Towards Identifying Causal Relation Between Instances and Labels

Tian-Zuo Wang * Sheng-Jun Huang † Zhi-Hua Zhou *

Abstract

Multi-Instance Multi-Label (MIML) learning is a popular framework in machine learning, where each object is represented by a bag of instances, and associated with multiple labels. While MIML learning has achieved success in many applications, it is less clear how the labels are related to the instances. In this paper, we propose to study the causal relation between instances and labels, which on one hand can improve the interpretability of complicated MIML models, and on the other hand may further improve the prediction performance at both instance and bag levels. We exploit prototypes in the instance space as a bridge to represent the examples, and then propose an efficient algorithm to identify the causal relations from prototypes to class labels, which are further utilized for model training and key instance detection. Experiments on various datasets show that in addition to superior classification performance, our approach can identify reasonable causal relations between instances and labels.

1 Introduction

Multi-Instance Multi-Label learning is a popular framework for learning from complicated objects. In MIML, each example is represented by a bag of instances, and at the same time is annotated with multiple class labels. For example, in the task of text categorization, a document consists of multiple paragraphs, each one is represented by an instance; and may be simultaneously related to *finance* and *politic*. Figure 1 is an illustration of the MIML framework [32].

During the past years, MIML has achieved success in various applications, such as image classification [4, 31], text categorization [29], music analysis [6], etc. However, given the high complexity of the learning objects, it becomes rather challenging to learn the many-to-many mapping from instance space to the label space. Many methods try to degenerate the MIML problem into traditional supervised learning problems and then solve them via existing approaches [31]. Such methods can reduce the complexity of MIML problems and scale well with the data size. However, they may loss important information by treating the labels or instances independently, and subsequently obtain sub-optimal performance. It thus attracts many research interests to exploit the relationship among instances or labels to help the learning task.

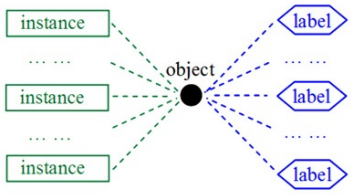


Figure 1: Illustration of MIML framework [32]. An object is represented by a bag of multiple instances, and associated with multiple class labels.

Some methods try to exploit the label relationship, mainly by constraints on label co-occurrence [7, 23] or model reusing among labels [11]. While exploiting high-order label correlation is difficult or even impossible due to the high computational cost, second-order correlation is commonly considered between label pairs, which however, can only partially capture relationship in the output space [7, 23]. Some other methods try to exploit the instance relationship in the input space. For example, the spatial relationships in a bag are exploited under the assumption that instances are non i.i.d. [30].

While exploiting the relationships either in input space or output space can somehow improve the performance of MIML, it is more essential to disclose the causal relation between instances and labels. Firstly, exploiting the causal relation can improve the interpretability of the learning models, which is especially important for the complicated MIML framework. For example, in medical image analysis, in addition to an accurate model that can predict the diseases based on the image, it is also crucial to let people know what causes the disease. Secondly, exploiting the causal relation may help to improve the learning performance. The causal relation describes the direct effect of an instance to a label, and may provide useful guidance to

*National Key Laboratory for Novel Software Technology, Nanjing University, China, {wangtz,zhouzh}@lamda.nju.edu.cn

†College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, {huangsj}@nuaa.edu.cn

the model training in addition to the empirical loss minimization. Actually, with causal relations considered, it would be naturally disclosed that the labels with shared cause will be correlated with each other. This provides us an efficient and effective approach to multi-instance multi-label learning without the time-consuming process of utilizing the label correlations.

In this paper, we propose to identify the causal relation between instances and labels with prototype based representation as a bridge. Specifically, prototypes are collected by performing clustering the original instance space, and then each instance can be represented by the similarities between itself and the prototypes. After that, based on a DAG structure of the prototypes, we propose an efficient algorithm to identify the causal relations from prototypes to class labels by examining the conditional independence. With the identified causations, on one hand, we select the feature dimensions corresponding to the cause prototypes to train a MIML model for better performance; and on the other hand, the key instances for each label can be detected from the bag based on the cause prototypes. Experiments are performed on different tasks; and the comparison results with some state-of-the-art methods show that by exploiting the causal relation, the proposed approach can achieve significant better performance with high interpretability.

The main contributions of this study are summarized as follows:

- For the first time we propose to study the causal relations between instances and labels under the MIML framework. We disclose that causal discovery is a more essential problem in MIML, and can help to improve both the prediction performance and interpretability.

- We propose an effective approach by firstly transforming the instances into prototype based representations, and then efficiently identify the causations from DAG structured prototype variables.

- Extensive experiments are performed to validate the superiority of the proposed method on the label prediction, the interpretability and the key instance detection.

The rest of this paper is organized as follows. Section 2 reviews some related studies on MIML and causality. Section 3 presents our approach in detail, followed by the experiments in Section 4. Finally, Section 5 concludes with future work discussed.

2 Related Work

Multi-Instance Multi-Label learning has attracted many research interests due to its advantages on learning with complicated objects [12,13,15,20,26,31,32]. A lot of efforts are made on designing effective algorithms for this high-complexity framework, mainly by exploiting the relationship among instances or labels. In [31], under the assumption of independence among the labels or the instances, authors proposed MIMLSVM by degenerating MIML to single-instance multi-label problems, and MIMLBooost by degenerating MIML to multi-instance single-label problems. After the degeneration, the correlations may be further exploited to improve the effectiveness. For example, instances are treated as non i.i.d. in [30], and different orders of label correlations may be utilized [7, 23]. There are also some methods try to exploit the correlations without degeneration [2, 19, 32]. For example, the cluster structure inside each class is considered in [19] . These methods can somehow benefit from the correlations considered to get better performance. However, they focus on the relationship within either the instance space or label space, but fail to study the causal relations from instances to labels, which is more essential. There are a few studies try to find what instances trigger what labels under the MIML framework [13, 20]. However, these methods focus on identifying the key instances based on the prediction model, instead of exploiting the real causal relations between instances and labels.

In the literature of causality, causal discovery is one of the most important problems. There are mainly two kinds of methods. One is constraint-based approaches [10, 14, 16, 18, 22], which discover the causal relation based on the conditional independence. The other is score-based approach [5], in which every possible causal structure is given a score and that with the highest score will be output.

Traditional constraint-based approach [22] can not identify the causal structure in the Markov equivalence class. In order to overcome this shortcoming, the additive noise model is considered in [10]. Based on the similar idea, it is extended to a more general case with post-nonlinear causal model in [27]. Then in [14] and [18], a novel causal discovery approach named Regression with Subsequent Independence Test (RESIT) is proposed, which is also based on noise additive model and employs a regression method to minimize the statistical dependence between the regressors and residuals. Also, causal discovery between discrete variables is studied in [17].

3 MIML with Causal Discovery

We denote by $\mathcal{D} = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$ a MIML dataset with n bags, where $X_i =$

$\{\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \cdots, \mathbf{x}_{i,n_i}\}$ is the i-th bag consisting of n_i instances, and $Y_i \subset \{y_1, y_2, \cdots, y_L\}$ is the relevant subset of all L possible labels. Let $\mathcal{X}$ denotes the space of instances and $\mathcal{Y}$ denotes the space of labels, the task of MIML is to learn the mapping function $f : 2^{\mathcal{X}} \to 2^{\mathcal{Y}}$ based on $\mathcal{D}$. In this paper, in addition to the function f, we also want to identify the causal relations between instances and labels. Formally, given a relevant label y_l of the bag X_i, we want to identify the instance $\mathbf{x}^*_{i,l} \in X_i$ that causes the label y_l.

Obviously, it is infeasible to directly apply causal discovery techniques to the instance level because they are designed to identify the causal relations between single variables. One alternative solution is to examine the causal relation between each feature dimension and a specific label. However, in many applications, one dimension of the feature space does not necessarily have a semantic meaning, which prevents us from understanding the model even causations. To overcome this challenge, we introduce prototype representations as a bridge, based on which the instances will be transformed into a new space with explicit semantic meaning in each dimension. After discovering the causal relations between prototypes and labels, the cause instances can be further identified easily with a simple classification model.

To transform instances into a new representation suitable for causal discovery, we firstly collect all the instances from the whole training set, and then perform clustering to obtain the cluster centers as prototypes. Each prototype represents some common properties shared by a group of similar instances and can be expected to has some semantic meanings [13]. Denoting by $\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_K$ the K prototypes, then an instance $\mathbf{x}$ can be transformed into the new representation $\phi(\mathbf{x})$ with K dimensions:

$$\phi(\mathbf{x}) = [sim(\mathbf{x}, \mathbf{c}_1), sim(\mathbf{x}, \mathbf{c}_2), \cdots, sim(\mathbf{x}, \mathbf{c}_K)], \tag{3.1}$$

where each dimension is the similarity between the instance and the corresponding prototype, measured by the function $sim(\mathbf{x}, \mathbf{c}_k) = \exp(-\|\mathbf{x} - \mathbf{c}_k\|_2^2/\delta_k^2)$. Here δ_k denotes the average distance between the prototype $\mathbf{c}_k$ and all instances within this cluster. Further, we can also have a K-dimensional feature vector to represent each bag X as follows:

$$\phi(X) = [sim(X, \mathbf{c}_1), sim(X, \mathbf{c}_2), \cdots, sim(X, \mathbf{c}_K)],$$

where $sim(X, \mathbf{c}_k) = \max_{\mathbf{x} \in X}\{sim(\mathbf{x}, \mathbf{c}_k)\}$.

We then try to discover the causal relations between feature variables and labels in the new representation space, where each feature variable corresponds to a prototype. For simplicity of presentation, we will refer to the variable corresponding to a dimension in the prototype based representation space as a prototype variable.

In traditional causal discovery studies, the task is to identify all the causal relations from a given set of variables, where all variables are treated equally. However, in our task, we have two different groups of variables: the continuous variables of prototype features and the discrete variables of class labels. Also, we have a strong prior knowledge that some features cause some labels, but not vice versa. This difference, on one hand prevents us from directly applying existing causal discovery methods to our task, but on the other hand brings a good news that we can focus on identifying causal relations only from prototype variables to label variables, instead of among all variables. Besides, we notice there are no causations between labels. Because the label we give only depends on the object, but not other labels. With these prior knowledge, we propose to efficiently discover the causal relations with two steps.

In the first step, we employ an existing approach RESIT [18] to determine the causal orders of prototype variables. RESIT tries to find a directed acyclic graph (DAG) to represent the causal relations, where each variable corresponds to each node, and the parent nodes have a direct cause effect to the children nodes. Given a candidate DAG, RESIT models a node z_i as a function of its parents as follows,

$$z_i = f_i(pa(z_i)) + \mu_i, i = 1, \ldots, n,$$

where n is the number of nodes, μ_i stands for the noise about the node z_i. Then the algorithm tries to approximate the unknown function f_i by minimizing the dependence between the residuals and the parents nodes. The dependence is measured by the empirical HSIC estimator [9]. After that, the residuals could be estimated. For every node, if its residuals are independent with its parents nodes, then we consider there is no causal relations between them.

We employ the RESIT algorithm to obtain the causal relations between the prototype variables. For clarity of presentation, we further record the causal relations in the form of a causal matrix, denoted by $\mathcal{C}$. Specifically, if the variable corresponding to the i-th prototype has a direct causal effect to the j-th prototype variable, then the element $\mathcal{C}_{i,j}$ at the i-th row and j-th column equals to 1; otherwise $\mathcal{C}_{i,j} = 0$.

In the second step, based on the causal relations between prototype variables obtained in the first step, we try to discover the causal relations from prototype variables to class labels. One straightforward way is to exhaustively examine all possible relations from prototype variables to labels. However, such a simple method

could be time consuming, what's worse, it could suffer from serious noises due to the disturbances from other prototype variables. Instead, we propose an efficient approach to consider the variables in an effective order by utilizing the previously obtained causal matrix $\mathcal{C}$.

First of all, we define the causal order for a group of variables as follows.

DEFINITION 3.1. (CAUSAL ORDER) *Given a set of variables $\{z_1, z_2, \cdots, z_K\}$, and a function $\mathcal{G}(z_k)$ returns the rank of the variable z_k among the K variables. $\mathcal{G}$ defines a causal order if $\mathcal{G}(z_i) > \mathcal{G}(z_j)$ for any z_i and z_j that there is no direct causal effect from z_i to z_j.*

Obviously, the causal order of prototype variables can be easily obtained from the causal matrix $\mathcal{C}$. Then we show in Theorem 3.1 that based on the faithfulness and causal sufficiency assumptions, causal relations can be effectively discovered without considering all variables. The faithfulness assumption states that the conditional independence between two variables holds if and only if there is no direct causal relation between them. The causal sufficiency assumption excludes the latent variable that has direct causal effect to more than one target variables. Both of them are basic assumptions commonly used in causal discovery literature.

THEOREM 3.1. *Given a set of prototype variables $\{z_1, z_2, \cdots, z_K\}$ with known DAG causal structure and causal order $\mathcal{G}$. For a prototype variable z_i and a label y_l, let $\mathcal{S}$ denotes the set consisting of the parents of z_i and variables ranked after z_i but have direct causal effect to y_l, i.e., $S = pa(z_i) \cup \{z | z \to y_l, \mathcal{G}(z) > \mathcal{G}(z_i)\}$. Under the faithfulness and causal sufficiency assumption, z_i has no direct causal effect to the label y_l if and only if z_i is conditional independent of y_l given the set S, i.e., $p(z_i|S) \times P(y_l|S) = P(z_i, y_l|S)$.*

Proof. $\Leftarrow$ If there is a directed edge from z_i to the label y_l, no matter what set $\mathcal{S}$ is given, the z_i will be not conditionally independent of the label (this is because of the faithfulness assumption in causality literature). It contradicts the condition. So there is no causal relation from z_i to the label y_l.
$\Rightarrow$ Assume that the variables except for $\mathcal{S}$, z_i and y_l constitute the set T, we will give a proof to $p(z_i|S) \cdot P(y_l|S) = P(z_i, y_l|S)$. At first, when z_i has no causal effect to the label y_l,

$$\begin{aligned} P(z_i, y_l|S) &= \sum_T P(z_i, y_l|S, T) P(T|S) \\ &= \sum_T P(z_i|S, T) \cdot P(y_l|S, T) P(T|S) \\ &= \sum_T P(z_i, T|S) \cdot P(y_l|S, T) \end{aligned}$$

The second line holds because there is no causal effect from y_l to any others so that z_i is conditional independent of y_l given all other variables. We divide T into two parts $\{T_1, T_2\}$, T_1 represents the variables that is before z_i in the causal order $\mathcal{G}$, T_2 represents the others. With the Markov condition, we have

$$\begin{aligned} P(z_i, T_1|S) &= P(z_i|S) P(T_1|S) \\ P(y_l|S, T) &= P(y_l|S, T_1) \end{aligned}$$

then we have

$$\begin{aligned} P(z_i, y_l|S) &= \sum_{T_1, T_2} P(z_i, T_1, T_2|S) \cdot P(y_l|S, T_1, T_2) \\ &= \sum_{T_1, T_2} P(z_i, T_1, T_2|S) \cdot P(y_l|S, T_1) \\ &= \sum_{T_1} P(z_i, T_1|S) \cdot P(y_l|S, T_1) \\ &= \sum_{T_1} P(z_i|S) \cdot P(T_1|S) \cdot P(y_l|S, T_1) \\ &= P(z_i|S) \sum_{T_1} P(T_1|S) \cdot P(y_l|S, T_1) \\ &= P(z_i|S) \sum_{T_1} P(y_l, T_1|S) \\ &= P(z_i|S) P(y_l|S). \end{aligned}$$

With Theorem 3.1, we can effectively detect the causal relations from prototype variables to labels as shown in Algorithm 1. For each label y_l, we maintain a set Q_l to collect the prototype variables that have a direct causal effect to y_l. The prototype variables are iteratively examined one by one in the reverse order of $\mathcal{G}$. At the k-th iteration, we first find the prototype variable ranked at the k-th position according to the causal order. Without loss of generality, we assume this prototype variable is z_i. Then the parents of z_i are collected as $pa(z_i)$ according to the causal matrix $\mathcal{C}$. We again define a function f to predict the value of z_i with $pa(z_i)$ and Q_l as inputs:

$$z_i = f(pa(z_i), Q_l).$$

After that, we train the function f in the rule of empirical risk minimization and get the prediction value. And with the actual value of z_i, we can obtain the residual, denoted by $\hat{\epsilon}_1$. Similarly, we have $\hat{\epsilon}_2$ as the residual between label y_l and the prediction value with the same input variables. At last, a statistical test is performed to examine the independence of $\hat{\epsilon}_1$ and $\hat{\epsilon}_2$. If the two groups of residuals are not dependent, then z_i is conditional independent of the label y_l given all the variables in Q_l and $pa(z_i)$. Then there is no causal relation between z_i and y_l. Otherwise, z_i has a direct causal effect to y_l, and will be added into Q_l.

Algorithm 1 Causal discovery from prototype variables to labels

input: A data set of n bags with prototype based representations; the threshold value α;
The causal matrix $\mathcal{C}$ and causal order $\mathcal{G}$ of the K prototype variables.

output: The set of prototype variables Q_l that have cause effect to each label y_l.

```
for l = 1 to L do
    Q_l ← ∅
    for k = K, K−1, ··· , 1 do
        locate the variable z_i with G(z_i) = k;
        pa(z_i) = {z_t | ∀t ∈ {1 : K} and C_{t,i} = 1}
        ε̂_1 ← Residuals(z_i, pa(z_i) ∪ Q_l)
        ε̂_2 ← Residuals(y_l, pa(z_i) ∪ Q_l)
        p_i ← TestIndependency(ε̂_1, ε̂_2)
        if p_i > α then
            continue
        else
            Q_l = Q_l ∪ z_i
        end if
    end for
end for
```

So far we have obtained the causal relations from the prototype variables to class labels. This naturally inspires us to train the MIML classifier by the causal relations exploited. As the bags have been transformed into vector representations based on the prototypes, the original MIML problem becomes a single-instance multi-label learning problem. As discussed before, label correlations are only derived from prototype variable's causal relations. Some prototype variables cause many labels at the same time, which take correlations to them. Hence, the label correlations could be ignored if the causal relations between the prototype variables and labels have been exploited. We thus try to train a classifier independently for each label. For each label y_l, we select the set of features corresponding to the causal variables in Q_l to train the classifier. Although we convert the original MIML problem into much simpler problems, it can be expected to not degenerate the performance by exploiting the causal relations. This is validated in our experiments.

At last, we discuss how to identify the key instance from a bag that causes a specific label. Assume that we already trained a classifier g_l for label y_l with the above method, then we can also get a prediction for any instance with the classifier, given that the instances and bags sharing the same representation space based on the prototypes. So the key instance can be easily identified by selecting the one with maximum prediction value.

Table 1: Discovered causal relations between the features and the labels from the synthetic dataset.

label	z_1	z_2	z_3	z_4	z_5	z_6	z_7
y_1	1	0	1	0	0	0	0
y_2	0	0	0	1	0	0	1
y_3	0	1	0	0	0	1	1

Table 2: The datasets used in the experiments.

Data	# instance	# bags	# label	# label per bag
Bird Song	10232	548	13	2.1
MSRC v2	1758	591	23	2.5
Letter Frost	565	144	26	3.6
Letter Carroll	717	166	26	3.9
Scene	18000	2000	5	1.2
News	24406	612	10	1.5

4 Experiment

We evaluate the effectiveness of our method from different aspects. Firstly, we test on a synthetic data to examine the causal relations identified by the proposed method. Then, 6 MIML datasets along with 20 multi-instance datasets are used to compare the proposed method with state-of-the-art methods on multiple performance measures. After that, comparison results on key instance detection and causal effect are reported.

4.1 Causal Discovery on Synthetic Data To validate the effectiveness of causal discovery for the proposed method, we generate a synthetic dataset with 7 feature variables $\{z_1, z_2, \cdots, z_7\}$ and 3 label variables $\{y_1, y_2, y_3\}$. The causal function relationships among the variables are described as follows:

$$
\begin{aligned}
z_1 &= \sin(z_4^2) + z_2^2 + \cos(z_7) + E_1, && E_1 \sim U(-0.1, 0.1)\\
z_2 &= z_3^2 + E_2, && E_2 \sim U(-0.5, 0.5)\\
z_3 &= E_3, && E_3 \sim U(-1.0, 1.0)\\
z_4 &= \sin(z_2) + \sin(2z_3) + E_4, && E_4 \sim U(-0.5, 0.5)\\
z_5 &= \tanh(z_6 + z_7 + z_2) + E_5, && E_5 \sim U(-0.3, 0.2)\\
z_6 &= \sin(z_2) + \cos(2z_4) + E_6, && E_6 \sim U(-0.5, 0.5)\\
z_7 &= \cos(z_6 + z_3) + E_7, && E_7 \sim U(-0.3, 0.3)\\
y_1 &= \mathbb{I}(z_3 + z_1^2 - 1.5 + \epsilon_1 > 0), && \epsilon_1 \sim N(0, 1)\\
y_2 &= \mathbb{I}(3z_4 + z_7^2 - 0.7 + \epsilon_2 > 0), && \epsilon_2 \sim N(0, 1)\\
y_3 &= \mathbb{I}(z_2^2 + 2z_6 + 4z_7 + \epsilon_3 > 6.2), && \epsilon_3 \sim N(0, 1)
\end{aligned}
$$

U stands for uniform distribution and N stands for normal distribution. We generate 2000 data points according to the function relationships on show and then discover the causal relations with our algorithm. The

Table 3: Comparison results (mean±std.) on six data sets. ↑(↓) indicates that the larger (smaller) the value, the better the performance; •(◦) indicates that MIMLcaus is significantly better(worse) than the corresponding method based on paired t-tests at 95%significance level.

	MIMLcaus	MIMLfast	MIMLkNN	MLR	KISAR
Bird song					
h.l. (↓)	0.054 ± 0.009	0.173 ± 0.011 •	0.089 ± 0.016 •	0.219 ± 0.013 •	0.107 ± 0.018 •
co. (↓)	1.681 ± 0.275	5.225 ± 0.517 •	2.420 ± 0.469 •	3.931 ± 0.600 •	2.677 ± 0.521 •
o.e. (↓)	0.055 ± 0.040	0.561 ± 0.072 •	0.163 ± 0.048 •	0.301 ± 0.077 •	0.205 ± 0.060 •
r.l. (↓)	0.026 ± 0.008	0.272 ± 0.030 •	0.074 ± 0.023 •	0.136 ± 0.024 •	0.085 ± 0.028 •
avg (↑)	0.938 ± 0.020	0.539 ± 0.037 •	0.852 ± 0.033 •	0.727 ± 0.041 •	0.822 ± 0.045 •
F1 (↑)	0.728 ± 0.042	0.088 ± 0.014 •	0.567 ± 0.040 •	0.398 ± 0.016 •	0.553 ± 0.104 •
MSRC v2					
h.l. (↓)	0.063 ± 0.010	0.094 ± 0.008 •	0.107 ± 0.007 •	0.076 ± 0.010 •	0.070 ± 0.007 •
co. (↓)	4.447 ± 0.643	6.583 ± 0.561 •	6.962 ± 0.835 •	4.742 ± 0.697	6.401 ± 0.660 •
o.e. (↓)	0.244 ± 0.059	0.357 ± 0.056 •	0.454 ± 0.092 •	0.264 ± 0.052	0.308 ± 0.078 •
r.l. (↓)	0.082 ± 0.016	0.146 ± 0.022 •	0.159 ± 0.028 •	0.086 ± 0.015	0.156 ± 0.020 •
avg (↑)	0.757 ± 0.037	0.617 ± 0.041 •	0.587 ± 0.055 •	0.731 ± 0.036	0.615 ± 0.052 •
F1 (↑)	0.466 ± 0.064	0.195 ± 0.036 •	0.165 ± 0.028 •	0.453 ± 0.035	0.401 ± 0.050 •
Letter Frost					
h.l. (↓)	0.102 ± 0.018	0.138 ± 0.020 •	0.126 ± 0.019 •	0.083 ± 0.022 ◦	0.114 ± 0.022 •
co. (↓)	7.328 ± 1.378	9.071 ± 1.110 •	9.621 ± 1.702 •	5.778 ± 1.112 ◦	9.835 ± 1.112 •
o.e. (↓)	0.142 ± 0.082	0.214 ± 0.121 •	0.128 ± 0.065	0.085 ± 0.094	0.142 ± 0.094
r.l. (↓)	0.114 ± 0.029	0.177 ± 0.019 •	0.179 ± 0.041 •	0.070 ± 0.024 ◦	0.198 ± 0.024 •
avg (↑)	0.769 ± 0.037	0.630 ± 0.029 •	0.713 ± 0.043 •	0.834 ± 0.046 ◦	0.643 ± 0.046 •
F1 (↑)	0.245 ± 0.053	0.052 ± 0.012 •	0.227 ± 0.041	0.444 ± 0.064 ◦	0.213 ± 0.064
Letter Carroll					
h.l. (↓)	0.135 ± 0.024	0.157 ± 0.024	0.206 ± 0.021 •	0.098 ± 0.018 ◦	0.152 ± 0.022
co. (↓)	10.63 ± 1.953	11.50 ± 1.293	14.22 ± 1.255 •	7.487 ± 2.064 ◦	13.02 ± 1.771
o.e. (↓)	0.225 ± 0.114	0.250 ± 0.102	0.606 ± 0.102 •	0.075 ± 0.049 ◦	0.243 ± 0.054
r.l. (↓)	0.189 ± 0.042	0.224 ± 0.025	0.360 ± 0.040 •	0.090 ± 0.036 ◦	0.311 ± 0.059 •
avg (↑)	0.670 ± 0.064	0.609 ± 0.042	0.417 ± 0.039 •	0.823 ± 0.048 ◦	0.544 ± 0.056
F1 (↑)	0.376 ± 0.073	0.244 ± 0.050 •	0.116 ± 0.036 •	0.548 ± 0.069 ◦	0.287 ± 0.051
Scene					
h.l. (↓)	0.178 ± 0.010	0.199 ± 0.013 •	0.171 ± 0.013 ◦	0.268 ± 0.012 •	0.187 ± 0.011 •
co. (↓)	0.989 ± 0.077	1.119 ± 0.079 •	0.937 ± 0.079 ◦	1.137 ± 0.076 •	1.069 ± 0.101 •
o.e. (↓)	0.333 ± 0.039	0.369 ± 0.032 •	0.324 ± 0.032	0.377 ± 0.038 •	0.365 ± 0.035 •
r.l. (↓)	0.179 ± 0.018	0.209 ± 0.017 •	0.166 ± 0.017 ◦	0.210 ± 0.018 •	0.198 ± 0.023 •
avg (↑)	0.783 ± 0.021	0.754 ± 0.017 •	0.793 ± 0.017 ◦	0.750 ± 0.021 •	0.761 ± 0.022 •
F1 (↑)	0.560 ± 0.029	0.569 ± 0.033	0.581 ± 0.033 ◦	0.570 ± 0.015	0.508 ± 0.032 •
News					
h.l. (↓)	0.095 ± 0.011	0.174 ± 0.014 •	0.162 ± 0.013 •	0.240 ± 0.010 •	0.147 ± 0.012 •
co. (↓)	2.114 ± 0.241	3.448 ± 0.333 •	2.587 ± 0.369 •	3.600 ± 0.312 •	3.019 ± 0.404 •
o.e. (↓)	0.350 ± 0.067	0.646 ± 0.070 •	0.510 ± 0.067 •	0.645 ± 0.064 •	0.527 ± 0.094 •
r.l. (↓)	0.159 ± 0.023	0.291 ± 0.037 •	0.216 ± 0.039 •	0.290 ± 0.028 •	0.256 ± 0.043 •
avg (↑)	0.735 ± 0.037	0.511 ± 0.044 •	0.631 ± 0.051 •	0.509 ± 0.033 •	0.594 ± 0.060 •
F1 (↑)	0.559 ± 0.060	0.077 ± 0.020 •	0.278 ± 0.036 •	0.132 ± 0.016 •	0.336 ± 0.065 •

result is shown in the Table 1. The entry at the i-th row and j-th column is equal to 1 if the variable z_j is considered the cause of the label y_i, and 0 otherwise. It can be observed that the discovered causal relations are consistent with the ground-truth relations.

4.2 Performance Comparison We then evaluate the learning performance on 6 MIML datasets, including Bird Song, MSRC v2, Letter Frost, Letter Carroll, Scene and News [8]. The News dataset is processed from the raw data [1]. Some rare labels are deleted and the dataset in our experiment has 612 documents from 10 classes: 'politics' , 'arts', 'sports', 'business', 'economy', 'terrorism', 'books', 'soccer', 'crime', 'at-war'. In a document, each sentence is represented by an instance of 50 dimensions, which is extracted with Word2Vec in the "gensim" package. The detailed characteristics about the datasets are summarized in Table 2.

We compare our approach MIMLcaus to some state-of-the-art approaches, including MIMLKNN [28], KiSar [13], MLR [20] and MIMLfast [12]. Six popular measures, i.e. hamming loss(*h.l.*), coverage(*co.*), one error(*o.e.*), ranking loss(*r.l.*), average precision(*a.p.*), macro-F1(*F1*) are used to evaluate the performance. Detailed description can be found in [21, 25, 32].

For each dataset, we separate it to ten groups in random. And then we train with nine groups and test with one group for ten times. For MIMLcaus, we set $\alpha = 0.05$ in the RESIT step. The parameters of other approaches is chosen according to related paper. The experiment result is given in Table 3.

We can observe that MIMLcaus outperforms other methods in most cases. It only loses to MIMLKNN on the Scene dataset, and MLR on the two letter datasets. It can be concluded that the causal relations contribute to the highly competitive performance. And exploiting the causations can help us get rid of the correlations between labels.

To validate that studying the causal relations is meaningful in further, in addition to the MIML tasks, we also test our algorithm in the multi-instance framework. The only difference is that only one label exists so that the proposed MIMLcaus approach can be directly applied. The compared approaches are MiSVM [1], MILES [3] and MIFV [24]. The datasets are collected from "20 Newsgroups". We divide every categorization data into ten groups. And we train with nine groups and test with one group for ten times. The accuracy of the compared methods is shown in Table 4. Again our MIMLcaus algorithm achieves the best performance on 19 of the 20 datasets.

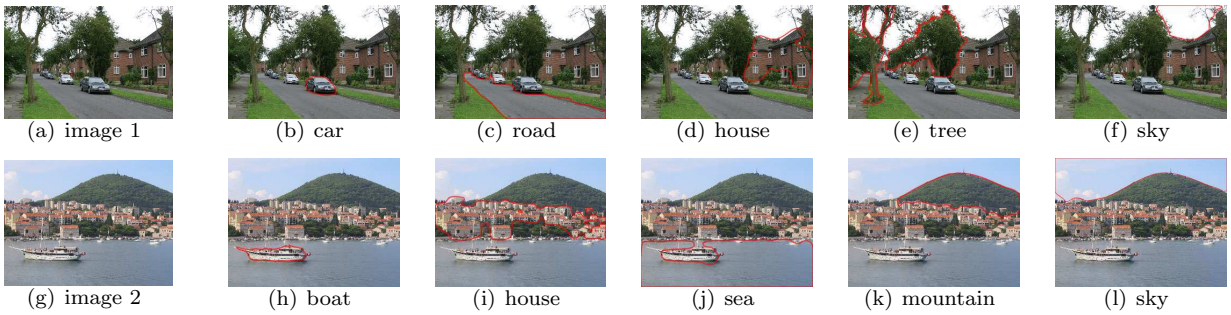
[1] https://github.com/kgohil/MultiLableClassification

Table 4: Accuracy comparison on the multi-instance datasets

	miSVM	MILES	MIFV	MIMLcaus
alt.atheism	0.67	0.48	0.72	**0.86**
comp.graphics	0.74	0.51	0.54	**0.82**
comp.os.ms-windows.misc	0.69	0.48	0.57	**0.72**
comp.sys.ibm.pc.hardware	0.77	0.51	0.51	**0.80**
comp.sys.mac.harware	0.73	0.48	0.52	**0.80**
comp.window.x	0.67	0.51	0.62	**0.81**
misc.forsale	0.63	0.48	0.60	**0.68**
rec.autos	0.64	0.48	0.60	**0.77**
rec.motorcycles	0.50	0.48	0.73	**0.81**
rec.sport.baseball	0.53	0.48	0.73	**0.84**
rec.sport.hockey	0.56	0.48	0.75	**0.88**
sci.crypt	0.64	0.48	0.69	**0.77**
sci.electronics	**0.91**	0.53	0.51	0.90
sci.med	0.59	0.48	0.72	**0.80**
sci.space	0.55	0.48	0.70	**0.79**
sci.religion.christian	0.50	0.48	0.74	**0.83**
talk.politics.guns	0.68	0.48	0.59	**0.73**
talk.politics.mideast	0.74	0.48	0.72	**0.80**
talk.politics.misc	0.74	0.48	0.65	**0.71**
talk.religion.misc	0.52	0.49	0.69	**0.80**

4.3 Results on Key Instance Detection One advantage of causal discovery in MIML is to understand what instances trigger a specific label. To provide a quantitive evaluation, we report the accuracy of key instance detection on Bird and MSRC-v2. These two datasets have the instance-level labels, and thus can be used to validate whether the key instance identified is associated with the class label. Given a label, we predict for each bag about which instance is most likely to cause the label. Then we examine whether the key instance has the label. Table 5 shows the comparison results between our algorithm and KISAR. It can be seen that MIMLcaus has a significant advantage on detecting the key instances.

In addition, we also provide a qualitative analysis of key instance detection on the text dataset News and image dataset MSRC-v2. In Table 6, we show some example causal pairs discovered on News dataset. For the label in the first column, we present the sentence corresponding to the key instance in the second column. Because of the limited space, we only show three examples. As we can see, it is reasonable that the detected key instance causes the corresponding label.



(a) image 1 (b) car (c) road (d) house (e) tree (f) sky

(g) image 2 (h) boat (i) house (j) sea (k) mountain (l) sky

Figure 2: The key instances detected for different labels of two example images on MSRC-v2.

Table 5: Key instance detection accuracy.

	MIMLcaus	KISAR
Bird song	0.676 ± 0.041	0.362 ± 0.107
MSRC-v2	0.704 ± 0.025	0.686 ± 0.045

Table 6: The causal instances discovered for some example labels on the News dataset.

label	**sentences corresponding to the key instance**
politics	Last year PolitiFact could find only eight Republicans in Congress, out of 278 in the caucus, who had made on-the-record comments accepting the reality of man-made global warming.
sports	Hellebuyck, playing in his second N.H.L. game, had a strong follow-up to his debut, a 3-1 victory over Minnesota on Friday.
business	Uber is spending a lot of that money on an aggressive plan to expand internationally.

In Figure 2, we show the key instances detected by our algorithm for different labels of two example images on MSRC-v2 dataset. It can be shown that the detected key instances well match the corresponding labels, which again validates that the proposed MIMLcaus approach can find the key instance that causes a specific label.

4.4 Examination on Causal Effect In causality literature, if one variable has a causal effect to another variable, then intervention on it will usually take a direct and significant influence to the variables caused by it. Based on this, we conduct an experiment to further examine the discovered causal relations. Specifically, for a specific label y^*, we first identify the key prototype variable z^* with the most positive causal effect to the label. Then for all the test bags without the label y^*, we modify z^* to the maximum possible value. We thus generate new bags which are more likely to have label y^* because it has a large value on the key prototype. Because we don't know the ground-truth of the modified sample, we use the label of its nearest neighbor in the training set to replace it. If most of generated samples are with label y^*, then it implies that the discovered variable z^* does have a significant causal effect to the label y^*. We then show the number of test bags that will have the expected label after modification on the key prototype variable. The results are shown in Table 7 with a comparison with KISAR. It can be observed that our MIMLcaus approach can make more labels change after modifying the key prototype variable, indicating that MIMLcaus discovers reasonable causal relations.

5 Conclusion

In this paper, we propose to study the causal relations between instances and labels for multi-instance multi-label learning. By exploiting prototypes in the instance to have a semantical representation of bags, we propose an effective algorithm with guaranteed asymptotically consistent to discover the feature variables that have direct causal effect to a specific label. These variables then are utilized to further improve the MIML classification model as well as detect the key instances. Extensive experiments validate that the proposed method can achieve superior performance and discover reasonable causal relations. In the future, more advanced MIML algorithms will be incorporated with causal discovery to further improve the interpretability of MIML models.

Acknowledgement. This research was supported by the National Science Foundation of China (61751306, 61876081).

Table 7: The number of bags with label changes after modification of the discovered causal variable.

dataset	MIMLcaus	KISAR
Bird song	162.7	64.8
MSRC v2	203.7	156.9
Letter Frost	118.6	57
Letter Carroll	221.6	83.5
Scene	398.2	155
News	50.6	46.9

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann, *Support vector machines for multiple-instance learning*, in NIPS, 2002.

[2] F. Briggs, X. Z. Fern, and R. Raich, *Context-aware MIML instance annotation: exploiting label correlations with classifier chains*, Knowledge and Information Systems, 43 (2015), pp. 53–79.

[3] Y. Chen, J. Bi, and J. Z. Wang, *MILES: multiple-instance learning via embedded instance selection*, IEEE TPMAI, (2006).

[4] Z. Chen, Z. Chi, H. Fu, and D. Feng, *Multi-instance multi-label image classification: A neural approach*, Neurocomputing, 99 (2013), pp. 298–306.

[5] D. M. Chickering, *Optimal structure identification with greedy search*, JMLR, 3 (2002), pp. 507–554.

[6] B. et al., *Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach*, The Journal of the Acoustical Society of America, (2012), pp. 4640–4650.

[7] N. Ghamrawi and A. McCallum, *Collective multi-label classification*, in CIKM, 2005, pp. 195–200.

[8] R. Goyal, *Natural language processing: Labelling new york times artices*, (2016).

[9] A. Gretton, O. Bousquet, A. J. Smola, and B. Schölkopf, *Measuring statistical dependence with hilbert-schmidt norms*, in ALT, 2005, pp. 63–77.

[10] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, *Nonlinear causal discovery with additive noise models*, in NIPS, 2008, pp. 689–696.

[11] S. Huang, Y. Yu, and Z. Zhou, *Multi-label hypothesis reuse*, in KDD, 2012, pp. 525–533.

[12] S.-J. Huang, W. Gao, and Z.-H. Zhou, *Fast multi-instance multi-label learning*, in AAAI, 2014, pp. 1868–1874.

[13] Y.-F. Li, J.-H. Hu, Y. Jiang, and Z.-H. Zhou, *Towards discovering what patterns trigger what labels*, in AAAI, 2012.

[14] J. M. Mooij, D. Janzing, J. Peters, and B. Schölkopf, *Regression by dependence minimization and its application to causal inference in additive noise models*, in ICML, 2009, pp. 745–752.

[15] N. Nguyen, *A new SVM approach to multi-instance multi-label learning*, in ICDM, 2010, pp. 384–392.

[16] J. Pearl, *Causality*, Cambridge University Press, 2009.

[17] J. Peters, D. Janzing, and B. Schölkopf, *Identifying cause and effect on discrete data using additive noise models*, in AISTATS, 2010, pp. 597–604.

[18] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf, *Causal discovery with continuous additive noise models*, JMLR, 15 (2014), pp. 2009–2053.

[19] A. T. Pham, R. Raich, and X. Z. Fern, *Simultaneous instance annotation and clustering in multi-instance multi-label learning*, in MLSP, 2015, pp. 1–6.

[20] ———, *Dynamic programming for instance annotation in multi-instance multi-label learning*, IEEE TPMAI, (2017).

[21] R. E. Schapire and Y. Singer, *Boostexter: A boosting-based system for text categorization*, Machine Learning, 39 (2000), pp. 135–168.

[22] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*, MIT Press, 2000.

[23] G. Tsoumakas, A. Dimou, E. Spyromitros, V. Mezaris, I. Kompatsiaris, and I. Vlahavas, *Correlation-based pruning of stacked binary relevance models for multi-label learning*, in Proceedings of the 1st International Workshop on Learning from Multi-label Data, 2009, pp. 101–116.

[24] X.-S. Wei, J. Wu, and Z.-H. Zhou, *Scalable algorithms for multi-instance learning*, IEEE TNNLS, (2017).

[25] X.-Z. Wu and Z.-H. Zhou, *A unified view of multi-label performance measures*, in ICML, 2017, pp. 3780–3788.

[26] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang, *Joint multi-label multi-instance learning for image classification*, in CVPR, 2008.

[27] K. Zhang and A. Hyvärinen, *On the identifiability of the post-nonlinear causal model*, in UAI, 2009, pp. 647–655.

[28] M.-L. Zhang, *A k-nearest neighbor based multi-instance multi-label learning algorithm*, in International Conference on Tools with Artificial Intelligence, 2010, pp. 207–212.

[29] M.-L. Zhang and Z.-H. Zhou, *M3MIML: A maximum margin method for multi-instance multi-label learning*, in ICDM, 2008, pp. 688–697.

[30] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, *Multi-instance learning by treating instances as non-i.i.d. samples*, in ICML, 2009, pp. 1249–1256.

[31] Z.-H. Zhou and M.-L. Zhang, *Multi-instance multi-label learning with application to scene classification*, in NIPS, 2006, pp. 1609–1616.

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