Multi-label Causal Feature Selection

Xingyu Wu[1], **Bingbing Jiang**[2], **Kui Yu**[3], **Huanhuan Chen**[1]*, **Chunyan Miao**[4]
[1]School of Computer Science and Technology, University of Science and Technology of China.
[2]Hangzhou Institute of Service Engineering, Hangzhou Normal University.
[3]School of Computer and Information, Hefei University of Technology.
[4]School of Computer Science and Engineering, Nanyang Technological University.
xingyuwu@mail.ustc.edu.cn, jiangbb@hznu.edu.cn, yukui@hfut.edu.cn, hchen@ustc.edu.cn, ascymiao@ntu.edu.sg.

Abstract

Multi-label feature selection has received considerable attentions during the past decade. However, existing algorithms do not attempt to uncover the underlying causal mechanism, and individually solve different types of variable relationships, ignoring the mutual effects between them. Furthermore, these algorithms lack of interpretability, which can only select features for all labels, but cannot explain the correlation between a selected feature and a certain label. To address these problems, in this paper, we theoretically study the causal relationships in multi-label data, and propose a novel Markov blanket based multi-label causal feature selection (MB-MCF) algorithm. MB-MCF mines the causal mechanism of labels and features first, to obtain a complete representation of information about labels. Based on the causal relationships, MB-MCF then selects predictive features and simultaneously distinguishes common features shared by multiple labels and label-specific features owned by single labels. Experiments on real-world data sets validate that MB-MCF could automatically determine the number of selected features and simultaneously achieve the best performance compared with state-of-the-art methods. An experiment in Emotions data set further demonstrates the interpretability of MB-MCF.

Introduction

In machine learning research, multi-label learning focuses on the problem that each instance is associated with multiple class labels simultaneously (Zhang and Zhou 2006), which is ubiquitous in many real-world applications, such as image annotation (Liu et al. 2018), text categorization (Liu et al. 2017), and gene function classification (Fodeh and Tiwari 2018). Similar to single-label learning, high dimensional data with an enormous amount of redundant features significantly increases the computational burden of multi-label learning, which could also lead to over-fitting and performance degradation of learning algorithms (Lin et al. 2015). Previous studies (Liu and Motoda 2007) have shown that only a subset of relevant features carry the most discriminative information. Thus, in recent years, many multi-label feature selection algorithms have been proposed to find a lower-dimensional representation of the original feature space, which can be broadly classified into transformation-based methods and direct methods (Pereira et al. 2018).

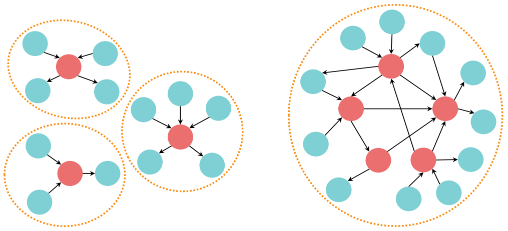

(a) Multiple irrelevant labels. (b) Multiple relevant labels.

Figure 1: Causal structures of two extreme cases. Labels are highlighted in red and features are highlighted in green.

Multi-label data contains three types of variable relationships, i.e., relationships between labels, between features, and between labels and features. Earlier feature selection methods, such as some transformation-based methods, only focus on the last two types of relationships, and transform a multi-label problem into one or several single-label problems with transformation techniques (Godbole and Sarawagi 2004; Read 2008). While some recent direct approaches revise traditional single-label feature selection algorithms to process the multi-label data directly, such as sub-feature uncovering with sparsity (SFUS) (Ma et al. 2012) and multi-label informed feature selection (MIFS) (Jian et al. 2016), which have taken label correlation into consideration. However, most of these algorithms consider the three types of relationships individually. For example, MIFS uses a matrix to encode label correlations only. Separately solving the three types of relationships would ignore the mutual effects between different types of relationships, which limits the effectiveness of feature selection. Simultaneously analyzing all the relationships needs to consider the underlying mechanism, while existing methods do not attempt to uncover it.

Another problem is that, existing methods lack of interpretability, i.e., cannot explain the correlation between a selected feature and a certain label. It is necessary to know which labels are influenced by a selected feature. For example, the recently presented topic, label-specific feature (Zhang and Wu 2015), aiming to the phenomenon that different class labels may carry specific characteristics of their

*Corresponding author: Huanhuan Chen.

own, will benefit from the interpretable feature section algorithm. Therefore, we need to propose a method that not only selects predictive features, but also distinguishes the common features shared by multiple labels and the label-specific features owned by some single labels.

To address above challenges, this paper investigates the multi-label feature selection from a causal perspective. The superiority of causality-aware methods reflects in two aspects. Firstly, causal mechanism treats labels and features as ordinary variables, focusing on the underlying cause-effect relationships between all variables. Therefore, by mining the causal mechanism, we can simultaneously consider all types of the relationships, and thereby solve the aforementioned first challenge. As depicted in Figure 1, using a directed acyclic graph (DAG) to represent the causal structure, we can intuitively 'read' from the causal structure that labels are all independent of each other in Figure 1 (a) but are all relevant to each other in Figure 1 (b). Similarly, the dependency between labels and features and the redundancy between different features can also be 'read' out. As for the interpretability, for any feature, it is easy to locate the labels influenced by a certain feature in the causal structure, which could help to distinguish the common features and label-specific features.

Given the superiority of causality, we need to find an effective means to represent the complex causal relationships. In this paper, Markov blanket (MB) is chosen, which is amenable to represent the local causal structure of a variable and has been used in single label causal feature selection (Yu, Liu, and Li 2018). In a faithful Bayesian network (BN), the MB of a variable consists of its parents (direct causes), children (direct effects) and spouses (other direct causes of direct effects). And given the MB of a variable, all other variables will be independent of this variable (Spirtes et al. 2000). In single-label learning, the MB of a label can be directly used as the selected feature set (Pellet and Elisseeff 2008). However, in multi-label scenario, directly using the MB of multiple labels is obviously unadvisable due to two problems. Firstly, real-world data always violates the faithfulness condition, and accordingly, some features will contain equivalent information (Statnikov et al. 2013) about labels, which might lead to more common features. Secondly, some relevant features might be excluded out of the local causal structure when there exist strong label relevances. In this paper, we take some DAGs as examples to elaborate these two problems and provide detailed theoretical analyses, which give an insight on algorithm design.

To solve these problems, we propose a MB-based multi-label causal feature selection algorithm (MB-MCF). MB-MCF first mines the local causal structure of each label to uncover the causal mechanism between both labels and features. Based on the causal relationships, MB-MCF then searches common features between relevant labels and label-specific features for single labels. The main contributions of this paper are summarized as follows:

1. Some theoretical contributions facilitate the multi-label causal feature selection research. 1) We analyze the causal structure for multi-label scenario, and find that the causal relationships between labels and its relevant features might be blocked by other strongly relevant labels. 2) Given the fact that real-world data sets violate the faithfulness condition, we study the equivalent information phenomenon in multi-label data, and formally give the property of the common features of multiple labels.
2. Different from existing methods, the proposed MB-MCF has at least three practical benefits: 1) Based on causal learning, MB-MCF simultaneously considers all dependencies between both labels and features, and selects features not only predictive but also causally informative. 2) MB-MCF possesses interpretability. 3) MB-MCF does not require the number of selected features to be predetermined. To the best of our knowledge, it is the first multi-label *causal* feature selection algorithm.

Notations and Definitions

In this paper, the capital letters (such as X) represent random variables and the lower-case letters (such as x) represent their values, the capital bold italic letters (such as $\boldsymbol{Z}$) denote variable sets. Specifically, let $\boldsymbol{U}$ denote the set of all the (discrete random) variables, and $\boldsymbol{T} = \{T_1, T_2, ..., T_l\} \subset \boldsymbol{U}$ denote the label set. In addition, the symbol $X \not\perp Y|\boldsymbol{Z}$ ($X \perp Y|\boldsymbol{Z}$) represents that variables X and Y are conditionally (in)dependent given a variable set **Z**. The symbol $I(X,Y)$ denotes the mutual information between X and Y.

*Definition 1 (**Bayesian network**)* (Pearl 1998). Let $\mathbb{P}$ denote the joint probability distribution over a variable set $\boldsymbol{U}$ of a directed acyclic graph (DAG) $\mathbb{G}$. The triplet $\langle \boldsymbol{U}, \mathbb{G}, \mathbb{P} \rangle$ constitutes a BN, if $\langle \boldsymbol{U}, \mathbb{G}, \mathbb{P} \rangle$ satisfies the Markov condition: every variable is independent of any subset including its non-descendant variables given its parents in $\mathbb{G}$.

*Definition 2 (**Faithfulness**)* (Spirtes et al. 2000). Given a BN $\langle \boldsymbol{U}, \mathbb{G}, \mathbb{P} \rangle$, $\mathbb{G}$ is faithful to $\mathbb{P}$ if and only if every conditional independence present in $\mathbb{P}$ is entailed by $\mathbb{G}$ and the Markov condition. $\mathbb{P}$ is faithful if and only if there exists a DAG $\mathbb{G}$ such that $\mathbb{G}$ is faithful to $\mathbb{P}$.

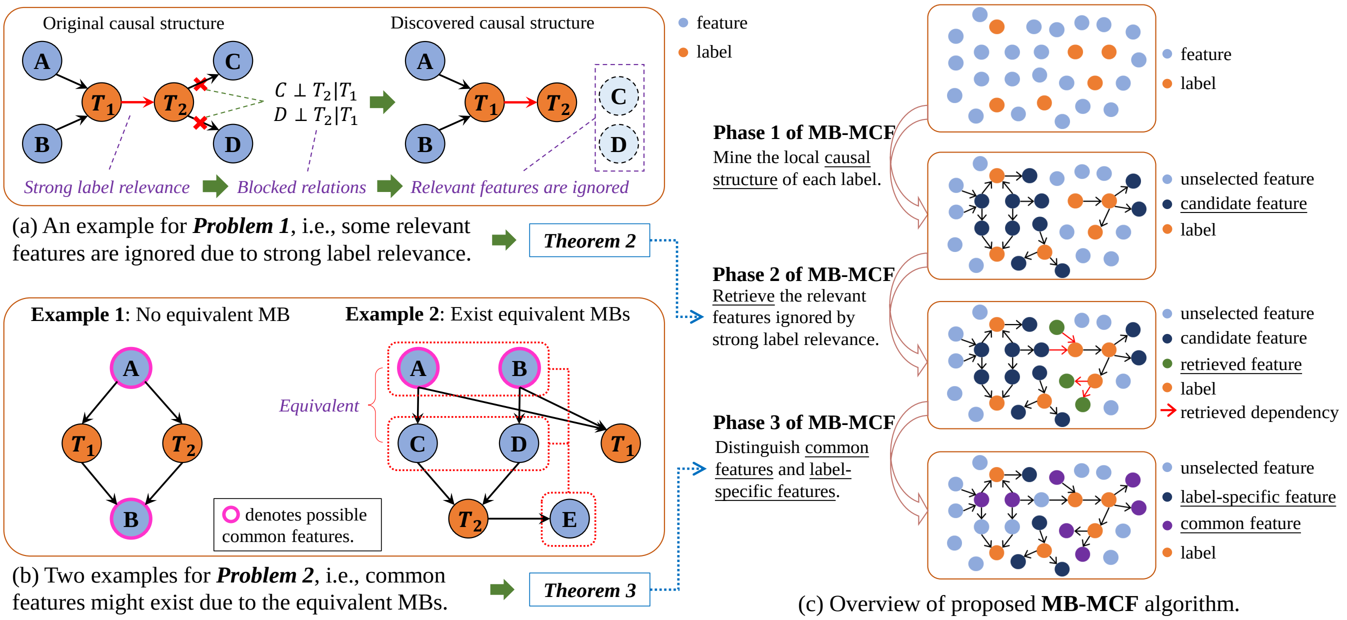
*Definition 3 (**Markov blanket**)* (Pearl 1998). In a faithful BN $\langle \boldsymbol{U}, \mathbb{G}, \mathbb{P} \rangle$, the Markov blanket of variable T in $\mathbb{G}$ is unique and consists of its parents, children, and spouses.

Theorem 1 (Pearl 1998). Given the $\boldsymbol{MB}(T)$, $X \perp T|\boldsymbol{MB}(T)$ for any $X \in \boldsymbol{U} - \boldsymbol{MB}(T) - \{T\}$.

Tsamardinos and Aliferis (2003) proved that MB is the theoretically optimal set of features if the faithfulness condition is satisfied, which confirms that we can transfer the feature selection problem into the MB discovery of the class attribute in a faithful BN. To understand the intuition in the perspective of causal leaning, we consider that the MB includes the direct causes (parents), direct effects (children), and other direct causes of direct effects (spouses) of the class attribute (Yu et al. 2019).

*Definition 4 (**Equivalent information**)* (Statnikov et al. 2013). Two subsets of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ from $\boldsymbol{U}$ contain equivalent information about a variable T iff the following conditions hold: $T \not\perp \boldsymbol{X}$, $T \not\perp \boldsymbol{Y}$, $T \perp \boldsymbol{X}|\boldsymbol{Y}$ and $T \perp \boldsymbol{Y}|\boldsymbol{X}$.

Equivalent information phenomenon occurs when the faithfulness condition is violated. It can be interpreted as $I(T,\boldsymbol{X}) = I(T,\boldsymbol{Y})$, where the symbol I denotes the mu-


Original causal structure
Discovered causal structure
A
B
C
D
T_1
T_2
$C \perp T_2|T_1$
$D \perp T_2|T_1$
Strong label relevance
Blocked relations
Relevant features are ignored
feature
label
Theorem 2
Example 1: No equivalent MB
Example 2: Exist equivalent MBs
Equivalent
E
○ denotes possible common features.
Theorem 3
Phase 1 of MB-MCF
Mine the local causal structure of each label.
Phase 2 of MB-MCF
Retrieve the relevant features ignored by strong label relevance.
Phase 3 of MB-MCF
Distinguish common features and label-specific features.
unselected feature
candidate feature
retrieved feature
retrieved dependency
label-specific feature
common feature

(a) An example for ***Problem 1***, i.e., some relevant features are ignored due to strong label relevance.

(b) Two examples for ***Problem 2***, i.e., common features might exist due to the equivalent MBs.

(c) Overview of proposed **MB-MCF** algorithm.

Figure 2: We discuss two problems for multi-label causal feature selection. (a) illustrates ***Problem 1***. When T_2 is completely dependent (red arrow) on T_1, the causal relationships (highlighted with red 'X') between T_2 and its causal features C and D might be blocked, making C and D excluded out of the local causal structure of T_2. To retrieve ignored features, we present ***Theorem 2***. (b) gives two examples to illustrate ***Problem 2***. For Example 2 in (b), if $\{A, B, E\}$ and $\{C, D, E\}$ are equivalent MBs of T_2, then A and B are possible common features of T_1 and T_2. To find more common features and decrease redundancy in feature set, we present ***Theorem 3*** to give the property of common features. (c) gives an overview of the proposed MB-MCF algorithm, where the phase 2 solves ***Problem 1*** based on ***Theorem 2***, and phase 2 solves ***Problem 2*** based on ***Theorem 3***.

tual information. In other words, $\boldsymbol{X}$ and $\boldsymbol{Y}$ contain the same information about T. In real-world applications, the faithfulness condition is usually not satisfied, and thus, equivalent information phenomenon is common in these data sets. In this paper, we will use the definition to help us find the redundant features in multi-label scenarios.

Problem Discussion and Analyses

As previously mentioned, MB is employed to represent the local causal structure of a variable. In single-label learning, the MB of a label is the minimal set which carries all the information about the class label, and thereby, is the theoretically optimal feature subset if the faithfulness condition is satisfied (Pellet and Elisseeff 2008). MB of multiple labels is the union of MB set of each label (Liu and Liu 2018), which can be formalized as:

$$\boldsymbol{MB}(\{T_1, T_2, ..., T_l\}) = \bigcup_{i=1}^{l} \boldsymbol{MB}(T_i) - \{T_1, T_2, ..., T_l\}. \quad (1)$$

However, there exist two problems making the MB of multiple labels unsuitable for direct use as a feature subset. We take some DAGs as examples to locate these two problems and further give some theoretical analyses to solve them.

Problem 1: *Dependencies between label and its relevant features might be blocked by other strongly relevant labels, which makes some relevant features be discarded.*

Using MB to define the variable relevance, two strongly relevant variables are included in the MB set of each other, while two weakly relevant variables are not, but have a connecting path (in BN) between them (Tsamardinos and Aliferis 2003). A special case might occur in strong relevance, called deterministic relations, where presence (or absence) of one variable implies presence (or absence) of the other, and vice versa. For example, labels $male$ and $female$ for a person can not coexist at the same instances. Statnikov et al. (2013) has proved that deterministic relations could result in equivalent information phenomenon (refer to *Definition 4*).

Due to this phenomenon, strongly relevant labels could block the dependencies between features and labels. Specifically, when a label and its relevant features contain equivalent information about another label, these relevant features might be excluded out of the local causal structure in the causal discovery process, mainly because the label is independent of these features given the other label. For example in Figure 2(a), assume that the relation between T_1 and T_2 is deterministic, and $\{C, D\}$ and T_2 contain equivalent information about T_1. If we implement a MB discovery algorithm to find the local causal structure of T_2, the dependence between T_2 and C, D will be tested as $T_2 \perp C|T_1$ and $T_2 \perp D|T_1$. Thus, C and D would be misjudged as non-parent-child variables and further be discarded as redundant features instead of causal features.

To solve this problem, we propose *Theorem 2* to give an insight on how to retrieve these ignored relevant features.

Theorem 2. Labels T_1 and T_2 are strongly relevant to each other. $\boldsymbol{MB}_i$ denotes the MB of T_i ($i \in \{1, 2\}$), and $\boldsymbol{S} \subset \boldsymbol{MB}_2$.

If T_2 and $\boldsymbol{S}$ contain equivalent information about T_1, then:

$$I(T_1, (\boldsymbol{MB}_1 - \{T_2\}) \cup \boldsymbol{S}) = I(T_1, \boldsymbol{MB}_1). \tag{2}$$

Proof. Since labels T_1 and T_2 are strongly relevant to each other, thus, $T_2 \in \boldsymbol{MB}_1$. According to the chain rule of mutual information (Cover and Thomas 2012), we have:

$$\begin{aligned}
&I(T_1, (\boldsymbol{MB}_1 - \{T_2\}) \cup \boldsymbol{S}) \\
&= I(T_1, \boldsymbol{S}|\boldsymbol{MB}_1 - \{T_2\}) + I(T_1, \boldsymbol{MB}_1 - \{T_2\}), \\
&I(T_1, \boldsymbol{MB}_1) \\
&= I(T_1, T_2|\boldsymbol{MB}_1 - \{T_2\}) + I(T_1, \boldsymbol{MB}_1 - \{T_2\}).
\end{aligned} \tag{3}$$

Then, subtracting $I(T_1, \boldsymbol{MB}_1 - \{T_2\})$ from both sides of the Eq. (2), we obtain:

$$I(T_1, \boldsymbol{S}|\boldsymbol{MB}_1 - \{T_2\}) = I(T_1, T_2|\boldsymbol{MB}_1 - \{T_2\}). \tag{4}$$

Eq. (2) will be proved by showing that Eq. (4) is established. Since T_2 and $\boldsymbol{S}$ contain equivalent information about T_1, thus $I(T_1, \boldsymbol{S}) = I(T_1, T_2)$. Therefore, Eq. (4) is established. We have thus proved the theorem. ■

In *Theorem 2*, features in feature subset $\boldsymbol{S}$ are the aforementioned discarded relevant features, i.e., $\{C, D\}$ in Figure 2(a). According to Eq. (2), if we remove the strongly relevant labels (i.e. T_2 in Eq. (2)) first and then search the causal features of T_1 again, then the discarded relevant features (i.e. $\boldsymbol{S}$ in Eq. (2)) will be retrieved. In the designed algorithm, we will retrieve the discarded relevant features in this way.

Problem 2: *More common features might exist due to the equivalent multiple MBs.*

Due to label relevance, some features are shared by multiple labels, called common features. For Example 1 in Figure 2(b), the common cause A, and the common effect B, are included in both $\boldsymbol{MB}(T_1)$ and $\boldsymbol{MB}(T_2)$. Obviously, A and B are common features of T_1 and T_2. Actually, since the real-world data usually violates the faithfulness condition, a label might have multiple MBs, and thereby, there exist common features as long as the intersection of one of the MB sets and the MB of another label is non-empty. We illustrate the problem with Example 2 in Figure 2(b). In the conventional sense, the intersection of $\boldsymbol{MB}(T_1)$, $\{A, B\}$, and $\boldsymbol{MB}(T_2)$, $\{C, D, E\}$, is empty. However, the common features might still exist only if $\{A, B, E\}$ is also the MB of T_2, and then $\{A, B\}$ is the common features of T_1 and T_2.

From the perspective of information theory, the above phenomenon in Example 2 in Figure 2(b) is mainly because $\{A, B\}$ and $\{C, D\}$ contain equivalent information about T_2. Intuitively, if all feature subsets containing equivalent information could be found, then we need not implement the time-consuming process to find all the MBs of a label. Nevertheless, there is no theoretical proof that has been presented to guarantee that there is no loss of information if part of the MB features are replaced with their equivalent features. In the following, we propose *Theorem 3* to illustrate this issue.

Theorem 3. Let $\boldsymbol{T} = \{T_1, T_2, ..., T_k\}$ denote the label subset, $\boldsymbol{MB}_i$ denote the MB of T_i ($i \in \{1, 2, ..., k\}$), $\boldsymbol{S}_i \subset \boldsymbol{MB}_i$. For $\forall i \in \{1, 2, ..., k\}$, if $\boldsymbol{S} \subset \boldsymbol{U}$ and $\boldsymbol{S}_i$ contain equivalent information about T_i, then $\boldsymbol{S}$ satisfy:

$$I(\bigcup_{i=1}^{k} \boldsymbol{MB}_i \cup \boldsymbol{S} - \bigcup_{i=1}^{k} \boldsymbol{S}_i, \boldsymbol{T}) = I(\bigcup_{i=1}^{k} \boldsymbol{MB}_i, \boldsymbol{T}). \tag{5}$$

And we call $\boldsymbol{S}$ is the ***common feature set*** of labels in $\boldsymbol{T}$.

Proof. It suffices to prove the case that $\boldsymbol{T} = \{T_1, T_2\}$, since any multi-label case is a direct consequence of two-label case using induction on the number of variables involved in $\mathbf{T}$. According to the *Definition 5*, we have:

$$I(\boldsymbol{S}, T_i) = I(\boldsymbol{S}_i, T_i). \tag{6}$$

We first prove that $I(\boldsymbol{MB}_i, T_i) = I(\boldsymbol{MB}_i \cup \boldsymbol{S} - \boldsymbol{S}_i, T_i)$. According to the chain rule, we have:

$$\begin{aligned}
&I(\boldsymbol{MB}_i \cup \boldsymbol{S}, T_i) = I((\boldsymbol{MB}_i \cup \boldsymbol{S} - \boldsymbol{S}_i) \cup \boldsymbol{S}_i, T_i) \\
&= I(\boldsymbol{MB}_i \cup \boldsymbol{S} - \boldsymbol{S}_i, T_i) + I(\boldsymbol{S}_i, T_i|\boldsymbol{MB}_i \cup \boldsymbol{S} - \boldsymbol{S}_i),
\end{aligned} \tag{7}$$

where $I(\boldsymbol{MB}_i \cup \boldsymbol{S}, T_i) = I(\boldsymbol{MB}_i, T_i)$ since $\boldsymbol{S} \perp T_i|\boldsymbol{MB}_i$, and $I(\boldsymbol{S}_i, T_i|\boldsymbol{MB}_i \cup \boldsymbol{S} - \boldsymbol{S}_i) = 0$ according to Eq. (6). Substituting them into Eq. (7), we get:

$$I(\boldsymbol{MB}_i \cup \boldsymbol{S} - \boldsymbol{S}_i, T_i) = I(\boldsymbol{MB}_i, T_i). \tag{8}$$

Let $\boldsymbol{A}_i = \boldsymbol{MB}_i \cup \boldsymbol{S} - \boldsymbol{S}_i$, then we transform Eq. (5) as:

$$I(\boldsymbol{A}_1 \cup \boldsymbol{A}_2, T_1 \cup T_2) = I(\boldsymbol{MB}_1 \cup \boldsymbol{MB}_2, T_1 \cup T_2). \tag{9}$$

According to the chain rules, we expand the left term and right term in Eq. (9) as $\mathcal{L}$ and $\mathcal{R}$, respectively.

$$\begin{aligned}
\mathcal{L} =& I(\boldsymbol{A}_1, T_1) + I(\boldsymbol{A}_2, T_2|\boldsymbol{A}_1, T_1) \\
&+ I(\boldsymbol{A}_1, T_2|T_1) + I(\boldsymbol{A}_2, T_1|\boldsymbol{A}_1), \\
\mathcal{R} =& I(\boldsymbol{MB}_1, T_1) + I(\boldsymbol{MB}_2, T_2|\boldsymbol{MB}_1, T_1) \\
&+ I(\boldsymbol{MB}_1, T_2|T_1) + I(\boldsymbol{MB}_2, T_1|\boldsymbol{MB}_1).
\end{aligned} \tag{10}$$

According to Eq. (8), we have $I(\boldsymbol{A}_1, T_1) = I(\boldsymbol{MB}_1, T_1)$ and $I(\boldsymbol{A}_2, T_2|\boldsymbol{A}_1, T_1) = I(\boldsymbol{MB}_2, T_2|\boldsymbol{MB}_1, T_1)$. According to the property of MB, we have $I(\boldsymbol{A}_2, T_1|\boldsymbol{A}_1) = I(\boldsymbol{MB}_2, T_1|\boldsymbol{MB}_1)$. We continue expand the third term in Eq. (10) as follows.

$$\begin{aligned}
&I(\boldsymbol{A}_1, T_2|T_1) = I(\boldsymbol{MB}_1 - \boldsymbol{S}_1, T_2) + I(\boldsymbol{S}, T_2|\boldsymbol{MB}_1 - \boldsymbol{S}_1), \\
&I(\boldsymbol{MB}_1, T_2|T_1) = I(\boldsymbol{MB}_1 \cup \boldsymbol{S}_1 - \boldsymbol{S}_1, T_2) \\
&= I(\boldsymbol{MB}_1 - \boldsymbol{S}_1, T_2) + I(\boldsymbol{S}_1, T_2|\boldsymbol{MB}_1 - \boldsymbol{S}_1).
\end{aligned}$$

Therefore, $I(\boldsymbol{A}_1, T_2|T_1) \geqslant I(\boldsymbol{MB}_1, T_2|T_1)$ according to Eq. (6), and thereby, $\mathcal{L} \geqslant \mathcal{R}$. According to Eq. (1), the MB set of $\{T_1, T_2\}$ is a subset of $\boldsymbol{MB}_1 \cup \boldsymbol{MB}_2$. Thus, we can directly prove that $\mathcal{L} \leqslant \mathcal{R}$ from Eq. (9) according to the property of MB. Hence, $\mathcal{L} = \mathcal{R}$ and Eq. (5) is established. ■

For better understanding, we map the elements in *Theorem 3* to the Example 2 in Figure 2(b). Feature set $\boldsymbol{S}$ is $\{A, B\}$ in Example 2, which contains the information of all labels. And $\boldsymbol{S}_1 = \{A, B\}$ (Note that, $\boldsymbol{S}$ and $\boldsymbol{S}_i$ need not be different.), $\boldsymbol{S}_2 = \{C, D\}$. And thereby, *Theorem 3* has proved that, if $\boldsymbol{S}$ and $\boldsymbol{S}_1$ contain equivalent information about T_1, and $\boldsymbol{S}$ and $\boldsymbol{S}_2$ contain equivalent information about T_2, then we can use $\boldsymbol{S}$ to simultaneously replace $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, without any information loss (as shown in Eq. (5)). Furthermore, the replacement process is interpretable, which points out the common features of multiple labels and label-specific features of single labels. For example, in Figure 2(b), removing $\{C, D\}$ and replacing it with $\{A, B\}$, we obtain the selected feature subset $\{A, B, E\}$, in which E is a specific feature of T_2, while A and B are common features of $\{T_1, T_2\}$. The idea in *Theorem 3* will be used in the proposed algorithm.

Our Algorithm

In this section, we propose the MB-based multi-label causal feature selection algorithm (MB-MCF, Algorithm 1) for detail. As shown in Figure 2(c), MB-MCF has three phases: Phase 1 (lines 2-4) mines the causal mechanism in data. Phase 2 (lines 5-14) retrieves the ignored relevant features influenced by the strongly relevant labels, which solves the aforementioned ***Problem 1*** with *Theorem 2*. Phase 3 (lines 15-23) finds the common features and the label-specific features, which uses *Theorem 3* to solve the ***Problem 2***.

Phase 1 (lines 2-4). This phase employs an up-to-date divide-and-conquer-based MB discovery algorithm $\mathbb{A}$ (such as CCMB (Wu et al. 2019)) to get the local causal structure of each label. We only find the direct causes and effects (i.e., $\boldsymbol{PC}_i$) of each label T_i, since the direct causes and effects carry most of the information about labels while the spouse discovery process is time-consuming. Note that, in the discovery process, we do not distinguish labels and features but consider them as ordinary variables. Thus, $\boldsymbol{PC}_i$ provides the relationships not only between labels and features but also between different labels.

Phase 2 (lines 5-14). This phase retrieves the features ignored in the Phase 1, which is influenced by strong label relevance as illustrated in ***Problem 1***. Directly testing the deterministic relations between labels is difficult. Nevertheless, testing strong label relevance is feasible, which can be 'read' from the local causal structure of a label (as $T_i \in \boldsymbol{PC}_j$ in line 5). According to *Theorem 2*, we can find the ignored features through removing the strongly relevant labels and retesting the dependencies between the ignored features and target label. Concretely, line 5 traverses all strongly relevant label pairs T_i, T_j and line 6 traverses all possible ignored features F_k. Given a conditioning set $\boldsymbol{Z}$ which includes the strongly relevant label T_j, if a feature F_k is independent of the target label T_i (line 7), then F_k might be the ignored feature. A further test is implemented in line 8, if any conditioning set $\boldsymbol{Z}$ which dose not include T_j can not block the dependency between the feature F_k and the target label T_j, then, according to Eq.(2), we can assert that F_k is ignored in the Phase 1 and retrieve it in line 9. Finally, in line 13, we need to remove the strongly relevant labels out of the feature set as preparation for next phase.

Phase 3 (lines 15-23). Based on the local causal structure obtained from the preceding phases, we find common features and label-specific features in this phase. According to *Theorem 3*, feature sets containing equivalent information can be used to replace each other without any information loss. Thus, we first find the equivalent feature sets for each label in lines 15-21. For each label (line 15), line 17 finds the feature set $\boldsymbol{Z}$ such that $\boldsymbol{Z}$ and $\boldsymbol{S}$ (a subset of $\boldsymbol{PC}_j$) contain equivalent information about label T_i. Actually, to improve efficiency, an upper limit should be set for the size of $\boldsymbol{Z}$ since large-size $\boldsymbol{Z}$ can be derived from the small-size $\boldsymbol{Z}$. According to *Definition 3*, $\boldsymbol{Z}$ and $\boldsymbol{S}$ meeting the condition in line 17 can be considered as the equivalent feature sets of target label T_i, thus, they will be recorded in equivalent feature table $\boldsymbol{EIF}_i$ of T_i in line 18. As we get $\boldsymbol{EIF}_i$ for all T_i, the common feature discovery problem can be transformed to a search problem, that is, searching the minimal set ***SelFea*** such that for $\forall i, \exists <\boldsymbol{S},\boldsymbol{Z}> \in \boldsymbol{EIF}_i, (\boldsymbol{PC}_i - \boldsymbol{S} \cup \boldsymbol{Z}) \subset$ ***SelFea*** (line 22). The constraint can be rephrased as, for label T_i, if we replace some of the features in $\boldsymbol{PC}_i$ with its equivalent feature set, then at least one of the substituted $\boldsymbol{PC}_i$ must be included in the selected feature subset ***SelFea***, to guarantee that there is no information loss of a label. The minimal set satisfying the constraint is the optimal feature subset. A greedy algorithm can be used to find an optimal or suboptimal solution. Note that, in the process of searching ***SelFea***, we can record the relationship between each selected feature and each label. For example in Figure 2(b), assuming $\boldsymbol{EIF}_{T_2} = \{<\{A,B\},\{C,D\}>\}$, then $\{A,B\}$ is recorded as the common features of $\{T_1,T_2\}$ when $\{C,D\}$ is replaced with $\{A,B\}$ in line 22. And E is a label-specific feature of T_2.

Algorithm 1 The *MB-MCF* Algorithm.

```
1: Input: Labels set $\boldsymbol{T} = \{T_1, T_2, \ldots, T_l\}$, features set $\boldsymbol{F} = \{F_1, F_2, \ldots, F_m\}$, a divide-and-conquer-based MB discovery algorithm $\mathbb{A}$, significance level $\alpha$.
   {Phase 1: Get the local causal structure of each label.}
2: for $i = 1 \ldots l$ do
3:     $\boldsymbol{PC}_i \leftarrow$ Find direct causes and effects of $T_i$ from $\boldsymbol{T} \cup \boldsymbol{F} - \{T_i\}$ with the parent-child discovery process of $\mathbb{A}$.
4: end for
   {Phase 2: Identify the strong label relevance and retrieve the ignored features.}
5: for $i, j = 1 \ldots l$ and $T_i \in \boldsymbol{PC}_j$ do
6:     for $k = 1 \ldots m$ do
7:         if $\exists \boldsymbol{Z} : \{T_i\} \subset \boldsymbol{Z} \subset \boldsymbol{PC}_j$ s.t. $F_k \perp T_j | \boldsymbol{Z}$ then
8:             if $\forall \boldsymbol{Z} \subset \boldsymbol{PC}_j - \{T_i\}$ s.t. $F_k \not\perp T_j | \boldsymbol{Z}$ then
9:                 $\boldsymbol{PC}_j = \boldsymbol{PC}_j \cup \{F_k\}$
10:            end if
11:        end if
12:    end for
13:    $\boldsymbol{PC}_j = \boldsymbol{PC}_j - \{T_i\}$
14: end for
   {Phase 3: Find common features and label-specific features.}
15: for $i = 1 \ldots l$ do
16:     for each $\boldsymbol{Z} \subset \boldsymbol{F} - \boldsymbol{PC}_i$ and $\boldsymbol{Z} \not\perp T_i$ do
17:         if $\exists \boldsymbol{S} \subset \boldsymbol{PC}_i$ s.t. $T_i \perp \boldsymbol{Z} | \boldsymbol{S}$ and $T_i \perp \boldsymbol{S} | \boldsymbol{Z}$ then
18:             $\boldsymbol{EIF}_i = \boldsymbol{EIF}_i \cup \{<\boldsymbol{S}, \boldsymbol{Z}>\}$
19:         end if
20:     end for
21: end for
22: Search the minimal set SelFea s.t. $\forall i, \exists <\boldsymbol{S}, \boldsymbol{Z}> \in \boldsymbol{EIF}_i, (\boldsymbol{PC}_i - \boldsymbol{S} \cup \boldsymbol{Z}) \subset$ SelFea.
23: Output: Selected feature subset SelFea.
```

Experiments

Experimental Settings

In this section, we present the experimental studies of the proposed MB-MCF algorithm on real-world data sets, which are from diverse application domains. Table 1 displays the details of the five multi-label data sets, including domain, standard statistics, and sizes of divided training and test data of each data set. In which, $cardinality$ denotes the average number of labels for per instance, and $density$ normalizes the label cardinality by the number of labels.

Table 1: Details of the multi-label data sets.

Data set	domain	#Training	#Test	#Features	#Labels	$cardinality$	$density$
Birds	audio	500	100	260	19	1.014	0.053
CAL500	music	300	100	68	174	26.044	0.150
EUR-Lex	text	5000	2000	5000	201	2.213	0.011
Mediamill	video	1000	1000	120	101	4.376	0.043
NUS-WIDE	images	10000	5000	500	81	1.869	0.023

Five state-of-the-art multi-label feature selection algorithms are compared, including SFUS (Ma et al. 2012), CSF-S (Chang et al. 2014), MIFS (Jian et al. 2016), CMFS (Braytee et al. 2017) and MCLS (Huang, Jiang, and Sun 2018). In addition, we also use the original data with no feature selection as a baseline in each experiment. To evaluate the effectiveness of the proposed methods, we employ a representative multi-label classification algorithm, ML-kNN (Zhang and Zhou 2007), to compute the classification accuracies archived by using selected features, and the number of nearest neighbors k is set to 10 with default setting.

Due to space limitation, we choose an example-based metric $HammingLoss$ and two label-based metrics F_{Macro} and F_{Micro} (macro-averaging and micro-averaging of F1-measure) to measure the performances of multi-label classification algorithm with selected features of each algorithm. $HammingLoss$ evaluates the fraction of misclassified instance-label pairs:

$$HammingLoss = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{q}|Z_i \Delta Y_i|. \tag{11}$$

where p and q denote the number of instances and labels, respectively. Z_i represents the predicted label set and Y_i is the correct label set in the i-th instance, and Δ stands for the symmetric difference between the two sets.

F_{Micro} can be considered as a weighted average of F1-measure over all q labels, while F_{Macro} is an arithmetic average of all output labels, which can be calculated by:

$$F_{Micro} = \frac{1}{q}\sum_{i=1}^{q}\frac{2TP_i}{2TP_i + FP_i + FN_i}. \tag{12}$$

$$F_{Macro} = \frac{\sum_{i=1}^{q} 2TP_i}{\sum_{i=1}^{q}(2TP_i + FP_i + FN_i)}. \tag{13}$$

where TP_i, FP_i and FN_i denote the number of true positives, false positives and false negatives in the i-th class label, respectively.

Performance Comparison

In this experiment, we first apply MB-MCF and other comparing algorithms to select features and then use ML-KNN to train a classifier with the selected features. Each experiment is repeated 10 times with different training and test data, and we report the average performances, i.e., $HammingLoss$, F_{Macro} and F_{Micro}. Since these comparing feature selection algorithms can not determine the optimal number of features, we gradually increase the percentage of the selected features from 2% to 20% with a step of 2%. For a fair comparison, the regularization parameters for all comparing algorithms are tuned from $\{0.01, 0.1, 0.3, \ldots, 0.9, 1\}$ by grid search. For the proposed MB-MCF, we employ Hiton-MB (Aliferis, Tsamardinos, and Statnikov 2003) as the MB discovery algorithm and use the G^2-test (Pearl 1998) to implement the conditional independence tests.

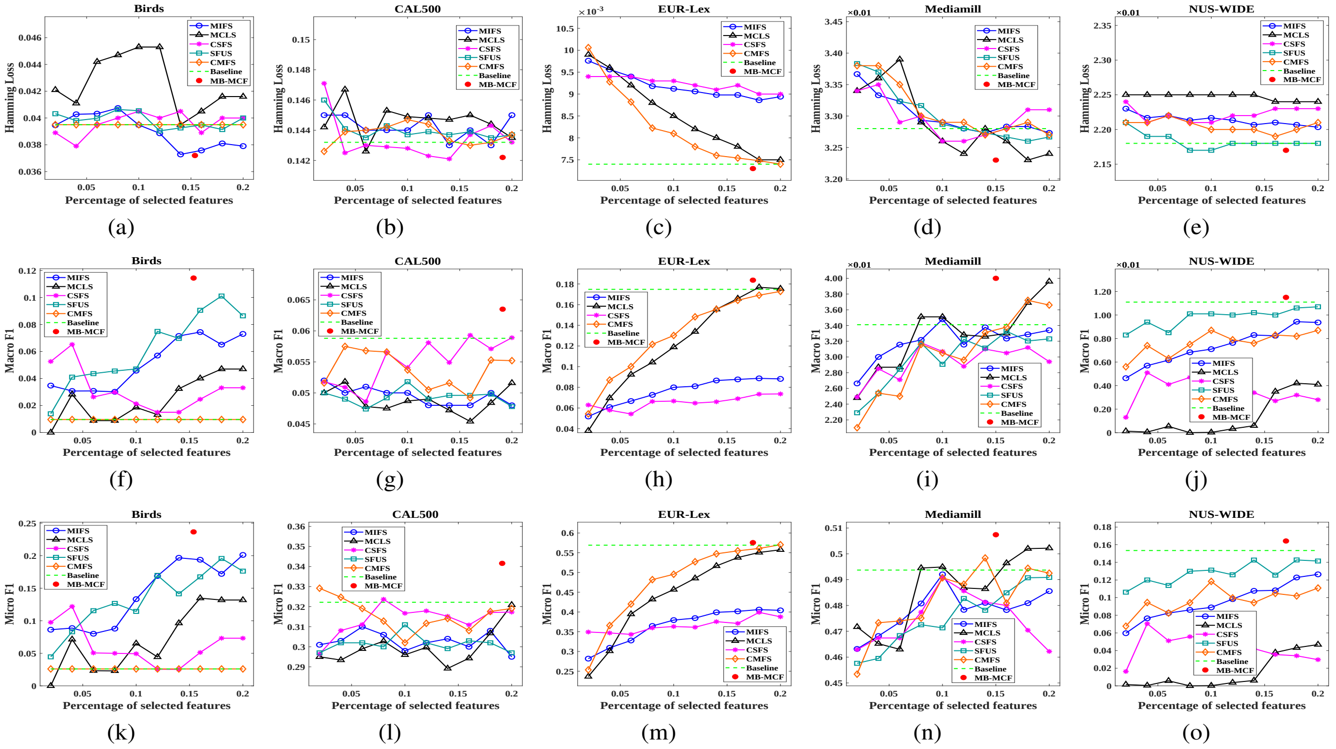
Table 2: The number of features selected by MB-MCF.

Data set	Birds	CAL500	EUR-Lex	Mediamill	NUS-WIDE
#Features	40	13	871	18	85

Figure 3 shows the average $HammingLoss$, F_{Macro} and F_{Micro} variation curves of different multi-label feature selection algorithms with respect to the percentage of selected features. As shown in Table 2 and Figure 3, our MB-MCF does not vary with the increasing selected feature percentage and can automatically determine the number of selected features, which is different from other algorithms requiring to predetermine the number of selected features. Clearly, MB-MCF consistently outperforms other algorithms in terms of each metric under the same number of selected features. Moreover, compared with the best result of the state-of-the-art methods, MB-MCF still achieves better or very competitive performances. Specifically, in Birds, CAL500 and Mediamill data sets, MB-MCF achieves significantly higher F_{Macro}, F_{Micro} and very competitive $HammingLoss$ compared with the state-of-the-art methods, which demonstrates that MB-MCF captures more effective features by considering the causal information between both features and labels. In large-scale data set EUR-Lex, SFUS can not be conducted on a 16-GB memory due to its high space complexity. From Figure 3 (c), (h), (m), we observe that the performances of state-of-the-art methods vary first and then tend to stable with the increase of the percentage of selected features, and the size of the feature set selected by MB-MCF exactly falls nearby the turning point, which demonstrates the effectiveness of MB-MCF to automatically determine the number of features. In large-scale data set NUS-WIDE, most of existing algorithms do not reach the performance of baseline since the insufficient training data influences the effectiveness of feature selection. However, MB-MCF is the only algorithm outperforming the baseline even when a few number of samples are used to train, which shows that MB-MCF is more data-efficient.

Interpretability

Compared with traditional multi-label feature selection methods, MB-MCF can select useful features and simultaneously possess interpretability. To illustrate the interpretability of MB-MCF, we implement MB-MCF on Emotions data set and provide the detail relationships between labels and selected features obtained from lines 15-22 in MB-MCF. The Emotions data set contains 6 labels, namely amazed-surprised (L_1), happy-pleased (L_2), relaxing-calm (L_3), quiet-still (L_4), sad-lonely (L_5), and angry-aggressive (L_6). We employ MB-MCF to select features on Emotions with 500 training samples, and record the relationship between each selected feature and each label as shown in Fig-



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

(k) (l) (m) (n) (o)

Figure 3: The $HammingLoss$, F_{Macro} and F_{Micro} of MB-MCF and other state-of-the-art algorithms. The result of MB-MCF is a red dot instead of a curve since MB-MCF could automatically determine the number of selected features.

ure 4. The selected features are listed at the top of the figure in the order of the serial numbers in Emotions data set, and each feature corresponds to a column of the grid, each label corresponds to a row. If a feature has an effect on a label, then the corresponding cell is dyed.

From Figure 4, we observe that features F_{21} and F_{40} carry the information about all labels, and F_2 is a label-specific feature of label L_2. From the distribution of shaded cells, we can conclude that the labels in label pairs (L_1, L_4), (L_2, L_5) and (L_3, L_6) have similar common features, and thereby, they have stronger correlation with each other than other label pairs, which is consistent with the Tellegen-Watson-Clark model [1] of mood in previous study (Tellegen, Watson, and Clark 1999).

On Emotions, MB-MCF achieves similar performance with existing methods (the details are not provided due to space limitation). However, MB-MCF can not only effectively select the relevant features containing discriminative information, but also simultaneously explain the relationships between variables (including both labels and features).

Conclusion

This paper investigates multi-label feature selection problem in the causal perspective. We study the causal structure of multi-label data, and discover that strong label relevance might block the dependencies between label and its relevant features. Furthermore, we present the property of common features shared by multiple labels. Based on the above theoretical contributions, we propose a novel algorithm, MB-MCF, which mines the causal mechanism first and then finds common features and label-specific features. Compared with traditional multi-label feature selection algorithms, MB-MCF possesses interpretability, and selects features not only predictive but also causally informative, while it does not require the number of selected features to be predetermined. Finally, we conduct extensive experiments to validate the effectiveness and superiority of proposed MB-MCF. Future work could strengthen the experimental evaluations and develop a joint algorithm (Jiang et al. 2019) to simultaneously select features and learn a classifier with the selected causal features.

	F_2	F_{18}	F_{21}	F_{23}	F_{24}	F_{25}	F_{29}	F_{31}	F_{39}	F_{40}	F_{46}	F_{48}	F_{53}	F_{54}	F_{65}	F_{70}
L_1		■	■	■		■				■		■	■		■	
L_2	■		■					■		■		■	■	■		■
L_3			■	■	■				■	■	■					■
L_4		■	■	■		■	■			■		■	■		■	
L_5			■					■		■		■	■	■		■
L_6			■	■	■		■		■	■	■					■

Figure 4: The relation between each selected feature (corresponding to a column) and each label (corresponding to a row) in Emotions. The shaded cell indicates that the corresponding feature has an effect on the corresponding label.

[1] The model was employed for labeling the data in the Emotions.

Acknowledgements

This research is supported in part by the National Key Research and Development Program of China under Grant No. 2016YFB1000905, the National Natural Science Foundation of China under Grant No. 61876206, 91846111, and 91746209.

References

Aliferis, C. F.; Tsamardinos, I.; and Statnikov, A. 2003. HITON: a novel Markov blanket algorithm for optimal variable selection. In *Proceedings of the American Medical Informatics Association Annual Symposium*, 21–25.

Braytee, A.; Liu, W.; Catchpoole, D. R.; and Kennedy, P. J. 2017. Multi-label feature selection using correlation information. In *Proceedings of the ACM Conference on Information and Knowledge Management*, 1649–1656.

Chang, X.; Nie, F.; Yang, Y.; and Huang, H. 2014. A convex formulation for semi-supervised multi-label feature selection. In *Proceedings of 28th AAAI Conference on Artificial Intelligence*, 1171–1177.

Cover, T. M., and Thomas, J. A. 2012. *Elements of information theory*. John Wiley & Sons.

Fodeh, S. J., and Tiwari, A. 2018. Exploiting MEDLINE for gene molecular function prediction via NMF based multi-label classification. *Journal of Biomedical Informatics* 86(10):160–166.

Godbole, S., and Sarawagi, S. 2004. Discriminative methods for multi-labeled classification. In *Proceedings of Pacific-Asia conference on knowledge discovery and data mining*, 22–30.

Huang, R.; Jiang, W.; and Sun, G. 2018. Manifold-based constraint laplacian score for multi-label feature selection. *Pattern Recognition Letters* 112(9):346–352.

Jian, L.; Li, J.; Shu, K.; and Liu, H. 2016. Multi-label informed feature selection. In *Proceedings of 26th International Joint Conference on Artificial Intelligence*, 1627–1633.

Jiang, B.; Wu, X.; Yu, K.; and Chen, H. 2019. Joint semi-supervised feature selection and classification through Bayesian approach. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*.

Lin, Y.; Hu, Q.; Liu, J.; and Duan, J. 2015. Multi-label feature selection based on max-dependency and min-redundancy. *Neurocomputing* 168(11):92–103.

Liu, X.-Q., and Liu, X.-S. 2018. Markov blanket and Markov boundary of multiple variables. *Journal of Machine Learning Research* 19(1):1658–1707.

Liu, H., and Motoda, H. 2007. *Computational methods of feature selection*. CRC Press.

Liu, J.; Chang, W.-C.; Wu, Y.; and Yang, Y. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM Conference on Research and Development in Information Retrieval*, 115–124.

Liu, Y.; Wen, K.; Gao, Q.; Gao, X.; and Nie, F. 2018. SVM based multi-label learning with missing labels for image annotation. *Pattern Recognition* 78(6):307–317.

Ma, Z.; Nie, F.; Yang, Y.; Uijlings, J. R.; and Sebe, N. 2012. Web image annotation via subspace-sparsity collaborated feature selection. *IEEE Transactions on Multimedia* 14(4):1021–1030.

Pearl, J. 1998. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann.

Pellet, J.-P., and Elisseeff, A. 2008. Using Markov blankets for causal structure learning. *Journal of Machine Learning Research* 9(7):1295–1342.

Pereira, R. B.; Plastino, A.; Zadrozny, B.; and Merschmann, L. H. 2018. Categorizing feature selection methods for multi-label classification. *Artificial Intelligence Review* 49(1):57–78.

Read, J. 2008. A pruned problem transformation method for multi-label classification. In *Proceedings of New Zealand Computer Science Research Student Conference*, 143–150.

Spirtes, P.; Glymour, C. N.; Scheines, R.; Heckerman, D.; Meek, C.; Cooper, G.; and Richardson, T. 2000. *Causation, prediction, and search*. MIT press.

Statnikov, A.; Lytkin, N. I.; Lemeire, J.; and Aliferis, C. F. 2013. Algorithms for discovery of multiple Markov boundaries. *Journal of Machine Learning Research* 14(2):499–566.

Tellegen, A.; Watson, D.; and Clark, L. A. 1999. On the dimensional and hierarchical structure of affect. *Psychological Science* 10(4):297–303.

Tsamardinos, I., and Aliferis, C. F. 2003. Towards principled feature selection: relevancy, filters and wrappers. In *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*.

Wu, X.; Jiang, B.; Yu, K.; Miao, C.; and Chen, H. 2019. Accurate markov boundary discovery for causal feature selection. *IEEE Transactions on Cybernetics, to be published, doi:10.1109/TCYB.2019.2940509*.

Yu, K.; Li, J.; Ding, W.; and Le, T. D. 2019. Multi-source causal feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, to be published, doi:10.1109/TPAMI.2019.2908373*.

Yu, K.; Liu, L.; and Li, J. 2018. A unified view of causal and non-causal feature selection. *[Online] Available: https://arxiv.org/abs/1802.05844*.

Zhang, M.-L., and Wu, L. 2015. Lift: multi-label learning with label-specific features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(1):107–120.

Zhang, M.-L., and Zhou, Z.-H. 2006. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering* 18(10):1338–1351.

Zhang, M.-L., and Zhou, Z.-H. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition* 40(7):2038–2048.