Article

Chinese Science Bulletin

Machine Learning

? 201? Vol. ? No. ?: 1–?
doi:

Learnability of Multi-Instance Multi-Label Learning

WANG Wei & ZHOU ZhiHua*

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210046, China

Received November 8, 2011

Multi-Instance Multi-Label learning (MIML) is a new machine learning framework where one data object is described by multiple instances and associated with multiple class labels. During the past few years, many MIML algorithms have been developed and many applications have been described. However, there lacks theoretical exploration to the learnability of MIML. In this paper, through proving a generalization bound for multi-instance single-label learner and viewing MIML as a number of multi-instance single-label learning subtasks with the correlation among the labels, we show that the MIML hypothesis class constructed from a multi-instance single-label hypothesis class is PAC-learnable.

machine learning, learnability, multi-instance multi-label learning (MIML)

Citation: Wang W, Zhou Z H. Learnability of multi-instance multi-label learning. Chinese Sci Bull, 201?, ?: ?–?, doi: ??

Multi-Instance Multi-Label learning (MIML) [1, 2] is a new machine learning framework. In contrast to traditional supervised learning where one data object is represented by one instance and associated with one class label, in MIML one object is described by multiple instances and associated with multiple class labels (Figure 1). Such a framework is particularly useful for handling complicated data objects with multiple semantic meanings. For example, in image annotation, an image contains many patches each can be represented by an instance, while the image can be assigned with multiple annotation terms simultaneously; in text categorization, one document contains multiple sections each can be represented by an instance, while the document can be classified into multiple categories simultaneously.

Formally, let $\mathcal{X}$ and $\mathcal{Y}$ denote the instance space and the set of class labels, respectively. The task of MIML is to learn a function $f : 2^{\mathcal{X}} \to 2^{\mathcal{Y}}$ from a given data set $\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_m, Y_m)\}$, where $X_i \subseteq \mathcal{X}$ is a set of instances $\{x_{i1}, x_{i2}, \cdots, x_{i,n_i}\}$, $x_{ij} \in \mathcal{X}$ $(j = 1, 2, \cdots, n_i)$ and Y_i is a set of labels $\{y_{i1}, y_{i2}, \cdots, y_{i,l}\}$. n_i denotes the number of instances in X_i, l denotes the number of candidate labels and $y_{ik} \in \{-1, +1\}$ $(k = 1, 2, \cdots, l)$. X_i is also called as a *bag* (of instances) and let $\mathcal{D}$ denote the distribution over the bags. $y_{ik} = +1$ means the instance x_i has the k-th label; otherwise, x_i does not have the k-th label.

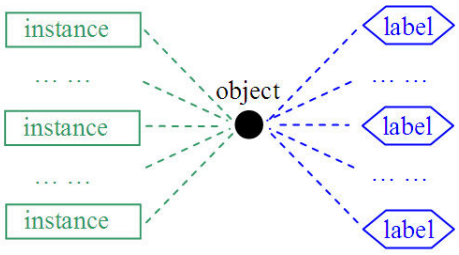

Figure 1 Illustration of the MIML framework

Many MIML algorithms have been proposed, for example, Boosting-style algorithm MimlBoost [1], SVM-style algorithm MimlSvm [1], D-MimlSvm [2] and M3Miml [3], CRF-style algorithm MlMil [4], Dirichlet-Bernoulli Alignment based algorithm Dba [5], single-label instance assumption based algorithm Sisl-Miml [6], etc. MIML techniques have also been applied to many tasks such as image annotation [1, 4, 6], text categorization [2, 3, 5], video annotation [7], bioimage informatics [8], etc.

*Corresponding author (email: zhouzh@nju.edu.cn)

This research was supported by the National Fundamental Research Program of China (2010CB327903) and the National Science Foundation of China (61073097,61021062).

Though there are significant advances in algorithms and applications, there lacks theoretical understanding of MIML. In this paper, based on Sebato and Tishiby's recent results on multi-instance single-label learning [9], we theoretically study the learnability of MIML. Our result shows that the MIML hypothesis class constructed from a multi-instance single-label hypothesis class is PAC-learnable.

1 Preliminaries

Multi-instance single-label learning, also called multi-instance learning [10, 11], can be viewed as a degenerated version of MIML where the most labels associated with a bag are neglected and only one label is concerned. Indeed, multi-instance single-label learning has been exploited as a bridge in degeneration-based MIML algorithms such as MimlBoost [1, 2]. Thus, we build our theoretical analysis by focusing on MIML hypothesis class constructed from a multi-instance single-label hypothesis class.

For the simplicity of discussion, we assume that $n_i = n$, i.e., the bags contain the same number of instances. Let $\mathcal{H}$ denote the hypothesis class where $h \in \mathcal{H}$ is a mapping from $\mathcal{X}$ to $\{-1, +1\}$. The classification rule φ_n^h over a bag $X_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$ can be defined as $\varphi_n^h(X_i) \triangleq \varphi(h(x_{i1}), \ldots, h(x_{in}))$, where $\varphi(\cdot, \cdot, \ldots, \cdot) \triangleq \max[\cdot, \cdot, \ldots, \cdot]$. Here, we consider the homogeneous multi-instance single-label learning setting, i.e., the instances in bag X_i can be classified by the same hypothesis $h \in \mathcal{H}$. Let $\varphi_n(\mathcal{H}) = \{\varphi_n^h | h \in \mathcal{H}\}$ denote the hypothesis class over bags generated from $\mathcal{H}$ by φ, d_I denote the finite VC-dimension of $\mathcal{H}$ and d_B denote the finite VC-dimension of $\varphi_n(\mathcal{H})$, based on the results in [9] it is easy to get eqs. (1-2).

$$d_B \geqslant n(d_I - 2) \tag{1}$$

$$d_B \leqslant \max\{2d_I(\log n - \log d_I + \log e), 4d_I^2, 16\} \tag{2}$$

Sabato and Tishby proposed a multi-instance single-label learning algorithm Misl in [9]. The Misl algorithm accepts as input a labeled and weighted (non-weighted) bag sample S_B and an algorithm $\mathcal{A}$ which receives a labeled and weighted (non-weighted) instance sample S_I derived from S_B and returns a hypothesis $\mathcal{A}(S_I)$ over instances. The accuracy of a hypothesis h over instances on a labeled and weighted instance sample S_I can be measured by its edge as eq. (3).

$$\Lambda(h, S_I) \triangleq \sum_{i \in |S|} w_i y_i h(x_i) / \sum_{i \in |S|} w_i, \tag{3}$$

Here $x_i \in S_I$, y_i is the label of x_i and w_i is the weight of x_i. If φ is a hypothesis over bags and S_B is a labeled and weighted bag sample, $\Lambda(\varphi, S_B)$ is defined identically except that w_i is the weight of the bag in S_B. Let h_{pos} denote the constant positive hypothesis, i.e., for any $x \in \mathcal{X}$, $h_{pos}(x) = +1$. Two hypotheses $\varphi_n^{\mathcal{A}(S_I)}$ and $\varphi_n^{h_{pos}}$ over bags are constructed by the Misl algorithm and the one which has the better edge on S_B is returned. The following Lemma 1 shows that under certain condition the Misl algorithm outputs an approximation to the optimal edge on the input bag sample [9].

Lemma 1. Let $\mathcal{H}$ denote the hypothesis class where $h \in \mathcal{H}$ is a mapping from $\mathcal{X}$ to $\{-1, +1\}$, h_M be the hypothesis returned by algorithm Misl when receiving S_B as input, $\omega \triangleq \Lambda(h_M, S_B)$ and $\omega^* \triangleq \max_{h \in \mathcal{H}} \Lambda(\varphi_n^h, S_B)$. If the following conditions hold:

i. for any instance sample S, $\Lambda(\mathcal{A}(S), S) \geqslant \max_{h \in \mathcal{H}} \Lambda(h, S)$,

ii. $\omega^* \geqslant 1 - \frac{1}{n^2}$,

then

$$\omega \geqslant \frac{n^2(\omega^* - 1) + 1}{2n - 1} \geqslant 0.$$

2 Multi-Instance Single-Label Generalization Bound

Sabato and Tishby [9] discussed the possibility of the generalization of the multi-instance single-label learning. Unfortunately, they did not provide a generalization bound. Here we derive a generalization bound by following their discussion, i.e., training number of T base classifiers $h_{M_1}, \ldots, h_{M_T}$ with algorithm Misl on the input bag sample S_B and then constructing a linear combination of the T base classifiers by using AdaBoost. To give the generalization bound in Theorem 1, we will use the following Lemmas 2 and 3 in [12].

Lemma 2. Let $\mathcal{P}$ be a distribution over $\mathcal{X} \times \{-1, 1\}$, S be a sample of m examples chosen independently at random according to $\mathcal{P}$ and $\delta > 0$. Suppose that the base classifier space $\mathcal{H}$ has VC-dimension d and that $m \geqslant d \geqslant 1$. Then with probability at least $1 - \delta$ over the random choice of the training set S, every weighted average function f in the form of

$$\{f : x \to \sum_{h \in \mathcal{H}} \alpha_h h(x) | \alpha_h \geqslant 0; \sum_h \alpha_h = 1\}$$

satisfies the following bound for all $\theta > 0$:

$$P_{\mathcal{P}}[yf(x) \leqslant 0] \leqslant P_S[yf(x) \leqslant \theta] + O\Big(\frac{1}{\sqrt{m}}\Big(\frac{d \log^2 \frac{m}{d}}{\theta^2} + \log \frac{1}{\delta}\Big)^{1/2}\Big).$$

Lemma 3. Suppose the base learning algorithm, when called by AdaBoost, generates classifiers with weighted training errors $\epsilon_1, \ldots, \epsilon_T$. Then for any θ, we have that

$$P_S[yf(x) \leqslant \theta] \leqslant 2^T \prod_{t=1}^T \sqrt{\epsilon_t^{1-\theta}(1 - \epsilon_t)^{1+\theta}}.$$

Theorem 1. Let $\mathcal{H}$ denote the hypothesis class where $h \in \mathcal{H}$ is a mapping from $\mathcal{X}$ to $\{-1, +1\}$, $h_{M_1}, \ldots, h_{M_T}$ be the hypotheses returned by algorithm Misl with T different $\mathcal{A}_1, \ldots, \mathcal{A}_T$ when receiving the bag sample S_B as input which contains m bags drawn randomly from $\mathcal{D}$. Let

d_I denote the finite VC-dimension of $\mathcal{H}$ and d_B denote the finite VC-dimension of $\varphi_n(\mathcal{H})$, where $n(d_I - 2) \leqslant d_B \leqslant \max\{2d_I(\log n - \log d_I + \log e), 4d_I^2, 16\}$. Suppose f is a linear combination constructed by using AdaBoost on T base classifiers $h_{M_1}, \ldots, h_{M_T}$. If the following conditions hold:

i. for any instance sample S, $\Lambda(\mathcal{A}_t(S), S) \geqslant \max_{h\in\mathcal{H}} \Lambda(h, S)$, $(t = 1, \ldots, T)$

ii. the multi-instance single-label learning task is separable,

then the following bound holds for $0 < \theta \leqslant \frac{1}{2n-1}$

$$error(f) \leqslant \Big(\frac{4n^{1+\theta}(n-1)^{1-\theta}}{(2n-1)^2}\Big)^{\frac{T}{2}} + O\Big(\frac{1}{\sqrt{m}}\Big(\frac{d_B \log^2(m/d_B)}{\theta^2} + \log\frac{1}{\delta}\Big)^{\frac{1}{2}}\Big).$$

Proof: Let ϵ_t $(t = 1, \ldots, T)$ denote the training error of h_{M_t} on S_B. It can be found that for binary hypothesis the edge of h on S equals $1 - 2\epsilon$ where ϵ is the error of h on S. From Lemma 1 we have eq. (4), where ϵ^* is the optimal training error on S_B.

$$1 - 2\epsilon_t \geqslant \frac{n^2(1 - 2\epsilon^* - 1) + 1}{2n - 1} \tag{4}$$

Considering that the multi-instance single-label learning task is separable, i.e., $\epsilon^* = 0$, we have $\epsilon_t \leqslant \frac{n-1}{2n-1}$ from eq. (4). So with Lemma 3 we get eq. (5) for $0 < \theta \leqslant \frac{1}{2n-1}$.

$$\begin{aligned} P_S[yf(x) \leqslant \theta] &\leqslant 2^T \prod_{t=1}^{T} \sqrt{\epsilon_t^{1-\theta}(1-\epsilon_t)^{1+\theta}} \\ &\leqslant \Big(\frac{4n^{1+\theta}(n-1)^{1-\theta}}{(2n-1)^2}\Big)^{\frac{T}{2}} \end{aligned} \tag{5}$$

Thus with eq. (5) and Lemma 2 we get Theorem 1 proved. □

Theorem 1 provides a generalization bound which depends on the number of bags m, the number of instances n in each bag, the VC-dimension d_B of multi-instance single-label hypothesis class $\varphi_n(\mathcal{H})$ and the confidence parameter δ for multi-instance single-label learning. It implies that the hypothesis class $\varphi_n(\mathcal{H})$ is PAC-learnable.

3 MIML Learnability

Let $\mathcal{H}_k \subseteq \mathcal{H}$ $(k = 1, 2, \ldots, l)$ denote the hypothesis class for multi-instance single-label learning with respect to the k-th label, where $h \in \mathcal{H}_k$ is a mapping from $\mathcal{X}$ to $\mathcal{Y}_k = \{-1, +1\}$. Let $\varphi_n(\mathcal{H}_k) = \{\varphi_n^h | h \in \mathcal{H}_k\}$ denote the hypothesis class over bags with respect to the k-th label generated from $\mathcal{H}_k$, d_{I_k} denote the finite VC-dimension of $\mathcal{H}_k$ and d_{B_k} denote the finite VC-dimension of $\varphi_n(\mathcal{H}_k)$, it is not difficult to get eqs. (6-7) with respect to eqs. (1-2).

$$d_{B_k} \geqslant n(d_{I_k} - 2) \tag{6}$$

$$d_{B_k} \leqslant \max\{2d_{I_k}(\log n - \log d_{I_k} + \log e), 4d_{I_k}^2, 16\} \tag{7}$$

Suppose we have l multi-instance single-label hypotheses $f_{M_1}, \ldots, f_{M_l}$ by using the multi-instance single-label learning algorithm and let $\varphi_n(\mathcal{H})^l$ denote the MIML hypothesis class. Now we analyze the MIML hypothesis $f_{MM} \in \varphi_n(\mathcal{H})^l$ constructed from $f_{M_1}, \ldots, f_{M_l}$ with respect to the correlation among them. Let $\widehat{Y}$ denote the predicted labels of bag X by f_{MM}, the error measured based on hamming loss can be shown as eq. (8).

$$error(f_{MM}) = E_{X\in\mathcal{D}}\Big(\frac{hloss(Y, \widehat{Y})}{l}\Big). \tag{8}$$

First, we consider the simplest case, i.e., the l candidate labels are independent to each other. In this case, MIML can be straightforwardly decomposed into l multi-instance single-label learning subtasks and the MIML hypothesis can be constructed in the form of $f_{MM} = (f_{M_1}, \cdots, f_{M_l})$. Let $error(f_{M_k})$ denote the error of f_{M_k} with respect to the k-th label, we have eq. (9) from eq. (8).

$$\begin{aligned} error(f_{MM}) &= E_{X\in\mathcal{D}}\Big(\frac{hloss(Y, \widehat{Y})}{l}\Big) \\ &= \frac{1}{l}\sum_{k=1}^{l} error(f_{M_k}) \end{aligned} \tag{9}$$

Now we give the generalization bound for f_{MM} in Theorem 2, which shows that when these l labels are independent to each other, MIML hypothesis class $\varphi_n(\mathcal{H})^l$ is PAC-learnable.

Theorem 2. Suppose the l labels are independent to each other and the conditions for the multi-instance single-label learning algorithm in Theorem 1 hold, let $d_{B_{\max}} = \max\{d_{B_1}, \ldots, d_{B_l}\}$, the following bound holds for $0 < \theta \leqslant \frac{1}{2n-1}$

$$error(f_{MM}) \leqslant \Big(\frac{4n^{1+\theta}(n-1)^{1-\theta}}{(2n-1)^2}\Big)^{\frac{T}{2}} + O\Big(\frac{1}{\sqrt{m}}\Big(\frac{d_{B_{\max}} \log^2(m/d_{B_{\max}})}{\theta^2} + \log\frac{1}{\delta}\Big)^{\frac{1}{2}}\Big).$$

Proof: With the condition that the l labels are independent to each other, the error of f_{MM} can be expressed as eq. (9). Considering Theorem 1 it is easy to get Theorem 2 proved. □

Secondly, we study the more realistic case where the l labels are not independent. Since there might exist many kinds of dependence, here we consider the following case: suppose these l labels can be divided into ϱ groups, i.e., $C_1, \cdots, C_\varrho$, and labels in the same group are positively correlated while labels in different groups are negatively correlated. In other words, if it is known that a label in C_r $(r \in \{1, \cdots, \varrho\})$ is a proper label, then the chances for other labels in C_r to be proper ones increase while the chances for labels in C_q $(q \in \{1, \cdots, \varrho\}, q \neq r)$ to be proper ones decrease. Note that, even for this case where these l labels can be divided into groups, it is still difficult to describe how to construct f_{MM} from l multi-instance single-label hypothses $f_{M_1}, \ldots, f_{M_l}$

trained by assuming that the labels are independent to each other. For the simplicity of discussion, we assume that the final f_{MM} constructed from $f_{M_1}, \ldots, f_{M_l}$ can be denoted in the form of $(f'_{M_1}, \cdots, f'_{M_l})$ and that if label $\mathcal{Y}_k \in C_r$, the positive correlation will lead to decrease the error by e_r^- for label $\mathcal{Y}_k$, while the negative correlation will lead to increase the error by $e_{r,q}^+$ for label $\mathcal{Y}_k$ $(r \neq q)$, i.e., eq. (10) holds.

$$error(f'_{M_k}) = error(f_{M_k}) - e_r^- + \sum_{q \neq r} e_{r,q}^+ \tag{10}$$

Based on these discussions and Theorem 2, we give the generalization bound for f_{MM} in Theorem 3 when these l labels are not independent to each other.

Theorem 3. Suppose the l labels satisfy above correlation assumption and the conditions for the multi-instance single-label learning algorithm in Theorem 1 hold, let $d_{B_{\max}} = \max\{d_{B_1}, \ldots, d_{B_l}\}$, the following bound holds for $0 < \theta \leqslant \frac{1}{2n-1}$

$$\begin{aligned} error(f_{MM}) \leqslant & \Big(\frac{4n^{1+\theta}(n-1)^{1-\theta}}{(2n-1)^2}\Big)^{\frac{T}{2}} \\ & + \frac{1}{l}\sum_{r=1}^{\varrho}\Big(\sum_{q \neq r}|C_q|e_{r,q}^+ - |C_r|e_r^-\Big) \\ & + O\Big(\frac{1}{\sqrt{m}}\Big(\frac{d_{B_{\max}}\log^2(m/d_{B_{\max}})}{\theta^2} + \log\frac{1}{\delta}\Big)^{\frac{1}{2}}\Big). \end{aligned}$$

Generally, for f_{M_k} where $\mathcal{Y}_k \in C_r$, the positive contribution e_r^- inside C_r may be larger than the negative contribution $\sum_{q \neq r} e_{r,q}^+$ from C_q $(q \neq r)$. So the bound in Theorem 3 is much lower than that in Theorem 2. Obviously, this implies that when the correlation among the labels is effectively exploited, MIML hypothesis class $\varphi_n(\mathcal{H})^l$ is also PAC-learnable.

4 Conclusions

In this paper, we prove that MIML hypothesis class $\varphi_n(\mathcal{H})^l$ constructed from a multi-instance single-label hypothesis class $\varphi_n(\mathcal{H})$ is PAC-learnable by showing the learnability of hypothesis class $\varphi_n(\mathcal{H})$. Note that, this paper just opens theoretical MIML study while there are many issues need a further exploration. Firstly, this study focuses on MIML hypothesis class constructed from a multi-instance single-label hypothesis class, while there are many MIML algorithms which do not rely on multi-instance single-label hypothesis. Secondly, the label correlation considered in this paper is very simple, while label correlation in real tasks is often much more complicated. These issues are well-worth studying in the future.

We thank Miao Xu for helpful discussions.

1 Zhou Z H and Zhang M L. Multi-instance multi-label learning with application to scene classification. In: Schölkopf B, Platt J C, Hoffman T, eds. Proceedings of the 20th Annual Conference on Neural Information Processing Systems, 2006 Dec 4–7, Vancouver, Canada. MIT Press, 2006. 1609–1616
2 Zhou Z H, Zhang M L, Huang S J, et al. Multi-instance multi-label learning. Artif Intell, 2012, 176(1): 2291–2320
3 Zhang M L, Zhou Z H. M3MIML: A maximum margin method for multi-instance multi-label learning. Proceedings of the 8th IEEE International Conference on Data Mining, 2008 Dec 15–19, Pisa, Italy. IEEE Computer Society, 2008. 688–697
4 Zha Z J, Hua X S, Mei T, et al. Joint multi-label multi-instance learning for image classification. Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, 2008 Jun 24–26, Anchorage, AK. IEEE Computer Society, 2008. 1–8
5 Yang S H, Zha H, Hu B G. Dirichlet-bernoulli alignment: A generative model for multi-class multi-label multi-instance corpora. In: Bengio Y, Schuurmans D, Lafferty J, et al, eds. Proceedings of the 23th Annual Conference on Neural Information Processing Systems, 2009 Dec 7–10, Vancouver, Canada. MIT Press, 2009. 2143–2150
6 Nguyen N. A new SVM approach to multi-instance multi-label learning. In: Webb G I, Liu B, Zhang C, et al, eds. Proceeding of the 10th IEEE International Conference on Data Mining, 2010 Dec 14–17, Sydney, Australia. IEEE Computer Society, 2010. 384–392
7 Xu X S, Xue X, Zhou Z H. Ensemble multi-instance multi-label learning approach for video annotation task. In: Candan K S, Panchanathan S, Prabhakaran B, et al, eds. Proceedings of the 19th ACM International Conference on Multimedia, 2011 Nov 28–Dec 1, Scottsdale, AZ. ACM, 2011. 1153–1156
8 Li Y X, Ji S, Kumar S, et al. Drosophila gene expression pattern annotation through multi-instance multi-label learning. IEEE ACM T Comput Bi, 2012, 9(1): 98–112
9 Sabato S, Tishby N. Homogenous multi-instance learning with arbitrary dependence. Proceeding of the 22nd Annual Conference on Learning Theory, 2009 Jun 18–21, Quebec, Canada. Omnipress, 2009. 93–104
10 Dietterich T G, Lathrop R H, Lozano-Pérez T, et al. Solving the multiple-instance problem with axis-parallel rectangles. Artif Intell, 1997, 89(1-2): 31–71
11 Foulds J R, Frank E. A review of multi-instance learning assumptions. Knowl Eng Rev, 2010, 25(1): 1–25
12 Schapire R E, Freund Y, Bartlett P, et al. Boosting the margin: A new explanation for the effectiveness of voting methods. Ann Stat, 1998, 26(5): 1651–1686