bound the expected error in the binary classification problem. By constructing a relationship between the VC-dimension and the Natarajan dimension of the original problem, the sample complexities can be given in terms of the Natarajan dimension.

For the SLL-I problem, the condition of small ambiguity degree guarantees learnability for problems with hypothesis spaces having finite Natarajan dimension. This condition leads to lower sample complexity bounds than those obtained from the general tied error condition. The sample complexity analysis with this condition generalizes the analysis of multiclass classification. For the SLL-B problem, we propose a reasonable assumption for the distribution of bags. With this assumption, we identify a practical condition stating that no two labels always co-occur in bag label sets.

There is more to explore in theoretical analysis of the superset label learning problem. Analysis is needed for the agnostic case. Another important issue is to allow noise in the training data by removing the assumption that the true label is always in the label superset.

Acknowledgments

We thank Qi Lou, Wei Wang and Shell Hu for useful discussions.

References

Aly, M. Survey on multi-class classification methods, 2005.

Anthony, M. and Biggs, N. *Computational Learning Theory*. Cambridge University Press, 1997.

Ben-David, S., Cesa-Bianchi, N., Haussler, D., and Long, P. Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. *Journal of Computer and System Sciences*, 50(1):74–86, 1995.

Blum, A. and Kalai, A. A note on learning from multiple-instance examples. *Machine Learning*, 30(1):23–29, 1998.

Briggs, F., Fern, X. Z., and Raich, R. Rank-loss support instance machines for MIML instance annotation. In *Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining (KDD'12)*, pp. 534–542, 2012.

Cour, T., Sapp, B., and Taskar, B. Learning from partial labels. *Journal of Machine Learning Research*, 12(May): 1501–1536, 2011.

Daniely, A., Sabato, S., Ben-David, S., and Shalev-Shwartz, S. Multiclass learnability and the ERM principle. In *The 24rd Annual Conference on Learning Theory (COLT'11)*, pp. 207–232, 2011.

Jin, R. and Ghahramani, Z. Learning with multiple labels. In *Advances in Neural Information Processing Systems 15 (NIPS'02)*, pp. 897–904, 2002.

Liu, L.-P. and Dietterich, T. G. A conditional multinomial mixture model for superset label learning. In *Advances in Neural Information Processing Systems 25 (NIPS'12)*, pp. 557–565, 2012.

Long, P. M. and Tan, L. PAC learning axis-aligned rectangles with respect to product distributions from multiple-instance examples. *Machine Learning*, 30(1):7–21, 1998.

Mukherjee, I. and Schapire, R. E. A theory of multiclass boosting. *Journal of Machine Learning Research*, 14 (Feb):437–497, 2013.

Natarajan, B.K. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

Nguyen, N. and Caruana, R. Classication with partial labels. In *Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining(KDD'08)*, pp. 551–559, 2008.

Sabato, S. and Tishby, N. Homogeneous multi-instance learning with arbitrary dependence. In *Proceeding of the 22nd Annual Conference on Learning Theory (COLT'09)*, pp. 93–104, 2009.

Wang, W. and Zhou, Z.-H. Learnability of multi-instance multi-label learning. *Chinese Science Bulletin*, 57(19): 2488–2491, 2012.

Zhang, M.-L. Disambiguation-free partial label learning. In *Proceedings of the 14th SIAM International Conference on Data Mining (SDM'14)*, pp. 37–45, 2014.

Zhou, Z.-H. and Zhang, M.-L. Multi-instance multilabel learning with application to scene classification. In *Advances in Neural Information Processing Systems 19 (NIPS'06)*, pp. 1609–1616, 2006.

Zhou, Z.-H., Zhang, M.-L., Huang, S.-J., and Li, Y.-F. Multi-instance multi-label learning. *Artificial Intelligence*, 176(1):2291–2320, 2012.