Applied Intelligence (2022) 52:13902–13915
https://doi.org/10.1007/s10489-022-03193-0

A theoretical analysis based on causal inference and single-instance learning

Chao Wang[1] · Xuantao Lu[1] · Wei Wang[1]

Accepted: 4 January 2022 / Published online: 28 February 2022

Abstract

Although using single-instance learning methods to solve multi-instance problems has achieved excellent performance in many tasks, the reasons for this success still lack a rigorous theoretical explanation. In particular, the potential relation between the number of causal factors (also called causal instances) in a bag and the model performance is not transparent. The goal of our study is to use the causal relationship between instances and bags to enhance the interpretability of multi-instance learning. First, we provide a lower bound on the number of instances required to determine causal factors in a real multi-instance learning task. Then, we provide a lower bound on the single-instance learning loss function when testing instances and training instances follow the same distribution and extend this conclusion to the situation where the distribution changes. Thus, theoretically, we demonstrate that the number of causal factors in the bag is an important parameter that affects the performance of the model when using single-instance learning methods to solve multi-instance learning problems. Finally, combining with a specific classification task, we experimentally validate our theoretical analysis.

Keywords Causal inference · Distribution change · Multi-instance learning · Single-instance learning

1 Introduction

Multi-instance learning (MIL) was originally used for the field of hand-printed numerals identification [1] and drug activity prediction [2]. Instead of considering a series of individually labeled instances, MIL focuses on the labels of *sets* (or called *bags*) of instances and demonstrate strong capabilities in many areas [3], e.g., speech localization [4], entity classification [5], protein structure determination [6], biometric authentication system [7–10], human pose estimation [11], medical image analysis [12], understanding chest CT imaging of COVID-19 [13], and clinical outcome prediction of COVID-19 [14].

However, the theoretical research of MIL still seriously lags behind the actual application speed. In other words, we are not very clear about some potential relationships between parameters (e.g., the number of instances in the bag) and the performance of the model. For example, users always neglect the influence of the number of instances in the bag. This makes the parameter settings of some experiments depend on the subjective intuition of the experimenters rather than interpretable principles. In most MIL tasks, we often overlook the following two issues: (1) *how does the number of positive instances in the bag affect the value of the loss function?* Most of the MIL tasks are based on a constraint premise, that the label of a bag is negative if and only if the bag does not contain any *positive* instance. However, this assumption ignores that the influence of the number of positive instances in the bag on the performance of the model. Another issue is that (2) *__t__esting instances tend to be assumed to follow the same __d__istribution as the __t__raining instances* (for the brevity of description, we refer to this assumption as the TTD). However, the TTD assumption is often violated in many real tasks [15], and whether the TTD assumption holds directly affects the performance of the model. For

✉ Wei Wang
weiwang1@fudan.edu.cn

Chao Wang
17110240038@fudan.edu.cn

Xuantao Lu
xtlu20@fudan.edu.cn

[1] Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China

example, when some real task scenarios cannot support the TTD assumption, the performance of some models will degrade [16].

Therefore, how to effectively deal with a series of MIL problems caused by the above issues has become a new research hotspot. To this end, many researchers put forward some methods in recent research, such as the approach based on the covariate shift setting [17], and test-distribution-based methods [18]. Unfortunately, the above methods rely on the prior distribution to improve the performance of the models and do not provide a rigorous theoretical explanation of the above two issues [19].

Besides, many researchers pay attention to the relation between MIL and causal inference. The advantage of using causal inference theory to study MIL is that causal inference can describe and explain the complex internal mechanism of the system by the causal relationship between data [20]. Moreover, causality helps the model make stable predictions in unknown environments [21], such as `rain causes slippery roads` and `excessive release of carbon dioxide` is one of the causes of `global warming`. This can be viewed as a stable mapping from cause to effect, therefore, a causality-based classifier is more stable than an association-based classifier [22]. Although causal inference theory opens up new opportunities for MIL problems, the mathematical principles behind some tasks have not been well clarified. Therefore, in a MIL task, we should not only focus on which instances cause changes in the bag labels (we denote the instances that affect the bag label change as *causal factors*), but also understand the possible connections between the causal factors and the performance of the model.

Recently, Zhang et al. [23] propose a novel MIL framework towards robust classification in distributional biased data. However, they neglect important metrics such as theoretical guarantees for obtaining causal factors. Feng et al. [24] attempt MIL from similar and dissimilar bags and obtain a series of performance analyses on MI classifiers and SI classifiers. However, they do not explicitly answer how the number of instances in the bag affects the model performance during the SIL method to solve MIL problems. In this paper, we combine the **s**ingle-**i**nstance **l**earning (SIL) method with the **p**otential **o**utcome **f**ramework (POF) [25] in causal inference to solve the two issues. We obtain a series of rigorous theoretical analysis results and verified our conclusions in a specific experimental task.

In brief, the contributions in this paper are summarized as follows:

- In the MIL task, we provide the lower bound on the number of samples required to determine causal factors (also called *causal instances* in [23]) within 95% confidence intervals (see Theorem 2).
- Based on the standard MI assumption [26], we capture that causal factors have a direct effect on the loss function of the single-instance learning method. We provide a lower bound on the loss function of the SIL method when the TTD assumption holds (see Theorem 3). And we extend the conclusion of Theorem 3 to the situation where the distribution changes (see Theorem 4).
- We provide a rigorous theoretical analysis of the effect of the parameters in the above theorem on model performance and validate our conclusions with an object classification experiment.

2 Related work

Distribution change In general, distribution change refers to the fact that the training and testing instances do not follow the TTD assumption, the causes of which are diverse. Ignoring the difference between the training and testing samples will lead to a decrease in the predictive ability of the model built based on the standard supervised approach. Therefore, how to solve the problem of MIL distribution change is a research hotspot [16, 27]. The covariate shift setting is a typical representative, based on which Sugiyama et al. [17] propose an estimation approach. The advantage is that this method does not rely on density estimation. Park et al. [28] propose an algorithm for calibrating prediction based on the probability of the covariate shift.

In addition, many researchers also try to resolve the problem of distribution change in MIL from different perspectives [15, 29]. For instance, Zhou et al. [30] present an effective way to analyze the case that the training instances are not independent identically distributed in MIL. Wang et al. [31] construct a new MIL-based neural network to improve its ability to diagnose diseases based on medical images in the presence of unbalanced data.

The connection between MIL and causal inference Recently, it has become a popular trend to explore the intrinsic connection between causal inference theory [20] and MIL problems [18, 32, 33]. Kuang et al. [21] develop an algorithm, called DGBR, which uses causality to achieve stable predictions. Shen et al. [22] propose an algorithm, named CRLR, which effectively improves the learning ability of the model in the presence of agnostic selection bias. Zhang et al. [23] obtain causal instances by evaluating the causal effects of the labels of bags and instances. They define a specific form of causal instance and propose a novel MIL algorithm that improves the robustness of the classifier.

The connection between SIL and MIL Approaches to solving MIL problems can be broadly classified into two categories: One class of approaches solves the MIL problem directly at the bag-level or instance-level. Another type of approach uses SIL methods to solve MIL problems (this paper focuses on this type of approach). We briefly summarize some recent related work in Table 1 that demonstrates the advantages and disadvantages of SIL and MIL for different tasks.

Although a large number of prior studies demonstrate the performance of MIL and SIL separately in different tasks (e.g., Table 1), most of the conclusions are summarized by the results of specific experiments, while few studies critically analyze the rationale behind solving MIL problems using SIL methods at the theoretical level [16, 23, 24]. Inspired by the above work, in this paper, we first provide a lower bound on the number of samples required to identify causal factors using causal inference theory. Subsequently, we demonstrate the superiority and limitations of the SIL approaches to solving MIL problems by rigorous theoretical analysis. Finally, we extend our conclusions to the case of distribution change. Our conclusions reveal to a certain extent why SIL methods can effectively solve MIL problems.

3 Preliminaries

In this section, the key notations and some basic notions about multi-instance learning, causal factor are reviewed. Table 2 summarizes the key notations commonly used in this paper and their descriptions.

Definition 1 ([26] **Multi-instance Learning (MIL)**) Let finite set $\mathcal{X}_{ins} = \{x\}$ denote the space consisting of all instances x, and let $\mathbb{N}(\mathcal{X}) = \{o(x)\}$, where $o(x)$ be the function that can be described by $o(x) : \mathcal{X}_{ins} \rightarrow \mathbb{N}$. The goal of MIL can be formalized as learning the function $\text{map}_{MIL} : \mathbb{N}(\mathcal{X}) \rightarrow \mathcal{Y}$, where $\mathcal{Y}$ represents the label set. In particular, the binary MIL problem can be described as:

$$\text{BIN}_{MIL} : \mathbb{N}(\mathcal{X}) \rightarrow \{Y_+, Y_-\}.$$

In our work, we only consider the binary MIL problem, unless otherwise specified. To simplify the presentation, for any finite set, we use $\{\cdot\}_n$ to denote the set containing n elements, i.e., $\{x_1, x_2, ..., x_n\} \triangleq \{x_i\}_n$, equivalently, $|\{x_i\}_n| = n$ (where '$|\cdot|$' denotes the cardinal number of the set). Let $\mathcal{B} = \{B_i\}_r$ be the set containing r pairs $(B_i, Y_{+/-})$, where $B_i \triangleq \{x_i\}_n$ represents that a bag contains n instances.

Table 1 A brief summary of the work related to MIL and SIL under different tasks

Method	Task	Description
SIL	Gene expression and text categorization [34]	SI classifiers outperform MI classifiers when there is not enough data to train a bag-level classifier.
SIL	Object class recognition and drug activity prediction [35]	MILES transforms MIL into the SIL, which improves classification accuracy and robustness to labeling uncertainty without satisfying the standard assumptions.
SIL	Prediction on agnostic test data[21]	The classifier constructed based on causality can achieve stable prediction for unknown test environments.
SIL	Classification tasks under positive instance sparsity [36, 37]	SI classifiers are inferior to MI classifiers when the bag with positive labels contains fewer positive instances.
SIL	Large-scale MIL problems[38]	As an extended version of [35], [38] still follows the transformation of MIL problems into SIL learning to solve and enable large-scale MI data efficiently.
MIL	MIL with key instance shift [15]	When the training and test instances do not follow the i.i.d. assumption, directly applying the SI distribution change method to MIL does not effectively improve the performance of the model.
MIL	Robust classification in distributional biased data [23]	[23] proposes a MIL framework that does not require access to unlabeled test data, based on the potential outcome framework in causal inference.
MIL	Medical diagnosis [31]	MIL can take advantage of incomplete, fragmented information in the model to generate reliable diagnostic results.
MIL	MI classification while ignoring standard assumptions [39]	Linear programming procedures can perform multi-instance classification tasks with adjusting instance contributions while ignoring standard assumptions.
MIL	Learning from similar and dissimilar bags [24, 40]	[24] learns from similar and dissimilar bags and obtains a series of performance analyses on MI classifiers and SI classifiers.

Table 2 Key notations and descriptions

Notation	Description
$\emptyset$	the empty set
$\mathcal{X}_{ins} = \{x\}$	the finite set of instance x
$\hat{x}$	the causal instances
x_{cnf}	the non-causal instances
$\mathcal{Y}$	label set, in this paper, we set $\mathcal{Y} = \{Y_+, Y_-\}$
B_i	the i-th bag
$\mathcal{B} = \{B_i\}_r$	the set containing r pairs $(B_i, Y_{+/-})$
$B(+)$	the set of all bags with positive label "+"
$B(-)$	the set of all bags with negative label"+"
$B_i^+ \in B(+)$	the i-th bag in $B(+)$, "+" means the bag label is "+"
$B_i^- \in B(-)$	the i-th bag in $B(-)$, "-" means the bag label is "-"
$\lvert \cdot \rvert$	the cardinal number of the set, e.g., $\lvert\emptyset\rvert = 0$, $\lvert\{a, b, c\}\rvert = 3$
# positive/total instances	the number of positive/total instances
$\{x_i\}_n$	the abbreviated form of the set $\{x_1, x_2, ..., x_n\}$
ψ	the labeling function of the bag
$(\cdot \vee \cdot)$	the Boolean OR function
$\xi(\cdot)$	the binary classification function
Ψ^+	the collection of causal instances from all B^+
Λ^+	the collection of non-causal instances from all B^+
Λ^-	the collection of non-causal instances from all B^-
$\Delta(x)$	the causal effect of x on the bag label in Theorem 1
$(1 - \alpha_c)$	the confidence interval in Theorem 2
$\mathbb{E}(\cdot)$	the expected value of the bag label
$\mathbb{H}(\cdot)$	the Heaviside step function in Theorem 3 and Theorem 4
$\mathbb{L}(x)$	the loss function in Theorem 3 and Theorem 4

In this paper, we assume that $|B_i| = |B_{j\neq i}|$, which is a reasonable extension. For example, for any B_i^-, the label of the bag has no relation with the number of instances inside the bag. We use $\mathcal{B}(+) = \{B_i^+\}_{k_+}$ and $\mathcal{B}(-) = \{B_i^-\}_{k_-}$ to represent that the set of all bags labeled Y_+ and Y_-, respectively. Given a bag $B_i = \{x_i\}_n$, and $\phi : \mathcal{X}_{ins} \to Y_{i,i\in\{+,-\}}$, the MIL function $\text{BIN}_{MIL} \triangleq \psi(\cdot)$ can be equivalently described in the following form [26]:

$$\psi(B_i) = (\cdots \vee \phi(x_i) \vee \phi(x_{i+1}) \vee \cdots)_n. \tag{1}$$

Where ψ is denoted as the labeling function of the bag and '$(\cdot \vee \cdot)$' is the Boolean OR function.

Definition 2 ([23] **Causal Factor)** If $\exists\, x \in B^+$, such that

$$\psi(x \cup B^-) = Y_+, \psi(B^-) = Y_-. \tag{2}$$

Then the instance x is denoted as causal factor $\hat{x}$.

By Definition 2, instances in the bag can now be divided into two categories: One category is the set consisting of *causal factors* (denoted as $\{\hat{x}\}$, we use *causal factors* instead of *causal instances* here to emphasize the importance of these instances in the MIL task because $\hat{x}$ is the unique factor that causes the bag to be labeled Y_+) and the other category is the set consisting of *non-causal instances* (denoted as $\{x_{ncf}\}$). Obviously, the instance x can be formalized as:

$$x = \begin{cases} \hat{x}, & \text{if } \psi(x \cup B^-) = Y_+ \\ x_{ncf}, & \text{otherwise.} \end{cases} \tag{3}$$

The causal factors ensure that $\psi(B) = Y_+$ holds for any bag B that contains them, and the non-causal instances do not affect the label of the bag. For example, we imagine such an image, a `cat` playing on the `lawn` (i.e., Fig. 1). The classifier will label images based on the presence or absence of the `cat` in the image. If there is at least one `cat` in the image, the classifier outputs Y_+, otherwise, the classifier outputs Y_-. Apparently, the `cat` in the image is a causal factor, while the rest, such as `lawn`, `flower`, `bamboo basket`, etc., are non-causal instances, and these non-causal instances do not affect the output of an oracle classifier. Therefore, in this paper, we follow the stander multi-instance assumption, i.e., the bag is labeled Y_-, if and only if the bag does not contain any causal factor.

Fortunately, causal factors can be obtained by estimating the causal effect of an instance on the label of the bag. Specifically, for the label Y, the causal effect of an

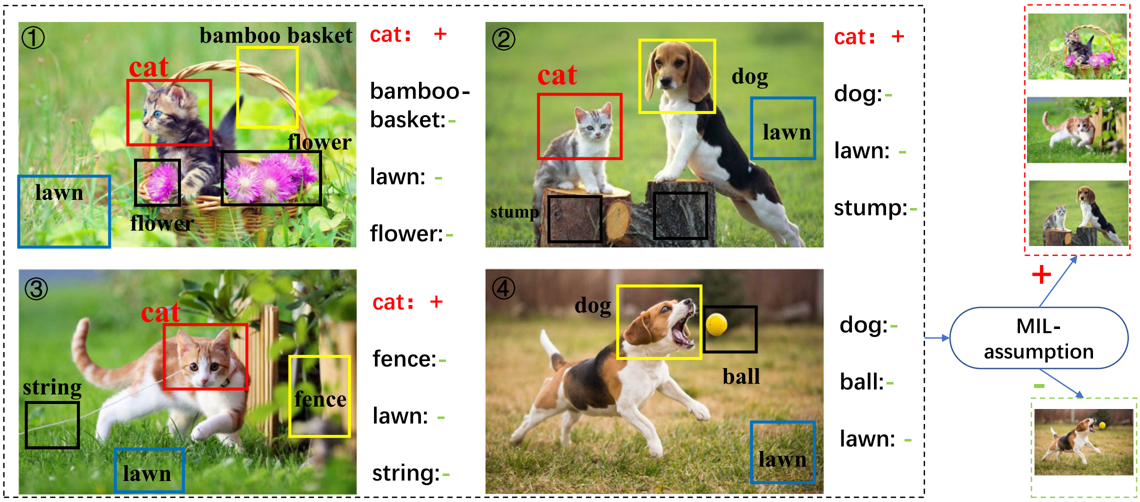

Fig. 1 (1) An example of the standard multi-instance assumption. In this example, the image is considered as a bag and the `cat` and other objects are considered as instances in the bag. As long as the image has a `cat` in it, the label of the image is positive (e.g., ①②③), and if the image does not contain any `cat`, the label of the image is negative (e.g., ④). (2) In the first picture on the left, the `cat` is the causal factor, `while flower`, `bamboo basket`, and `lawn` are non-causal instances

intervention (e.g., adding or removing a candidate instance x to or from a bag) is a comparison of the potential outcomes of the two label states under the POF [23]. Inspired by the application of causal inference to MIL tasks, in the next section, we aim to capture the connection between causal factors and MIL tasks.

4 Theoretical analysis based on causal inference

In this section, we focus on three main problems. Specifically, in Section 4.1, we obtain a minimum number of candidate instances required to determine the causal factor. In Section 4.2, we analyze the connection between causal factors and the SIL loss function. In Section 4.3, we analyze the effect of parameters on the decision threshold of the classifier.

4.1 Determining the causal factors by estimating causal effects

In this section, we determine whether an instance x is a causal factor by estimating the ***a****verage* ***c****ausal* ***e****ffect* (ACE) [41] (also called ***a****verage* ***t****reatment* ***e****ffect* (ATE), denoted as $\Delta(x)$) on the bag label. Specifically, let $Y_{T(x\in B)}$ be the potential label of bag B if $x \in B$. $T(x \in B)$ can be viewed as a **t**reatment of B. (i.e., adding x to B). Similarly, let $Y_{T(x\notin B)}$ be the potential label of bag B if the instance x is not present in B. Therefore, the $\Delta(x)$ can be defined as:

$$\Delta(x) = \mathbb{E}[Y_{T(x\in B)}] - \mathbb{E}[Y_{T(x\notin B)}], \tag{4}$$

where $\mathbb{E}(\cdot)$ represents the expected value of the bag label.

Note that $T(x \in B)$ can be obtained by adding the candidate instance to B (we use $\mathcal{C}^+ = \{x|x \in B^+\} = \{x_i\}_q$ to denote the *candidate set* and the $\hat{x}$ to denote the causal factor in $\mathcal{C}^+$). Similarly, $T(x \notin B)$ can be obtained by removing the candidate instance from bag B. Therefore, in the MIL task, (4) can be equivalently described as:

$$\Delta(x) = \mathbb{E}[\widehat{Y}|B_{T(x\in B)}] - \mathbb{E}[\widehat{Y}|B_{T(x\notin B)}], \tag{5}$$

where $\widehat{Y}$ be the new label of bag B after being treated. Note that if the performance of the classifier is good enough (e.g., a perfect classifier), (5) can assist the classifier to effectively distinguish whether an instance x is a causal factor or a non-causal instance. The specific conclusion is shown in Theorem 1.

Theorem 1 ([23] ***Causal effect of instance*** x ***on label*** $Y_{+/-}$) *Given an instance* x*, the estimated value of* $\Delta(x)$ *can be approximated as:*

$$\Delta(x) \approx \mathbb{E}[\widehat{Y}|Y_B = Y_-, B_{x\in B}] \cdot \Pr(Y_B = Y_-) \tag{6}$$

This theorem is elaborated upon in [23]. Theorem 1 describes the causal relationship between the label of the bag and the instances it contains under an ideal experimental situation.

The key is how to estimate the value of $\mathbb{E}[\widehat{Y}|Y_B = Y_-, B_{x\in B}]$ in actual tasks. An intuitive idea is to traverse all the instances, and finally obtain the value of $\Delta(x)$. However, even in a relatively small dataset, the cost of this calculation is very huge. Therefore, a certain number of samples are usually collected to estimate the value of $\Delta(x)$.

For example, we use the sample average to estimate the first term in Theorem 1 (i.e., $\mathbb{E}[\widehat{Y}|Y_B = Y_-, B_{x\in B}]$) as:

$$\widehat{\Delta(x)} = \frac{1}{q}\sum_{i=1}^{q} \xi(B_i^-(x_i)), \tag{7}$$

where $\xi(\cdot)$ is a classifier with $B_i^-(x_i)$ ($B_i^-(x_i) = \{B_i^- \cup x_i\}, x_i \in \mathcal{C}^+, i = \{1, 2, ..., q\}$) as input and a label as output. The core idea of this approach is to determine whether x_i is a causal factor by using the label of $B_i^-(x_i)$. Because if x_i is (not) a causal factor, then the ideal classifier will predict the label of $B_i^-(x_i)$ as 1 (0). In the MIL task, we can determine which instances are causal factors by estimating (7). For example, by setting the threshold t, and we select the higher-scoring x as the causal factor $\hat{x}$ (this goal can be achieved by Algorithm 1 proposed in [23]).

However, [23] ignores a crucial constraint on the $\mathcal{C}^+$. In other words, to obtain a valuation of (7) within an appropriate confidence interval, we need to determine how many instances the candidate set $\mathcal{C}^+$ contains at least. To achieve the determination of causal factors at the lowest possible cost. We prove a lower bound of the samples required to estimate (7), which is the minimum value of $q = |\mathcal{C}^+|$ of the candidate set, by the following theorem.

Theorem 2 ***(Minimum sample cost of obtaining causal factors within 95% confidence interval)*** *In an ideal MIL task, the sample lower bound for acquiring the 95%-confidence interval at a sub-linear cost is* $q \geq \frac{\log 40}{2\epsilon^2}$.

Proof Given the instance $x_i \in \mathcal{C}^+$, we assume $\mathbb{E}[\frac{1}{q}\sum_{i=1}^{q} \xi(B_i^-(x_i))] = \delta_c$. Utilizing Hoeffding's inequality [42], we have

$$\Pr\left(\frac{1}{q}\sum_{i=1}^{q} \xi(B_i^-(x_i)) - \delta_c \geq \epsilon\right) \leq e^{-2\epsilon^2 q}. \tag{8}$$

Furthermore, (8) can be extended to the form of the two-sided variant as follows [43]:

$$\Pr\left(\left|\frac{1}{q}\sum_{i=1}^{q} \xi(B_i^-(x_i)) - \delta_c\right| \geq \epsilon\right) \leq 2e^{-2\epsilon^2 q}. \tag{9}$$

Let $[\delta_c - \epsilon, \delta_c + \epsilon]$ be the confidence interval and let α_c be the level of significance for $[\delta_c - \epsilon, \delta_c + \epsilon]$ (i.e., the probability of making an error), then we have

$$\alpha_c = \Pr\left(\frac{1}{q}\sum_{i=1}^{q} \xi(B_i^-(x_i)) \notin [\delta_c - \epsilon, \delta_c + \epsilon]\right) \leq 2e^{-2\epsilon^2 q}. \tag{10}$$

Solving the above inequality, we require at least $q \geq \log(\frac{2}{\alpha_c})/2\epsilon^2$ samples to acquire $(1 - \alpha_c)$ confidence interval $[\delta_c - \epsilon, \delta_c + \epsilon]$. In particular, let $\alpha_c = 0.05$, we have $q \geq \frac{\log 40}{2\epsilon^2}$. □

The above theorem provides the cost of acquiring the confidence interval. Therefore, we can use Theorem 2 to obtain a lower bound (i.e., $\frac{\log 40}{2\epsilon^2}$) on the number of samples (i.e., q) used to determine the causal factor $\hat{x}$ within 95% confidence interval.

4.2 Connection between causal factor and loss function

Using the SIL approach to solve MIL problems is a popular way [21]. Therefore, in this section, we follow the framework proposed by [23] for further theoretical analysis of MIL using the SIL approach. Specifically, we will explore the potential connection between the number of causal factors in the bag and the loss function of the SIL method. Please note that in our subsequent analysis, we do not care about the specific experimental details of the model in determining causal factors (this problem is elaborated upon in [23]). Therefore, we assume that the model has determined the causal factor $\hat{x}$, i.e., there exists an ideal classifier $\xi(\cdot)$ to determine causal factors.

The core of our study is to describe the effect of the number of causal factors in the bag on the loss function in different situations where the TTD assumption holds or not. And further explains why using SIL methods to study MIL problems is an effective tool. Specifically, in Theorem 3, based on the TTD assumption, we use rigorous mathematical language to perfect the conclusion in [16]. Furthermore, we extend the conclusion of Theorem 3 to the case where the distribution changes (see Theorem 4). These conclusions help us to capture more thoroughly the connection between data distribution, causal factors, and loss function.

For a clearer description, we complement some notations and their descriptions. Let (B_i^+, B_i^-) denote the pair of B_i^+ and B_i^-, and one B_i^+ corresponds to one B_i^-. Let Ψ^+ denote the collection of causal factors from positive bags, i.e.,

$$\Psi^+ = \{\hat{x}|\hat{x} \in B^+\}, \text{ and } |\Psi^+| = \tau_c^+, \tag{11}$$

where '+' indicates that the label of the bag is Y_+, the subscript 'c' means that the causal factors. Similarity, let

$$\Lambda^- = \{x_{ncf}|x_{ncf} \in B^-\} \text{ and } |\Lambda^-| = \tau_n^-, \tag{12}$$

where '−' indicates that the label of the bag is Y_-, the subscript 'n' represents the non-causal instances.

For the sake of computational simplicity, we assume that all bags contain the same number of instances, i.e., $|B_i| = |B_{i\neq j}|$. Note that, according to the definition of causal factor, for any B_i^+, no matter whether there are one or more causal factors in B_i^+, the label of bag B_i^+ is always Y_+.

Therefore, we set each bag B_i^+ to have the same number of causal factors (This setting is just to simplify the theoretical calculations, and in real experimental scenarios in Section 5, we estimate the proportion of causal factors in each B^+). Formally, let $P_i^+ \subseteq \Psi^+$ ($P_j^+ \subseteq \Psi^+$) denote the subset of all causal factors in B_i^+ (B_j^+), we can set $|P_i^+| = |P_j^+| < \tau_c^+$. Similarly, the number of remaining non-causal instances in B_i^+ is defined as $|N_i^+|$. Since B^- does not contain any causal factors, i.e., $P_i^- = \emptyset$ ($P_i^- \subseteq \Lambda^-$ represents the subset of all causal factors in Λ^-), based on the above analysis, we have $|B_i^-| = |B_j^-| < \tau_n^-$.

This is a reasonable extension of n in $B = \{x_i\}_n$. Specifically, for B_i^+, no matter how many non-causal instances are included in B_i^+, they do not affect $Y_{B_i^+} = Y_+$. Similarly, for B_i^-, no matter how many non-causal instances are included in B_i^-, they do not affect $Y_{B_i^-} = Y_-$.

Let

$$\Lambda^+ = \{x_{ncf} | x_{ncf} \in B^+\} \tag{13}$$

be the set of non-causal instances in the B^+, and $|\Lambda^+| = \tau_n^+$, where '+' means the label of the bag is Y_+, the subscript 'n' means the non-causal instances. Intuitively, we have that

$$|\Psi^+| + |\Lambda^+| = |\Lambda^-| \text{ and } |\Lambda^+| = \tau_n^+ = \tau_n^- - \tau_c^+. \tag{14}$$

Assuming that the instances in bag B^+ contain causal factors and non-causal instances, we have that for each B_i^+, $|B_i^+| = |P_i^+| + |N_i^+|$ holds. To sum up, we have the following assumptions, i.e.,

$$|\Lambda^+| = \tau_n^+, |\Lambda^-| = \tau_n^-, |\Psi^+| = \tau_c^+, \tau_c^+ + \tau_n^+ = \tau_n^- . \tag{15}$$

Based on the above analysis, in the next, we consider a situation that the instances in Λ^+ and Λ^- follow the TTD assumption. Based on the TTD assumption, we obtain the lower bound of the SIL loss function.

Theorem 3 ***(Minimal value of the SIL loss function when the IID assumption holds)*** *Let $\xi(\cdot)$ denote the Heaviside step function such that $\xi(x \in \Psi^+) = 1$, and $\xi(x \notin \Psi^+) = 0$. If the instances in Λ^+ and Λ^- follow TTD assumption, then the SIL loss function $\mathbb{L}(x)$ can be minimized by linear function ξ_m of ξ such that*

$$\inf\{\mathbb{L}(x)\} = -\log\left(\beta^{\tau_n^+} \cdot (1-\beta)^{\tau_n^-}\right). \tag{16}$$

where $\beta = \tau_n^+ \cdot (\tau_n^+ + \tau_n^-)^{-1}$.

Proof First, we introduce the Heaviside step function[1] $\xi(\cdot)$ and a linear function $\xi_m(\cdot)$ of $\xi(\cdot)$. Since $\xi(x) = 1$ if and only if $x \in \Psi^+$, i.e., $\xi(\hat{x}) = 1$, $\xi(\cdot)$ can be denoted as $\xi(\cdot) \triangleq \mathbb{H}[\cdot]$, where $\mathbb{H}[\cdot]$ be the Heaviside step function. Given the linear function

$$\xi_m(x) = \alpha\xi(x) + \beta(1 - \xi(x)), \tag{17}$$

where α is the weight of the causal factors in the B^+ and β is the weight of non-causal instances. Intuitively, for α, since all causal factors are only in B^+, $\alpha = 1$ in (17). For β, we have at least two alternative forms for the value of β, which are

$$\beta = \begin{cases} \beta_1 = |\Lambda^+| \cdot (|\Lambda^+| + |\Lambda^-|)^{-1} \\ \beta_2 = |\Lambda^-| \cdot (|\Lambda^+| + |\Lambda^-|)^{-1} \end{cases} \tag{18}$$

Next, we need to determine β (whether $\beta = \beta_1$ or $\beta = \beta_2$) by calculating the extreme value of loss functions $\mathbb{L}(x)$ [16],

$$\begin{aligned} \mathbb{L}(x) &= -\sum_{x=\hat{x}} \log \xi_\sigma(x) \\ &\quad - \left(\sum_{x\in\Lambda^+} \log \xi_\sigma(x) + \sum_{x\in\Lambda^-} \log(1 - \xi_\sigma(x)) \right). \end{aligned} \tag{19}$$

We first determine the specific form of $\xi(\cdot)$. Assuming that $\xi(\cdot)$ be a composite function about $\xi_\sigma(\cdot)$,

$$\begin{aligned} &\xi(\cdot) \triangleq \mathbb{H}(\xi_\sigma(\cdot)) \\ &\text{s.t.} \begin{cases} \xi(x) = \xi_\sigma(x) = 1, & \text{if } x = \hat{x} \\ \xi(x) = 0, 0 < \xi_\sigma(x) < 1, & \text{otherwise.} \end{cases} \end{aligned} \tag{20}$$

where $0 < \xi_\sigma(x) < 1$ ensures that the loss function $\mathbb{L}(x)$ has extreme value in the interval $(0, 1)$.

According to the definition of $\xi(\cdot)$, (20) can be equivalently described as $\sum_{x=\hat{x}} \log \xi_\sigma(x) = 0$. $\xi_\sigma(x)$ can be expressed as a scalar that follows the distribution Dis (denoted as $\xi_\sigma \sim Dis$). Therefore, $\mathbb{L}(x)$ can be represented as:

$$\begin{aligned} -\mathbb{L}(x) &= |\Lambda^+| \cdot \left(\frac{\sum_{x\in\Lambda^+} \log \xi_\sigma(x)}{|\Lambda^+|} \right) \\ &\quad + |\Lambda^-| \cdot \left(\frac{\sum_{x\in\Lambda^-} \log(1 - \xi_\sigma(x))}{|\Lambda^-|} \right). \end{aligned} \tag{21}$$

Assuming that $\sum_i \log \xi_\sigma(x_i)$ be a large sample data, thus we can use the expected value of data to replace the average of $\sum_i \log \xi_\sigma(x_i)$ as:

$$\frac{1}{i} \sum_i \log \xi_\sigma(x_i) \approx \mathbb{E}_{\xi_\sigma \sim Dis}[\log \xi_\sigma(x_i)]. \tag{22}$$

Then we have

$$\begin{aligned} -\mathbb{L}(x) &= |\Lambda^+| \cdot \mathbb{E}_{\xi_\sigma \sim Dis} \log \xi_\sigma(x) \\ &\quad + |\Lambda^-| \cdot \mathbb{E}_{\xi_\sigma \sim Dis} \log(1 - \xi_\sigma(x)) \\ &= \mathbb{E}_{\xi_\sigma \sim Dis} \big[|\Lambda^+| \cdot \log \xi_\sigma(x) \\ &\quad + |\Lambda^-| \cdot \log(1 - \xi_\sigma(x)) \big]. \end{aligned} \tag{23}$$

[1] The Heaviside function may be defined as a piecewise function, which can describe the ideal binary classification.

To simplify the presentation, we let

$$h(\xi_\sigma(x)) = |\Lambda^+| \cdot \log \xi_\sigma(x) + |\Lambda^-| \cdot \log(1 - \xi_\sigma(x)). \quad (24)$$

Next, we obtain the extreme value $h(\xi_\sigma(x))$ by solving the derivative $h'(\xi_\sigma(x))$. After a simple calculation, we obtain that

$$\begin{aligned} & h(\xi_\sigma(x))_{max} = h\left(\tfrac{|\Lambda^+|}{|\Lambda^+|+|\Lambda^-|}\right) = h(\beta_1) \\ & \text{s.t. } 0 < \xi_\sigma(x) < 1. \end{aligned} \quad (25)$$

Therefore, the coefficient β in $\xi_m(x)$ can be determined to be β_1.

Next, we calculate the minimum value of $\mathbb{L}(x)$. Specifically,

$$\begin{aligned} \min\{\mathbb{L}(x)\} &= -\mathbb{E}_{\xi_\sigma \sim Dis}\left[h(\xi_\sigma(x))\right]_{\xi_\sigma(x)=\beta_1} \\ &= -\mathbb{E}_{\xi_\sigma \sim Dis}\left[|\Lambda^+| \cdot \log \beta_1 + |\Lambda^-| \cdot \log(1-\beta_1)\right] \\ &= -\mathbb{E}_{\xi_\sigma \sim Dis}\left[\tau_n^+ \cdot \log \beta_1 + \tau_n^- \cdot \log(1-\beta_1)\right] \\ &= -\mathbb{E}_{\xi_\sigma \sim Dis}\left[\log\left(\beta_1^{\tau_n^+} \cdot (1-\beta_1)^{\tau_n^-}\right)\right]. \end{aligned} \quad (26)$$

Obviously, all parameters in $\mathbb{E}_{\xi \sim Dis}[\cdot]$ are constants, thus (26) is equivalent to

$$\inf\{\mathbb{L}(x)\} = -\log\left(\beta_1^{\tau_n^+} \cdot (1-\beta_1)^{\tau_n^-}\right), \quad (27)$$

which proves the theorem. □

In particular, we consider an extreme case where there is only one causal factor in the bag, i.e., $\tau_c^+ = 1$. According to our previous assumption, there is $\tau_n^+ \approx \tau_n^-$, Thus, we obtain a minimum value of $\mathbb{L}(x)$ is $\tau_n^- \log 4$.

Based on the TTD assumption, Theorem 3 shows that the function $\xi_m(\cdot)$ can optimize the loss objective function $\mathbb{L}(x)$ to a minimum value. However, the TTD assumption is often violated in most real-world applications. Therefore, we extend the conclusion of Theorem 3 to the situation where the distribution changes. The conclusion is shown in Theorem 4.

Theorem 4 ***(Minimal value of the SIL loss function when the TTD assumption does not hold)*** *Let $\xi(\cdot)$ denote a Heaviside step function such that $\xi(x \in \Psi^+) = 1$, and $\xi(x \notin \Psi^+) = 0$. If instances in Λ^+ and Λ^- are drawn from the different distributions $\mathcal{D}_{\Lambda^+}(x)$ and $\mathcal{D}_{\Lambda^-}(x)$, respectively, then the SIL loss function $\mathbb{L}(x)$ can be minimized by linear function ξ_m of ξ such that*

$$\begin{aligned} \inf\{\mathbb{L}(x)\} = & -\mathop{\mathbb{E}}_{\xi_\sigma \sim \widehat{Dis}}\left[\tau_n^+ \cdot \delta \cdot \log(\beta_1 \cdot \delta)\right. \\ & \left. +(\tau_n^+ + \tau_n^- - \tau_n^+ \cdot \delta) \cdot \log(1 - \beta_1 \cdot \delta)\right]. \end{aligned} \quad (28)$$

Where $\widehat{Dis} = \beta_1 \cdot \mathcal{D}_{\Lambda^+}(x) + \beta_2 \cdot \mathcal{D}_{\Lambda^-}(x)$, and $\mathcal{D}_{\Lambda^+}(x) = \delta \cdot \widehat{Dis}$.

Proof According to Theorem 3 and (22), we know that

$$\begin{aligned} & -\mathbb{L}(x) = \mathbb{E}_{\xi_\sigma \sim Dis}\left[|\Lambda^+| \cdot \log \xi_\sigma(x) + |\Lambda^-| \cdot \log(1 - \xi_\sigma(x))\right], \\ & \text{s.t. } \xi_\sigma(x = \hat{x}) = 1. \end{aligned} \quad (29)$$

Since $\mathcal{D}_{\Lambda^+}(x) \neq \mathcal{D}_{\Lambda^-}(x)$, we assume that the new distribution $\widehat{Dis}$ is an affine combination of distribution $\mathcal{D}_{\Lambda^+}(x)$ and $\mathcal{D}_{\Lambda^-}(x)$, i.e.,

$$\begin{aligned} & \widehat{Dis} = \beta_1 \cdot \mathcal{D}_{\Lambda^+}(x) + \beta_2 \cdot \mathcal{D}_{\Lambda^-}(x) \\ & \text{s.t. } \beta_1 + \beta_2 = 1, \end{aligned} \quad (30)$$

where β_1 and β_2 are the coefficients in (18).

Given a Heaviside step function $\xi(\cdot) \triangleq \mathbb{H}[\xi_\sigma(\cdot)]$, which implies that

$$\Psi^+ \cap (\Lambda^+ \cup \Lambda^-) = \emptyset. \quad (31)$$

According to (17) in Theorem 3, we can determine that the function $\xi_m(\cdot)$ has the following form,

$$\xi_m(x) = \widehat{\alpha}\xi(x) + \widehat{\beta}(1 - \xi(x)),$$

where $\widehat{\alpha} = 1$(because all causal factors are only in B^+). Similar to the setting of parameters β_1 and β_2 in Theorem 3, we assume that

$$\widehat{\beta} = \begin{cases} \widehat{\beta}_1 = \frac{\tau_n^+ \cdot \mathcal{D}_{\Lambda^+}(x)}{\tau_n^+ \cdot \mathcal{D}_{\Lambda^+}(x) + \tau_n^- \cdot \mathcal{D}_{\Lambda^-}(x)}, \text{ or} \\ \widehat{\beta}_2 = \frac{\tau_n^- \cdot \mathcal{D}_{\Lambda^-}(x)}{\tau_n^+ \cdot \mathcal{D}_{\Lambda^+}(x) + \tau_n^- \cdot \mathcal{D}_{\Lambda^-}(x)} \\ \widehat{\beta}_1 + \widehat{\beta}_2 = 1. \end{cases} \quad (32)$$

According to (30), it is not difficult to find that

$$\widehat{\beta}_1 = \beta_1 \cdot \frac{\mathcal{D}_{\Lambda^+}(x)}{\widehat{Dis}(x)} \text{ and } \widehat{\beta}_2 = \beta_2 \cdot \frac{\mathcal{D}_{\Lambda^-}(x)}{\widehat{Dis}(x)}. \quad (33)$$

Note that the coefficient $\widehat{\beta}$ is a coefficient function $\widehat{\beta} \triangleq \beta(\widehat{Dis})$ about the distribution $\widehat{Dis}$. Therefore, $\xi_m(\hat{x}) = 1$ and $\xi_m(x \notin \Psi^+) = \widehat{\beta}_1$ or $\widehat{\beta}_2$.

According to (22) and the $\widehat{Dis}$ contains the distribution information of $\mathcal{D}_{\Lambda^+}(x)$ and $\mathcal{D}_{\Lambda^-}(x)$, we have

$$\begin{aligned} -\mathbb{L}(x) &= \mathbb{E}_{\xi_\sigma \sim \widehat{Dis}}\left[|\Lambda^+| \frac{\mathcal{D}_{\Lambda^+}(x)}{\widehat{Dis}(x)} \log \xi_\sigma(x)\right. \\ &\qquad \left. + |\Lambda^-| \frac{\mathcal{D}_{\Lambda^-}(x)}{\widehat{Dis}(x)} \log(1 - \xi_\sigma(x))\right] \\ &= \mathbb{E}_{\xi_\sigma \sim \widehat{Dis}}\left[h_1(\xi_\sigma(x))\right]. \end{aligned} \quad (34)$$

We obtain the extreme value of the function $h_1(\xi_\sigma(x))$ by solving the derivative $h_1'(\xi_\sigma(x))$. For a determined instance x, $\frac{\mathcal{D}_{\Lambda^+}(x)}{\widehat{Dis}(x)}$ and $\frac{\mathcal{D}_{\Lambda^-}(x)}{\widehat{Dis}(x)}$ can be regarded as ratio constants, the extreme solution of $h_1(\xi_\sigma(x))$ is

$$\begin{aligned} h_1(\xi_\sigma(x))_{max} &= h_1\left(\frac{|\Lambda^+| \cdot \mathcal{D}_{\Lambda^+}(x)}{(|\Lambda^+| + |\Lambda^-|) \cdot \widehat{Dis}(x)}\right) \\ &= h_1\left(\beta_1 \frac{\mathcal{D}_{\Lambda^+}(x)}{\widehat{Dis}(x)}\right) = h_1(\widehat{\beta}_1) \\ &\quad \text{s.t. } 0 < \xi_\sigma(x) < 1. \end{aligned} \quad (35)$$

Therefore, the coefficient $\widehat{\beta}$ in the function $\xi_m(x)$ can be determined to be $\widehat{\beta_1}$.

Furthermore, according to (30), we have that

$$1 = \beta_1 \cdot \frac{\mathcal{D}_{\Lambda^+}(x)}{\widehat{Dis}(x)} + \beta_2 \cdot \frac{\mathcal{D}_{\Lambda^-}(x)}{\widehat{Dis}(x)}. \tag{36}$$

To make the expression more concise, we set $\frac{\mathcal{D}_{\Lambda^+}(x)}{\widehat{Dis}(x)} = \delta$, then we have

$$\frac{\mathcal{D}_{\Lambda^-}(x)}{\widehat{Dis}(x)} = \frac{1}{\beta_2} \cdot \left(1 - \beta_1 \cdot \frac{\mathcal{D}_{\Lambda^+}(x)}{\widehat{Dis}(x)}\right) = \frac{1 - \beta_1 \delta}{\beta_2}. \tag{37}$$

Thus $h_1(\xi_\sigma(x))_{max}$ can be rewritten as:

$$\begin{aligned} h_1(\xi_\sigma(x))_{max} &= h_1(\widehat{\beta_1}) \\ &= \frac{\mathcal{D}_{\Lambda^+}(x)}{\widehat{Dis}(x)} |\Lambda^+| \cdot \log\left(\beta_1 \cdot \frac{\mathcal{D}_{\Lambda^+}(x)}{\widehat{Dis}(x)}\right) \\ &\quad + \frac{\mathcal{D}_{\Lambda^-}(x)}{\widehat{Dis}(x)} |\Lambda^-| \cdot \log\left(1 - \beta_1 \cdot \frac{\mathcal{D}_{\Lambda^+}(x)}{\widehat{Dis}(x)}\right) \\ &= \tau_n^+ \cdot \delta \cdot \log(\beta_1 \cdot \delta) + \tau_n^- \cdot \left(\frac{1 - \beta_1 \delta}{\beta_2}\right) \cdot \log(1 - \beta_1 \cdot \delta) \\ &= \tau_n^+ \cdot \delta \cdot \log(\beta_1 \cdot \delta) + (\tau_n^+ + \tau_n^- - \tau_n^+ \cdot \delta) \cdot \log(1 - \beta_1 \cdot \delta). \end{aligned} \tag{38}$$

Therefore, we obtain that

$$\begin{aligned} \inf\{\mathbb{L}(x)\} &= -\mathop{\mathbb{E}}_{\xi_\sigma \sim \widehat{Dis}} [h_1(\xi_\sigma(x))_{max}] \\ &= -\mathop{\mathbb{E}}_{\xi_\sigma \sim \widehat{Dis}} \big[\tau_n^+ \cdot \delta \cdot \log(\beta_1 \cdot \delta) \\ &\quad + (\tau_n^+ + \tau_n^- - \tau_n^+ \cdot \delta) \cdot \log(1 - \beta_1 \cdot \delta)\big], \end{aligned} \tag{39}$$

which completes the proof. □

4.3 Effect of parameters on the decision threshold of classifier

In this section, we discuss the effect of different parameters on the decision threshold (denoted as d_t) of the classifier when the TTD assumption holds and does not hold, respectively.

(1) The effect of parameters τ_n^-, τ_n^+ on function $\xi_m(x)$ when TTD assumption holds As Theorem 3 reveals, we find that τ_n^- and τ_n^+ have a direct effect on the extremes of the loss function $\mathbb{L}(x)$, which can be minimized by $\xi_m(x)$. This implies that the performance of the classifier can be improved with the aid of $\xi_m(x)$. Therefore, we explore the essential connection between the parameters τ_n^- and τ_n^+ and $\xi_m(x)$.

Note that $\xi_m(x)$ is a new classifier developed from $\xi(x)$, and $\xi_m(x)$ can be formalized as:

$$\xi_m(x) = \begin{cases} 1, & \text{if } x = \hat{x} \\ \beta_1, & \text{otherwise} \end{cases}, \tag{40}$$

where $\beta_1 = |\Lambda^+| \cdot (|\Lambda^+| + |\Lambda^-|)^{-1} = \tau_n^+ \cdot (\tau_n^+ + \tau_n^-)^{-1}$. Since τ_n^+ and τ_n^- take different values for different bags, this directly affects the range of β_1 and the determination of the decision threshold d_t for the classifier in real task. Therefore, our goal is to find a stable decision threshold d_t such that $\max(\beta_1) < d_t < 1$. Specifically, if the instances in Λ^+ and Λ^- follow TTD assumption, in terms of Theorem 3, we have

$$\beta_1 = \frac{\tau_n^+}{\tau_n^+ + \tau_n^-} = \frac{1}{1 + \frac{\tau_n^-}{\tau_n^+}}. \tag{41}$$

For the sake of descriptive brevity, we set $\frac{\tau_n^-}{\tau_n^+} = k_\tau$, then we obtain that

$$\beta_1 = \frac{1}{1 + k_\tau}, \tag{42}$$

As discussed previously, we set the number of instances in all bags to the same number, hence, according to (15), k_τ can be rewritten as:

$$k_\tau = \frac{\tau_n^-}{\tau_n^- - \tau_c^-}. \tag{43}$$

Note that in a real experiment, τ_n^- is a certain number (e.g., if the image is considered as a bag, the image can be partitioned into a certain number of patches), and the more causal factors in the bag (i.e., τ_c^+), the closer to 0 β_1 is. Therefore, in this case, the decision threshold d_t is empirically chosen as $d_t > 0.5$. Besides, we consider another special case where the bags with positive labels contain fewer causal instances (i.e., sparse positive instances), and we introduce the **w**itness **r**ate (WR [44]) to analyze the d_t selection of the classifier in the case of sparse causal instances (i.e., low witness rate).

Decision threshold selection at low witness rate. In MIL, the *witness rate* describes the ratio of the number of positive instances to the total number of instances. Therefore, in this paper, WR can be equivalently written as:

$$\text{WR} = \frac{\text{\#Causal factors}}{\text{\#Total instances}} = \frac{|\Psi^+|}{|\Psi^+| + |\Lambda^+| + |\Lambda^-|}. \tag{44}$$

When the WR is high, the classification task can be trained by a conventional supervised learning algorithm. However, the performance of the algorithm is affected at a low WR [45].

Obviously, in our task, when the total number of instances are determined, the value of WR depends only on $|\Psi^+|$. According to (15), we have that

$$|\Psi^+| = \tau_c^+ = \tau_n^- - \tau_n^+, \tag{45}$$

which is equivalent to

$$\text{WR} \propto \frac{\tau_n^-}{\tau_n^+} = k_\tau. \tag{46}$$

Thus, if $|\Psi^+|$ is small, we have that $\tau_n^- \approx \tau_n^+$, i.e., $k_\tau \approx 1$. This means that even at a low WR, β_1 can still reach 0.5. In summary, the decision threshold $d_t > 0.5$ can stably satisfy the classifier in the experiment. In other words, $\xi_m(x)$ **shows a more stable classification performance at a low witness rate.**

(2) The effect of parameters τ_n^-, τ_n^+, $\mathcal{D}_{\Lambda^-}(x)$ and $\mathcal{D}_{\Lambda^+}(x)$ on function $\xi_m(x)$ when TTD does not hold. Similar to the analysis of (40) in Theorem 3, we analyze the effect of Λ^+ and Λ^- on $\xi_m(x)$ if the instances in Λ^+ and Λ^- are drawn from the different distributions $\mathcal{D}_{\Lambda^+}(x)$ and $\mathcal{D}_{\Lambda^-}(x)$, respectively.

In this scenario, $\xi_m(x)$ is a new classifier constructed by $\xi(x)$, and $\xi_m(x)$ can be formalized in the form of

$$\xi_m(x) = \begin{cases} 1, & \text{if } x = \hat{x} \\ \widehat{\beta}_1, & \text{otherwise} \end{cases}, \tag{47}$$

where $\widehat{\beta}_1 = \frac{\tau_n^+ \cdot \mathcal{D}_{\Lambda^+}(x)}{\tau_n^+ \cdot \mathcal{D}_{\Lambda^+}(x) + \tau_n^- \cdot \mathcal{D}_{\Lambda^-}(x)}$.

Specifically, if the instances in Λ^+ and Λ^- do not follow the TTD assumption, then we have

$$\widehat{\beta}_1 = \frac{\tau_n^+ \cdot \mathcal{D}_{\Lambda^+}(x)}{\tau_n^+ \cdot \mathcal{D}_{\Lambda^+}(x) + \tau_n^- \cdot \mathcal{D}_{\Lambda^-}(x)} = \frac{1}{1 + \left(\frac{\tau_n^-}{\tau_n^+}\right) \cdot \left(\frac{\mathcal{D}_{\Lambda^-}(x)}{\mathcal{D}_{\Lambda^+}(x)}\right)}. \tag{48}$$

In contrast to (41), the value of function $\xi_m(x)$ depends on both $\frac{\tau_n^-}{\tau_n^+}$ (denoted as k_τ) and $\frac{\mathcal{D}_{\Lambda^-}(x)}{\mathcal{D}_{\Lambda^+}(x)}$ (denoted as k_Λ).

We visualize the effect of parameters $\frac{\tau_n^-}{\tau_n^+}$ and $\frac{\mathcal{D}_{\Lambda^-}(x)}{\mathcal{D}_{\Lambda^+}(x)}$ on the decision threshold of the classifier by a simple example. Suppose that $\mathcal{D}_{\Lambda^-}(x) \sim \mathcal{N}_1(\mu_1, \sigma_1)$ and $\mathcal{D}_{\Lambda^+}(x) \sim \mathcal{N}_2(\mu_2, \sigma_2)$. For the sake of computational simplicity, we assume that distributions $\mathcal{N}_1$ and $\mathcal{N}_2$ share an identical variance, i.e., $\sigma_1 = \sigma_2 = \sigma_c$. In this scenario, we have

$$\begin{aligned} k_\Lambda = \frac{\mathcal{D}_{\Lambda^-}(x)}{\mathcal{D}_{\Lambda^+}(x)} &= \frac{\frac{1}{\sigma_c\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_c}\right)^2}}{\frac{1}{\sigma_c\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_c}\right)^2}} \\ &= \exp\left(\frac{\mu_1 - \mu_2}{2\sigma_c^2}(2x - \mu_1 - \mu_2)\right). \end{aligned} \tag{49}$$

Let $\frac{\mu_2}{\mu_1} = k_\mu$, then (49) can be rewritten as:

$$\begin{aligned} k_\Lambda &= \exp\left[\frac{(1-k_\mu)\mu_1}{2\sigma_c^2} \cdot \left(2x - (1+k_\mu)\mu_1\right)\right] \\ &= \exp\left[\left(\frac{\mu_1^2}{2\sigma_c^2}\right) k_\mu^2 + \left(\frac{-x\mu_1}{\sigma_c^2}\right) k_\mu + \left(\frac{-\mu_1^2+2x\mu_1}{2\sigma_c^2}\right)\right] \\ &\triangleq \exp\left[f(k_\mu \mid \mu_1, \sigma_c, x)\right]. \end{aligned} \tag{50}$$

Therefore, for fixed instance x, mean μ_1 and standard deviation σ_c, $\frac{-x\mu_1}{\sigma_c^2}$ and $\frac{-\mu_1^2+2x\mu_1}{2\sigma_c}$ in (50) are considered constants. $f(k_\mu \mid \mu_1, \sigma_c, x)$ is a quadratic function with k_μ as the independent variable, and since $\frac{\mu_1}{2\sigma_c} > 0$, there exists a minimal value of the function $f(k_\mu \mid \mu_1, \sigma_c, x)$. After a simple calculation, we can obtain that the minimum value of the function $f(k_\mu \mid \mu_1, \sigma_c, x)$, i.e.,

$$\min k_\Lambda = \min\left(\frac{\mathcal{D}_{\Lambda^-}(x)}{\mathcal{D}_{\Lambda^+}(x)}\right) = \exp\left[-\frac{1}{2}\left(\frac{\mu_2 - \mu_1}{\sigma_c}\right)^2\right], \tag{51}$$

when $k_\mu = \frac{x}{\mu_1}$ (i.e., $x = \mu_2$).

Through the previous analysis, we confirm that both $\frac{\tau_n^-}{\tau_n^+}$ and $\frac{\mathcal{D}_{\Lambda^-}(x)}{\mathcal{D}_{\Lambda^+}(x)}$ have minimum values. Thus, we conclude that

- When $\left(\frac{\tau_n^-}{\tau_n^+}\right) \cdot \left(\frac{\mathcal{D}_{\Lambda^-}(x)}{\mathcal{D}_{\Lambda^+}(x)}\right)$ is large, $\widehat{\beta}_1$ tends to 0, and d_t can be empirically chosen as $d_t > 0.5$.
- When one of $\frac{\tau_n^-}{\tau_n^+}$ and $\frac{\mathcal{D}_{\Lambda^-}(x)}{\mathcal{D}_{\Lambda^+}(x)}$ is small, the result for $\left(\frac{\tau_n^-}{\tau_n^+}\right) \cdot \left(\frac{\mathcal{D}_{\Lambda^-}(x)}{\mathcal{D}_{\Lambda^+}(x)}\right)$ is difficult to determine, fortunately, it is then only necessary to set the d_t to a larger value (between 0.5 and 1). For example, if $\left(\frac{\tau_n^-}{\tau_n^+}\right) \cdot \left(\frac{\mathcal{D}_{\Lambda^-}(x)}{\mathcal{D}_{\Lambda^+}(x)}\right) \geq 1$, then $d_t > 0.5$ is still a valid decision threshold.
- When $\frac{\tau_n^-}{\tau_n^+}$ and $\frac{\mathcal{D}_{\Lambda^-}(x)}{\mathcal{D}_{\Lambda^+}(x)}$ are small, we consider an extreme case where $\frac{\mathcal{D}_{\Lambda^-}(x)}{\mathcal{D}_{\Lambda^+}(x)}$ reaches a minimum value and $\frac{\tau_n^-}{\tau_n^+}$ tends to 1 (i.e., low witness rate). In this case, the value of $\widehat{\beta}_1$ will be close to 1, and the d_t should be relatively large, e.g., $d_t > \frac{1}{1+\min(k_\Lambda)}$. In other words, the decision threshold of the classifier should be set as large as possible when the TTD assumption is violated.

5 Experiments

5.1 Task and remark

In this section, we perform experiments on the object classification task to validate our theoretical conclusion. Without exact coordinates for each object in the image, the object classification task can be seen as a classic standard MIL problem, the overview of the classification framework is shown in Fig. 2.

The following experiments will focus on verifying the conclusion of Theorem 3 in a real task scenario. Since Theorem 4 is a reasonable extension on Theorem 3 and the only difference is that the training data and the testing data are from different distributions. This means that experimentally verifying Theorem 4 is only a minor difference from verifying Theorem 3 at the level of experiment setup and model training, and is easy to implement. Therefore, in the rest of this paper, we do

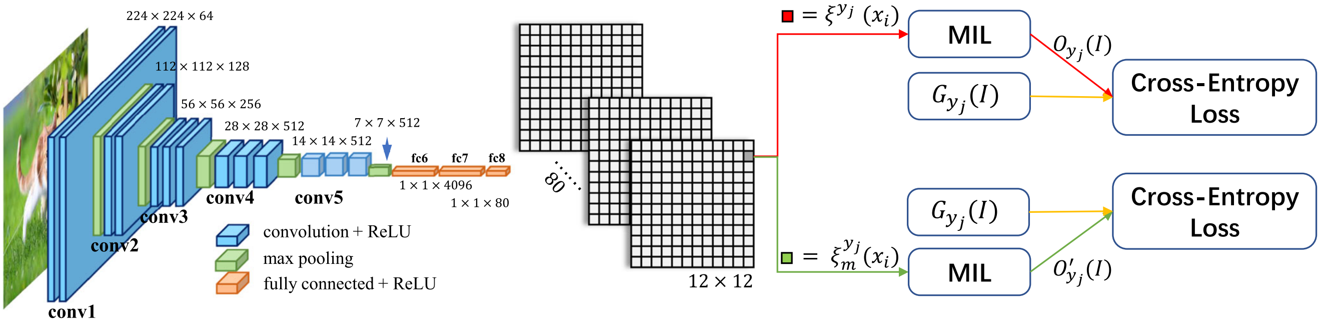

Fig. 2 An overview of the main differences between our framework (green) versus the comparison framework (red) in the object classification task. Best viewed in colors

not repeat similar experimental repetitions to verify the conclusion of Theorem 4.

5.2 Dataset

COCO [46]. The experiments to verify our conclusions are conducted on dataset COCO,[2] which is a large-scale dataset for object detection, segmentation, and captioning tasks published by Microsoft. COCO contains 1.5 million object instances. In addition, this dataset provides 5 captions per image. We use all nouns in the captions as labels of the object classification task.

5.3 Experimental setting

- **Image Processing:** we consider each image I (size of 576×576) as a bag containing a certain number of objects (12×12 patches of size 224×224 with stride 32, i.e., each image contains 144 patches).
- **Model Setup:** we adopt the VGG16 network [47] (the parameters are shown in Fig. 2) to obtain the feature maps (size of 12×12), and we use a one-dimension convolution followed by a sigmoid to generate the probability of 80 words for every image patch.
- **Labels:** captions are parsed to tokens, and 80 noun tokens with bounding boxes are selected as class labels.
- **Training and Testing dataset:** the number of training images is 82,783, the number of validation images is 20,252, and the number of testing images is 20,252.
- **Learning rate:** we set the learning rate to 0.0005. Training runs until the performance on the validation set does not improve.
- **Parameters:** an important parameter in the comparison experiment is $\frac{\tau_n^+}{\tau_n^+ + \tau_n^-}$ in Theorem 3. In the multi-label classification task of images, the number of positive and negative samples in the bag cannot be directly determined. Therefore, we have the area ratio of object region and non-object region in the COCO as the values of τ_n^+ and τ_n^-.

5.4 Experimental method

Let $\{y_i\}_l$ be the label set, according to the experimental setup, we have $l = 80$. For a fixed label y_j, the MIL objective can be obtained by

$$O_{y_j}(I) = 1 - \prod_{i=1}^{144} \left(1 - \xi^{y_j}(x_i)\right), \tag{52}$$

where $\xi^{y_j}(x_i)$ is the output of the classifier $\xi(\cdot)$ (in Theorem 3) with x_i as input, under the label y_j, x_i is a patch (144 in total) of an image I. Hence, we can obtain the total cost by summing all $O_{y_j}(I)$, i.e.,

$$\mathrm{MIL}(I) = \sum_{y_j, j=1}^{80} \mathrm{CE}(O_{y_j(I)}, G_{y_j}(I)), \tag{53}$$

where $G_{y_j}(I)$ is ground truth corresponding to label y_j, CE stands for cross-entropy.

Similarly, for a specific label y_j, the SIL objective can be obtained by

$$\mathrm{SIL}(I) = \sum_{y_j, j=1}^{80} \sum_{x_i, i=1}^{144} \mathrm{CE}(\xi_m^{y_j}(x_i), G_{y_j}(I)), \tag{54}$$

where $\xi_m^{y_j}(x_i)$ is the output of classifier $\xi_m(\cdot)$ (i.e, (17) in Theorem 3) corresponding to the label y_j. Then, we compare the performance of the model at the bag level. Specifically, for an image I and a specific label y_j, we use

$$S_{y_j}(I) = \max_i \xi_m^{y_j}(x_i) \tag{55}$$

as the bag level score. Finally, we compare $S_{y_j}(I)$ with $G_{y_j}(I)$ using four popular metrics, which will be given in the next section.

[2] https://cocodataset.org/#home

5.5 Metrics

We compare the performance of the model with $\xi(\cdot)$ and the optimized model with $\xi_m(\cdot)$ by four metrics [48], **m**ean **a**verage **p**recision (mAP), **h**amming **l**oss (HL), **cov**erage (COV), and **r**anking **l**oss (RL).

- **Mean average precision** evaluates the mean of average precision for all classes, where **a**verage **p**recision (AP) can be formalized as:

$$\mathrm{AP} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|Y_i|} \times \sum_{y\in Y_i}\frac{\left|\left\{\hat{y} \mid \mathrm{rank}\left(x_i,\hat{y}\right) \leqslant \mathrm{rank}\left(x_i,y\right), \hat{y}\in Y_i\right\}\right|}{\mathrm{rank}\left(x_i,y\right)}, \tag{56}$$

  where $\hat{y}$ is the ground truth and $\mathrm{rank}\left(x_i,y\right)$ refers to the ranking of the object x_i predicted to be y. n is the total number of samples in the testing data. The value of mAP ranges from 0 to 1. The closer to 1 the value is, the better the performance of the model is.
- **Hamming loss** measures the number of times a label is misclassified, e.g., a label belonging to a sample is not correctly predicted, or a label that does not belong to a sample is incorrectly predicted as belonging to that sample. It can be formalized as:

$$\mathrm{HL} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\left|f\left(x_i\right)\mathrm{SD}(Y_i)\right|, \tag{57}$$

  where $f(\cdot)$ is a multi-label classifier and $\mathrm{SD}(\cdot)$ evaluates the symmetric difference between two sets. T is the total number of labels. The value of HL ranges from 0 to 1. The closer to 0 the value is, the better the performance of the model is.
- **Coverage** denotes the mean of the least ranked true label among all samples. It can be formalized as:

$$\mathrm{COV} = \frac{1}{n}\sum_{i=1}^{n}\max_{y\in Y_i}\mathrm{rank}\left(x_i,y\right) - 1. \tag{58}$$

  The value of CE ranges from 0 to 1. The closer to 0 the value is, the better the performance of the model is.
- **Ranking loss** measures the number of times that the predicted probability value of a relevant label is smaller than the predicted probability value of an irrelevant label. **R**anking **l**oss (RL) can be formalized as:

$$\mathrm{RL} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\left\{(y_1,y_2) \mid f\left(x_i,y_1\right) \leqslant f\left(x_i,y_2\right)\right|}{|Y_i||Y_i^c|}, \quad (y_1,y_2)\in Y_i\times Y_i^c. \tag{59}$$

  Where Y_i^c is the complementary set of Y_i. The value of RL ranges from 0 to 1. The closer to 0 the value is, the better the performance of the model is.

5.6 Experimental results and analysis

The core of the experiment is to compare $\xi(\cdot)$ in the MIL approach (i.e., Fig. 2) with $\xi_m(\cdot)$ in the SIL approach (i.e., Theorem 3). The results are given in Table 3.

As shown in Table 3, we find that the model $\xi_m(\cdot)$ outperforms the original multi-instance classification model in all four metrics. Overall, the results in Table 3 indicate that:

- The model with $\xi_m(\cdot)$ improves mAP by 2.92%. This is a good indication of the validity of the model with $\xi_m(\cdot)$.
- Model with $\xi_m(\cdot)$ also achieves 0.035, 0.353, and 0.003 improvements in metrics HL, COV, and RL. The results show that the improvement of mAP does not come at the expense of other metrics, which sufficiently proves the conclusion of our theoretical analysis.
- *Causal factors have two effects on multi-instance learning, one is that the causal factors in a bag determine the bag label, and the other is that their number affects the performance of the model.* Specifically, we know that the main difference between the model with $\xi(\cdot)$ and the model with $\xi_m(\cdot)$ is the coefficient $\beta = \tau_n^+ \cdot (\tau_n^+ + \tau_n^-)^{-1}$. When the number of instances in the bag is determined (e.g., each image is divided into 144 patches in our experiments), then the coefficient β will strictly depend on the number of causal factors in the bag (because $\tau_n^+ = \tau_n^- - \tau_c^+$). To our surprise, the coefficient τ_n^+ plays a positive role in all four of these main evaluation metrics.
- Together, these results provide important insights into capturing the inner relationship between multi-instance learning and causal factors.

Table 3 Comparison of model performance between $\xi_m(\cdot)$ and $\xi(\cdot)$

Model	Method	mAP(%)↑	HL↓	COV↓	RL↓
The model with $\xi(\cdot)$	MIL (i.e., (52))	57.20	0.172	9.583	0.045
The model with $\xi_m(\cdot)$	SIL (i.e., (55))	**60.12**	**0.137**	**9.230**	**0.042**

Larger mAP indicates better model performance, while smaller HL (Hamming loss), COV (Coverage), and RL (Ranking loss) indicate better model performance

6 Conclusion

In this paper, we conduct a theoretical analysis of MIL problems by using causal inference theory and the SIL method. We first analyze the role of positive factors in the bag using causal inference theory. In Theorem 2, we prove a lower bounder of the number of sample instances needed to effectively determine the causal factors within a certain confidence interval. We then analyze the impact of data distribution on the multi-instance learning problem. Most of the previous MIL tasks are based on a strong constraint (i.e., the TTD assumption) on data distribution. However, in many real applications, the TTD assumption is not followed. We capture the relationship between the number of instances in the bag and the loss function when the TTD assumption holds and does not hold, respectively i.e., Theorem 3 and Theorem 4.

Specifically, we exhaustively analyze the number of instances in the bag and the advantages of using a single-instance approach to solve multi-instance tasks. Although some previous studies have illustrated the drawbacks of the single-instance approach in multi-instance learning tasks, our conclusions show that the single-instance approach demonstrates good robustness in solving multi-instance learning problems with acceptable time costs. In addition, an important detail is that we tend to ignore the effect of the number of positive and negative instances in the bag on the model when constructing the bag. Our conclusions show that although the number of positive instances in the bag does not affect the bag label (based on the definition of bag label in multi-instance learning), it has a direct impact on the performance of the model.

In addition, we analyze the effect of some parameters on the decision threshold of the classifier. The theoretical analysis shows that when the training and testing samples follow independent identical distributions, the decision threshold can be empirically chosen to be 0.5, while when the training and testing samples do not follow independent identical distributions, the decision threshold should be appropriately increased to ensure the performance and effectiveness of the classifier.

Finally, we verify our above theoretical analysis by a classical MIL task, the experimental results and the theoretical analysis provide important insights for researchers using causal inference theory and the SIL method to study multi-instance learning tasks.

Acknowledgements This work was supported by the National Key R&D Program of China under Grant 2018YFB1403200.

References

1. Keeler JD, Rumelhart DE, Leow WK (1990) Integrated segmentation and recognition of hand-printed numerals. In: conference on advances in neural information processing systems
2. Dietterich TG, Lathrop RH, Lozano-Perez T (1997) Solving the multiple instance problem with axis-parallel rectangles. Artif Intell 89(1-2):31–71
3. Zhang M-L, Zhou Z-H (2009) Multi-instance clustering with applications to multi-instance prediction. Appl Intell 31(1):47–68
4. Hebbar R, Papadopoulos P, Reyes R, Danvers AF, Polsinelli AJ, Moseley SA, Sbarra DA, Mehl MR, Narayanan S (2021) Deep multiple instance learning for foreground speech localization in ambient audio from wearable devices. EURASIP J Audio Speech Music Process 2021(1):1–8
5. Yaghoobzadeh Y, Adel H, Schütze H (2018) Corpus-level fine-grained entity typing. J Artif Intell Res 61:835–862
6. Alam FF, Shehu A (2021) Unsupervised multi-instance learning for protein structure determination. J Bioinforma Comput Biol 19(01):2140002
7. Morampudi MK, Veldandi S, Prasad MVNK, Raju USN (2020) Multi-instance iris remote authentication using private multi-class perceptron on malicious cloud server. Appl Intell 50(9):2848–2866
8. Morampudi MK, Prasad MVNK, Raju USN (2021) Privacy-preserving and verifiable multi-instance iris remote authentication using public auditor. Appl Intell:1–14
9. Fei L, Zhang B, Tian C, Teng S, Wen J (2021) Jointly learning multi-instance hand-based biometric descriptor. Inf Sci 562:1–12
10. Tarek M, Hamouda E, Abohamama AS (2021) Multi-instance cancellable biometrics schemes based on generative adversarial network. Appl Intell:1–13
11. Shamsolmoali P, Zareapoor M, Zhou H, Yang J (2020) Amil: Adversarial multi-instance learning for human pose estimation. ACM Trans Multimed Comput Commun Appl 16(1s):23
12. Schwab E, Gooßen A, Deshpande H, Saalbach A (2020) Localization of critical findings in chest x-ray without local annotations using multi-instance learning. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, pp 1879–1882
13. He K, Zhao W, Xie X, Ji W, Liu M, Tang Z, Shi Y, Shi F, Gao Y, Liu J et al (2021) Synergistic learning of lung lobe segmentation and hierarchical multi-instance classification for automated severity assessment of covid-19 in ct images. Pattern Recogn 113:107828
14. Brand L, Baker LZ, Wang H (2021) A multi-instance support vector machine with incomplete data for clinical outcome prediction of covid-19. In: proceedings of the 12th ACM conference on bioinformatics, computational biology, and health informatics, pp 1–6
15. Zhang YL, Zhou ZH (2017) Multi-instance learning with key instance shift. In: twenty-sixth international joint conference on artificial intelligence
16. Kozdoba M, Moroshko E, Shani L, Takagi T, Katoh T, Mannor S, Crammer K (2018) Multi instance learning for unbalanced data. arXiv:1812.07010
17. Sugiyama M, Suzuki T, Kanamori T (2012) Density ratio estimation in machine learning. Cambridge University Press
18. Liu A, Ziebart BD (2014) Robust classification under sample selection bias. Adv Neural Inf Process Syst 1:37–45

19. Zhang W-J, Zhou Z-H (2014) Multi-instance learning with distribution change. In: twenty-eighth AAAI conference on artificial intelligence
20. Pearl J, Mackenzie D (2018) The book of why: the new science of cause and effect. Basic Books
21. Kuang K, Cui P, Athey S, Xiong R, Li B (2018) Stable prediction across unknown environments. In: proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp 1617–1626
22. Shen Z, Cui P, Kuang K, Li B, Chen P (2018) Causally regularized learning with agnostic data selection bias. In: proceedings of the 26th ACM international conference on multimedia, pp 411–419
23. Zhang W, Liu L, Li J (2020) Robust multi-instance learning with stable instances. In: ECAI 2020, 24th european conference on artificial intelligence: 29 August–8 September 2020, Santiago de Compostela, Spain: including 10th conference on prestigious applications of artificial intelligence (PAIS 2020): proceedings. Ios Press, pp 1682–1689
24. Feng L, Shu S, Cao Y, Tao L, Wei H, Xiang T, An B, Niu G (2021) Multiple-instance learning from similar and dissimilar bags. In: proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pp 374–382
25. Blumberg CJ (2016) Causal inference for statistics, social, and biomedical sciences: An introduction. Int Stat Rev 84(1):159–159
26. Foulds JR, Frank E (2010) A review of multi-instance learning assumptions. Knowl Eng Rev 25(1):1–25
27. Fernández A, Garcia S, Herrera F, Chawla NV (2018) Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res 61:863–905
28. Park S, Bastani O, Weimer J, Lee I (2020) Calibrated prediction with covariate shift via unsupervised domain adaptation. arXiv:2003.00343
29. Carbonneau M-A, Cheplygina V, Granger E, Gagnon G (2018) Multiple instance learning: A survey of problem characteristics and applications. Pattern Recogn 77:329–353
30. Zhou Z-H, Sun Y-Y, Li Y-F (2009) Multi-instance learning by treating instances as non-iid samples. In: Proceedings of the 26th annual international conference on machine learning. ACM, pp 1249–1256
31. Wang Z, Poon J, Poon SK (2019) Ami-net+: A novel multi-instance neural network for medical diagnosis from incomplete and imbalanced data. Aust J Intell Inf Process Syst 15(3):8–15
32. Hu W, Niu G, Sato I, Sugiyama M (2018) Does distributionally robust supervised learning give robust classifiers? In: International Conference on Machine Learning. PMLR, pp 2029–2037
33. Tan Y, Sun D, Shi Y, Gao L, Gao Q, Lu Y (2021) Bi-directional mapping for multi-label learning of label-specific features. Appl Intell:1–20
34. Alpaydın E, Cheplygina V, Loog M, Tax DMJ (2015) Single-vs. multiple-instance classification. Pattern Recogn 48(9):2831–2838
35. Chen Y, Bi J, Wang JZ (2006) Miles: Multiple-instance learning via embedded instance selection. IEEE Trans Pattern Anal Mach Intell 28(12):1931–1947
36. Bunescu RC, Mooney RJ (2007) Multiple instance learning for sparse positive bags. In: Proceedings of the 24th international conference on Machine learning, pp 105–112
37. Yuan M, Xu Y, Feng R, Liu Z (2021) Instance elimination strategy for non-convex multiple-instance learning using sparse positive bags. Neural Netw 142:509–521
38. Wei X-S, Wu J, Zhou Z-H (2016) Scalable algorithms for multi-instance learning. IEEE Trans Neural Netw Learn Syst 28(4):975–987
39. Küçükaşci EŞ, Baydoğan MG, Taşkin ZC (2021) A linear programming approach to multiple instance learning. Turkish J Electr Eng Comput Sci 29(4):2186–2201
40. Shimada T, Bao H, Sato I, Sugiyama M (2021) Classification from pairwise similarities/dissimilarities and unlabeled data via empirical risk minimization. Neural Comput 33(5):1234–1268
41. Holland PW (1986) Statistics and causal inference. J Amer Stat Assoc 81(396):945–960
42. Fisher NI, Sen PK (1963) Probability inequalities for sums of bounded random variables. Publ Am Stat Assoc 58(301):13–30
43. Serfling RJ (1974) Probability inequalities for the sum in sampling without replacement. Ann Stat 2(1):39–48
44. Li F, Sminchisescu C (2010) Convex multiple-instance learning by estimating likelihood ratio. In: NIPS, vol 10. Citeseer, pp 1360–1368
45. Li Y, Tax DMJ, Duin RPW, Loog M (2013) Multiple-instance learning as a classifier combining problem. Pattern Recogn 46(3):865–874
46. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision. Springer, pp 740–755
47. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. Computer Science
48. Zhang M-L, Zhou Z-H (2007) Ml-knn: A lazy learning approach to multi-label learning. Pattern Recogn 40(7):2038–2048