Multi-Target Markov Boundary Discovery: Theory, Algorithm, and Application

Xingyu Wu, Bingbing Jiang, Yan Zhong, and Huanhuan Chen, *Senior Member, IEEE*

Abstract—Markov boundary (MB) has been widely studied in single-target scenarios. Relatively few works focus on the MB discovery for variable set due to the complex variable relationships, where an MB variable might contain predictive information about several targets. This paper investigates the multi-target MB discovery, aiming to distinguish the common MB variables (shared by multiple targets) and the target-specific MB variables (associated with single targets). Considering the multiplicity of MB, the relation between common MB variables and equivalent information is studied. We find that common MB variables are determined by equivalent information through different mechanisms, which is relevant to the existence of the target correlation. Based on the analysis of these mechanisms, we propose a multi-target MB discovery algorithm to identify these two types of variables, whose variant also achieves superiority and interpretability in feature selection tasks. Extensive experiments demonstrate the efficacy of these contributions.

Index Terms—Markov boundary (MB), Markov blanket, feature selection, common MB variable, target-specific MB variable

1 Introduction

Markov boundary (MB) is of fundamental importance in statistical machine learning, which contains critical information about a given target. As shown in Fig. 1, in a faithful Bayesian network (BN), MB consists of the *parents*, *children*, and *spouses (other parents of the children)* of the target [1], [2], [3]. MB variables have the potential ability to imply the underlying mechanism around the target [1], [2], and are widely applied to real-world tasks. For example, MB discovery is the first step in BN structure learning, where the skeleton of the BN without orientation is constructed by MB [4], [5], [6]. Another important application is feature selection [7], [8], since all other features are independent of the class attribute conditioned on its MB [7]. Some studies [9], [10], [11], [12] have proved that the MB set is the theoretically optimal subset for learning and inference tasks. Due to the practical benefits, extensive algorithms are proposed to search the MB of a single target. Some of these algorithms [13], [14], [15], [16], [17], [18], [19] learn the MB based on Unique MB Assumption[1] [1], which is always violated in real-world data. Other algorithms [12], [16], [20] relax this assumption to detect multiple MBs of a target, whereas it is still intractable to find all of the possible MBs due to the unpredictable number of MBs.

However, few works consider the MB discovery for a variable set despite the ubiquity of multi-target data. This problem occurs when the joint probability distribution of several targets conditioned on other variables is analyzed, such as common features discovery of multiple targets, dimensionality reduction for multi-label learning, etc. Contrary to single-target scenarios, multi-target scenarios involve extra relationships between multiple targets, leading to two types of MB variables. As shown in Fig. 1, some MB variables simultaneously contain the predictive information about several targets, which are called common MB variables in the following, and correspondingly, others in the MB set are called target-specific MB variables. Both of them can facilitate the comprehension of the underlying mechanism, yet their focuses are different. Intuitively, target-specific MB variables reflect the differences among the local relations around different targets, which would assist the prediction or inference tasks on their corresponding targets [21]. While the common MB variables represent the connections between these targets, which are naturally capable of providing information about multiple targets with the minimal number of variables. Hence, they have extensive application prospects in dimension reduction, such as multi-label feature selection [22]. To drive a biased model[2] where different types of variables can contribute to

- *Xingyu Wu and Huanhuan Chen are with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China. E-mail: xingyuwu@mail.ustc.edu.cn, hchen@ustc.edu.cn.*
- *Bingbing Jiang is with the School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China. E-mail: jiangbb@hznu.edu.cn.*
- *Yan Zhong is with the School of Data Science, University of Science and Technology of China, Hefei 230027, China. E-mail: zhongyan@mail.ustc.edu.cn.*

Manuscript received 5 November 2020; revised 1 December 2021; accepted 15 August 2022. Date of publication 18 August 2022; date of current version 6 March 2023.
This work was supported in part by the National Key R&D Program of China under Grant 2021ZD0111700, in part by NSFC under Grants 62176245, 62137002, and 62006065, in part by Anhui Province under Grants 202104a05020011 and 202103a07020002, and in part by Fundamental Research Funds for the Central Universities.
(Corresponding author: Huanhuan Chen.)
Recommended for acceptance by R. Dechter.
Digital Object Identifier no. 10.1109/TPAMI.2022.3199784

1. A basic assumption of the MB discovery, supposing that each target has a unique MB set (Refer to Theorem 2 in *Section 2* for details).
2. A biased model means that different types of variables in the model have different effects on it. For example, in a prediction model for a certain target, target-specific MB variables of this target have a greater impact on predicting results.

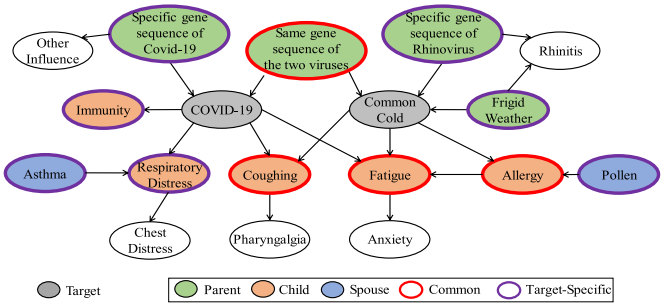


Fig. 1. Examples to illustrate the basic concepts in this paper. (1) MB of a target (e.g., Common Cold) contains *parents* (Frigid Weather, Specific gene sequence of Rhinovirus, and Same gene sequence of the two viruses), *children* (Coughing, Fatigue, and Allergy), and *spouses* (Pollen) of the target. (2) For multiple targets, *common* MB variables simultaneously influence multiple targets while *target-specific* MB variables influence a single target. For example, Coughing is the common MB variable of COVID-19 and Common Cold, while Respiratory Distress is the target-specific MB variable of COVID-19.

the model with varying degrees, it is necessary to study the multi-target MB discovery problem with the following goals:

- To discover all of the MB variables for a target set;
- To distinguish between common MB variables and target-specific MB variables.

Due to more complex relationships in multi-target scenarios, existing single-target methods cannot be applied to multi-target scenarios directly. Different from the single-target scenario, the Unique MB Assumption should be relaxed in multi-target MB discovery since the multiple MBs lead to uncertain solutions. For example, in meteorology, *the dropping in sea level pressure* (F_1), *the convergence of the winds near the surface* (F_2), and *the divergence of the winds at the top of the atmosphere* (F_3), are spatially adjacent [23], and just any one of them can be equivalently used to predict the *tornado* (T_1). And only F_1 needs to be used to predict the *extreme precipitations* (T_2). Thus, there exist three equivalent MBs of T_1 including F_1 or F_2 or F_3 respectively, making F_1 become a common MB variable of T_1 and T_2. However, as mentioned before, it is scarcely possible for existing methods to mine all MBs due to the low statistical reliability [12], making some common variables undetected. In the example above, if we search common MB variables of T_1 and T_2 through the intersection of their MBs but only find the MB of T_1 including F_2 or F_3, then F_1 can not be identified. Besides this, directly finding all MBs may suffer from high time complexity and low accuracy due to the numerous conditional independence tests (CI-tests), especially with large conditioning sets [18]. Fortunately, the non-unique MBs coexist with equivalent information[3] [12], which are easier and more efficient to detect. Furthermore, due to the target relationships in multi-target scenarios, dependence between targets should be examined.

In this article, we discover that common MB variables are determined by equivalent information following different mechanisms with or without the existence of target relationships. Explicitly, we prove that if any target does not contain the predictive information about another target, then the equivalent information about targets might induce new common MB variables, while the equivalent information about non-target variables does not. For this purpose, we introduce a Target-relation Assumption to simplify the discussion, which supposes that a target is not included by the MB of another target. Based on this assumption, the discussion is divided into two parts, satisfying and violating the assumption. We start from the simple case satisfying the assumption, and provide the general characteristics of common MB variables. Afterwards, we relax the assumption and find that some unidentified common MB variables are influenced by equivalent information about non-target variables. Based on the theoretical analyses, we subsequently develop a Common and Target-specific MB variable discovery (CTMB) algorithm to achieve the following benefits:

1) Practicability: CTMB can identify most MB variables and simultaneously distinguish the two types;
2) Robustness: CTMB is always effective in the case satisfying or violating the Target-relation Assumption and Unique MB Assumption;
3) Generality: CTMB can be directly extended to facilitate some real-world applications.

To demonstrate the generality of CTMB, we apply it to feature selection and propose a novel CTMB-driven multi-Label Feature Selection algorithm (CLFS). Through learning the MBs around multiple labels, CLFS possesses three superiorities over traditional algorithms:

1) Interpretability: CLFS can explain which labels a selected feature influences.
2) Practicability: Under the premise of ensuring the relatively higher accuracy, CLFS automatically predetermines the number of selected features via mining the underlying mechanism.
3) Theoretical Reliability: We will prove that CLFS achieves the maximum relevance and minimum redundancy in Section 4.2.

The remainder of this paper is organized as follows. We first introduce the related work in Section 2, including the basic theories and classical MB discovery algorithms. Then, the theoretical properties of common and target-specific MB variables are discussed in Section 3, with and without consideration of relationships between targets. Based on these theories, CTMB is also proposed in Section 3. We extend the CTMB to solve multi-label feature selection and propose the CLFS algorithm in Section 4, where the maximum relevance and minimum redundancy achieved by CLFS are also proved. In Section 5, we conduct extensive experiments to validate the proposed algorithms on various synthetic and real-world data sets. Finally, we conclude this paper and propose some future directions in Section 6.

2 Synopsis of Theories and Methods Motivating Present Work

In this section, we introduce some basic definitions and theories motivating the present work. Additionally, some classical and state-of-the-art methods related to this work are introduced. In this paper, the 'target' is used to denote the variable being studied when the MB discovery is discussed,

3. A Phenomenon that two variable sets contain equivalent information about a target (Refer to Definition 3 in Section 2 for details).

and the 'label' is used to replace the 'target' when the feature selection application is discussed. Common upper-case letters denote random variables and upper-case bold letters denote random variable sets. Specifically, $\boldsymbol{V} = \boldsymbol{U} \cup \boldsymbol{T}$ represents the set of all variables, in which $\boldsymbol{U}$ represents the non-target variable set (or feature set), and $\boldsymbol{T}$ represents the target set (or label set). When the discussion is about a certain target, T is used to denote the target. Hollow upper-case letter $\mathbb{G}$ denotes a directed acyclic graph (DAG), and $\mathbb{P}$ denotes the joint probability distribution over $\boldsymbol{V}$.

2.1 Basic Properties of Probability Distribution

Definition 1. *(Conditional Independence) Variable sets* $\mathbf{X}$ *and* $\mathbf{Y}$ *are conditionally independent given a variable set* $\mathbf{Z}$ *if* $\mathbb{P}(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = \mathbb{P}(\mathbf{X}|\mathbf{Z})\mathbb{P}(\mathbf{Y}|\mathbf{Z})$*, denoted as* $\mathbf{X} \perp \mathbf{Y}|\mathbf{Z}$*. Inversely,* $\mathbf{X} \not\perp \mathbf{Y}|\mathbf{Z}$ *denotes the conditional dependence relationships.*

Some important basic properties of joint probability distribution will be used to prove the theorems in this paper.

Theorem 1. *[3], [24] Let variable sets* $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{Z} \subset \mathbf{V}$*, six properties hold in any joint probability distribution* $\mathbb{P}$ *over* $\mathbf{V}$*:*

1) *Self-conditioning:* $\mathbf{A} \perp \mathbf{Z}|\mathbf{Z}$.
2) *Symmetry:* $\mathbf{A} \perp \mathbf{B}|\mathbf{Z} \Leftrightarrow \mathbf{B} \perp \mathbf{A}|\mathbf{Z}$.
3) *Decomposition:* $\mathbf{A} \perp \mathbf{B} \cup \mathbf{C}|\mathbf{Z} \Rightarrow \mathbf{A} \perp \mathbf{B}|\mathbf{Z}$ *and* $\mathbf{A} \perp \mathbf{C}|\mathbf{Z}$.
4) *Weak union:* $\mathbf{A} \perp \mathbf{B} \cup \mathbf{C}|\mathbf{Z} \Rightarrow \mathbf{A} \perp \mathbf{B}|\mathbf{Z} \cup \mathbf{C}$.
5) *Contraction:* $\mathbf{A} \perp \mathbf{B}|\mathbf{Z} \cup \mathbf{C}$ *and* $\mathbf{A} \perp \mathbf{C}|\mathbf{Z} \Rightarrow \mathbf{A} \perp \mathbf{B} \cup \mathbf{C}|\mathbf{Z}$.
6) *Intersection: If* $\mathbb{P}$ *is strictly positive, then:* $\mathbf{A} \perp \mathbf{B}|\mathbf{Z} \cup \mathbf{C}$ *and* $\mathbf{A} \perp \mathbf{C}|\mathbf{Z} \cup \mathbf{B} \Rightarrow \mathbf{A} \perp \mathbf{B} \cup \mathbf{C}|\mathbf{Z}$.

Using mutual information [25] to measure the conditional independence relationship, we have $I(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}) = 0$ if $\boldsymbol{X} \perp \boldsymbol{Y}|\boldsymbol{Z}$. There are five frequently-used CI-tests methods: λ^2 test, G^2 test, mutual information for discrete features [26], Fishers Z tests for continuous features with linear relations with additive Gaussian errors [27], and kernel-based tests for continuous features with nonlinearity and non-Gaussian noise [28]. In this paper, discussion about CI-tests is based on Assumption 1:

Assumption 1. *(Statistical Sufficiency Assumption) [1], [2] The learner has access to a sufficiently large training set and reliable statistical tests for determining conditional dependencies and independencies in the original distribution where the data are sampled from.*

Assumption 1 states that the CI-tests adopted in algorithms are correct, which requires that the data set is a sufficiently large independent and identically distributed sample of the underlying probability distribution. And Obviously, small-scale samples will influence the correctness of the statistical tests.

2.2 Markov Blanket and Markov Boundary

Definition 2. *(Markov Blanket and Markov Boundary) [3] The Markov blanket* $\mathbf{Mb}$ *of target* T *is a subset of* $\mathbf{V}$ *satisfying the condition:* $\forall X \in \mathbf{V} - \mathbf{Mb}, X \perp T|\mathbf{Mb}$ *in the joint probability distribution* $\mathbb{P}$*. Markov boundary* $\mathbf{MB}$ *of* T *is the minimum Markov blanket of* T *satisfying:* $\forall \mathbf{Z} \subset \mathbf{MB}$, $\mathbf{Z}$ *is not a Markov blanket of* T.

In this paper, the Markov boundary is abbreviated as MB. According to Definition 2, the mutual information $I(T, \boldsymbol{V} - \{T\}) = I(T, \boldsymbol{MB})$, and thus, MB set carries all of the predictive information about the corresponding target. MB provides a complete picture of the local relationship around the target [18], which could be intuitively understood in BN [3]. An example of MB with DAG is shown in Fig. 1. In a faithful BN, MB of a variable includes its parents, children, and spouses [3]. Thus, MB of Common Cold contains its parents (Frigid Weather, Specific gene sequence of Rhinovirus, Same gene sequence of the two viruses), children (Coughing, Fatigue, Allergy), and spouses (Pollen). The remaining variables are independent of Common Cold conditioned on its MB.

Extensive works further assume that the target has a unique MB, and propose many effective methods, which can be broadly classified into two types according to the review [29], i.e., simultaneous MB learning algorithms and divide-and-conquer MB learning algorithms. Some early proposed methods, such as Incremental Association MB (IAMB) [13] and its variants [13], [30], are simultaneous MB learning algorithms. These algorithms do not distinguish between the parent-child variables and spouse variables and learn them simultaneously. Thus, they are time-efficient but require the number of samples to be exponential to the size of the MB, which means that insufficient samples will result in performance degradation [16]. Divide-and-conquer MB learning algorithms are proposed to further improve the MB discovery accuracy with a reasonable time cost, which first search the parent-child variables and then the spouse variables of a target. Classical methods include Max-Min MB (MMMB) [14], HITON-MB [15], and Parents-and-Children-based MB (PCMB) [16], and the state-of-the-art ToLerant MB (TLMB) [19], Separation and Recovery MB (SRMB) [17], and Cross-check and Complement MB (CCMB) [18]. Most of these algorithms are efficient to seek an approximate MB set with a reasonable time cost. The aforementioned algorithms assume that the probability distribution is strictly positive:

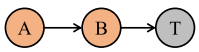
Theorem 2. *[31] If the joint probability distribution* $\mathbb{P}$ *satisfies the Intersection property, then a target has a unique MB.*

Under certain assumptions, these algorithms have good performances and also have been widely applied in feature selection. However, in real-world applications, the unique MB assumption is always violated, leading to multiple equivalent MBs for a target. For example, if a variable is completely determined by another, then the Intersection property is violated when taking one of the two as the target. Some relevant theories and algorithms about multiple MBs are reviewed next.

2.3 Multiple MBs and Equivalent Information

When the joint probability distribution $\mathbb{P}$ does not satisfy the Intersection property, there exists a phenomenon, called equivalent information.

Definition 3. *(Equivalent information) [12] Variable subsets* $\mathbf{X}$ *and* $\mathbf{Y}$ *contain equivalent information about target variable* T *conditioned on* $\mathbf{Z}$ *if and only if* $T \not\perp \mathbf{X}|\mathbf{Z}$, $T \not\perp \mathbf{Y}|\mathbf{Z}$, $T \perp \mathbf{X}|\mathbf{Y} \cup \mathbf{Z}$, $T \perp \mathbf{Y}|\mathbf{X} \cup \mathbf{Z}$.



$P(A)$	
$A=0$	0.6
$A=1$	0.4

$P(B\|A)$	$A=0$	$A=1$
$B=0$	0.0	1.0
$B=1$	1.0	0.0

$P(T\|B)$	$B=0$	$B=1$
$T=0$	0.9	0.1
$T=1$	0.1	0.9

(A,B,T)	(0, 0, 0)	(0, 0, 1)	(0, 1, 0)	(0, 1, 1)
$P(A,B,T)$	0.0	0.0	0.06	0.54
(A,B,T)	(1, 0, 0)	(1, 0, 1)	(1, 1, 0)	(1, 1, 1)
$P(A,B,T)$	0.36	0.04	0.0	0.0

Fig. 2. A simple example of Equivalent information. The response variable is T, and all variables take values $\{0,1\}$. Variables A and B, highlighted with the same color, contain equivalent information about T.

Fig. 2 provides an example of equivalent information. According to the probability distribution in the probability table, $\mathbb{P}(A,T|B) = \mathbb{P}(A|B)\mathbb{P}(T|B)$ and $\mathbb{P}(B,T|A) = \mathbb{P}(B|A)$ $P(T|A)$. Then, we can conclude that $A \not\perp T$, $B \not\perp T$, $A \perp T|B$, and $B \perp T|A$. According to Definition 3, A and B contain equivalent information about T. Moreover, it can be seen from the probability distribution in Fig. 2 that, the MB set of T could be $\{A\}$ or $\{B\}$, which verifies the coexistence of the equivalent information phenomenon and multiple MBs. The phenomenon is formally stated as Theorem 3.

Theorem 3. *[12] The intersection property holds if no information equivalence occurs.*

According to Theorem 3, if there exists equivalent information, a target might have multiple MBs. Some algorithms are proposed to detect multiple MBs. Earlier solutions are stochastic algorithms, which find multiple MBs through running a single MB discovery algorithm multiple times initialized with a random seed. For example, KIAMB [16] is a stochastic extension of IAMB, which tries to get all MBs by running IAMB K times. Ensemble gene selection by grouping [32], another multiple MB discovery strategy, groups variables into multiple clusters first and then randomly samples a representative from each cluster, constituting different MBs. The aforementioned approaches are highly heuristic and can not guarantee the correctness of the output. Iterative Removal HITON-PC (IR-HITON-PC) [33] applies Semi-Interleaved HITON-PC to identify an MB set, and then removes the variables in the discovered MB from the variable set and repeatedly invokes the MB discovery process. IR-HITON-PC is theoretically correct but not practical to induce multiple MBs in high-dimensional but relatively small-scale data sets [12]. Target Information Equivalence algorithm (TIE*) [12] is an algorithmic framework for multiple MB discovery. It uses a single MB discovery algorithm to find an initial MB first, and then repeats the following three steps: (1) Remove a subset $\boldsymbol{G}$ of a previously discovered MB $\boldsymbol{MB}_{pre}$ from the full variable set $\boldsymbol{V}$, and obtain $\boldsymbol{V} - \boldsymbol{G}$; (2) Learn a new MB $\boldsymbol{MB}_{new}$ from the remaining variable set $\boldsymbol{V} - \boldsymbol{G}$; (3) Estimate the correctness of $\boldsymbol{MB}_{new}$. They are repeated until all possible MB subsets $\boldsymbol{G}$ are considered. TIE* dictates to consider removing from $\boldsymbol{V}$ only certain subsets $\boldsymbol{G}$ of the previously found MB, and thus is more efficient than other methods. MB discovery under the Weak Markov Local Composition assumption (WLCMB) [20] uses a similar strategy to search multiple MBs. It improves step (2) in TIE* with a novel single MB discovery algorithm LCMB to avoid incorrect CI-tests. Although the efficiency and accuracy of multiple MB discovery algorithms continue to improve, they are still intractable to find all possible MBs since the unpredictable number of MBs makes the process time-consuming.

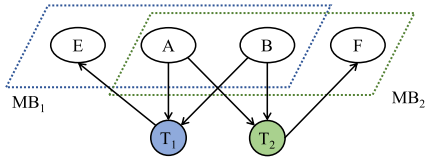


Fig. 3. An example of common MB variables and target-specific MB variables without multiple MBs.

3 MB Discovery for a Variable Set

Compared with the single-target problem, the additional information introduced to the multi-target problem is the relationship between targets, which is crucial for the analysis in multi-target scenarios [34]. Due to the target relationships, MB variables in multi-target data include two types, i.e., aforementioned common MB variables and target-specific MB variables. In the following, we formally define these variables in Section 3.1, and discuss their properties with and without consideration of relationships between targets in Sections 3.2 and 3.3. Based on the theoretical property, a discovery and distinguishing algorithm is proposed in Section 3.4 with a toy example in Section 3.5.

3.1 Definition and Assumption: Common & Target-Specific MB Variables, and Target-Relation Assumption

As a concept derived from MB, we formally give straightforward definitions of common MB variables and target-specific MB variables based on the concept of MB.

Definition 4. *For a target set* $\mathbf{T}$*, variable* X *is a common MB variable of target set* $\mathbf{T}$ *if* $\forall T \in \mathbf{T}$*, there exists an MB of* T *including* X*. Variable* X *is a target-specific MB variable if there is only one* $T \in \mathbf{T}$ *whose MB includes* X*.*

The common and target-specific MB variables are defined for a target set and a single target, respectively. According to Definition 4, they can be easily distinguished as long as the MB set of each target is previously known. For example in Fig. 3, if the intersection property holds and each target has a unique MB, then the intersection ($\{A,B\}$) of the MB sets of T_1 and T_2 is the common MB variables for $\{T_1,T_2\}$. E is a target-specific MB variable of T_1, and F is a target-specific MB variable of T_2. Fig. 4 further presents a case with multiple MBs, where $\{A,B\}$ and $\{C,D\}$ contain equivalent information about T_1, and $\{A,B\}$ and $\{G,H\}$ contain equivalent information about T_2. In this case, $\{A,B,E\}$ and $\{C,D,E\}$ are MBs of T_1, $\{A,B,F\}$ and $\{G,H,F\}$ are MBs of T_2[4]. Then, we obtain the common MB variables of $\{T_1,T_2\}$ are A and B, and others are target-specific. The second example indicates that learning all of the

4. Note that the case in this assumption is possible under the DAG in Fig. 4, which can be understood from the probability table in Fig. 2

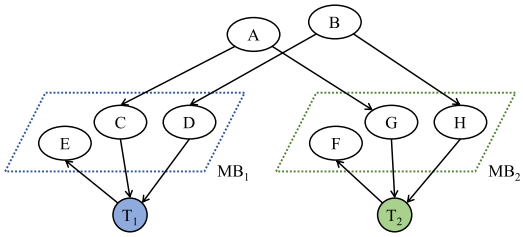


Fig. 4. An example of common MB variables and target-specific MB variables with multiple MBs.

multiple MB sets is the premise of distinguishing these two types of variables. Hence, Definition 4 is still difficult to use as a criterion for identification. Given the constant predictive information, we now present another definition of common MB variables from the perspective of information theory.

Definition 5. *Let* $\mathbf{T}$ *denote a target set,* $\mathbf{MB}_i$ *denote the MB of* $T_i \in \mathbf{T}$*, and* $\mathbf{Z}_i \subset \mathbf{MB}_i$ *denote a nonempty subset of* $\mathbf{MB}_i$*. If there exists* $\mathbf{Z} \subset \mathbf{U}$*, such that*

$$I(\mathbf{MB}_i - \mathbf{Z}_i \cup \mathbf{Z}, T_i) = I(\mathbf{MB}_i, T_i), \tag{1}$$

and any subsets of $\mathbf{Z}$ *do not satisfy Eq. (1), then all variables in* $\mathbf{Z}$ *are common MB variables of the target set* $\mathbf{T}$*.*

Intuitively, Eq. (1) in Definition 5 means that the common MB variables in $\boldsymbol{Z}$ can be used to replace the MB subset of each target without any information loss. Thus, $\boldsymbol{Z}$ carries the information of all targets in $\boldsymbol{T}$, while $\boldsymbol{Z}_i$ only carries the information of the target T_i. Moreover, $\boldsymbol{MB}_i - \boldsymbol{Z}_i \cup \boldsymbol{Z}$ is an MB of T_i, and $\boldsymbol{Z}$ is also an MB subset of T_i. Although $\boldsymbol{Z}$ and $\boldsymbol{Z}_i$ are both subsets of a certain MB of T_i, $\boldsymbol{Z}_i$ does not have to be different from $\boldsymbol{Z}$, which depends on whether they are from the same MB. For example, in Fig. 3, $\boldsymbol{MB}_1 = \{A, B, E\}$, $\boldsymbol{MB}_2 = \{A, B, F\}$, then $\boldsymbol{Z}_1 = \boldsymbol{Z}_2 = \boldsymbol{Z} = \{A\}$, or $\boldsymbol{Z}_1 = \boldsymbol{Z}_2 = \boldsymbol{Z} = \{B\}$. While in Fig. 4, assume that the MBs of T_1 and T_2 have been known as $\boldsymbol{MB}_1 = \{C, D, E\}$ and $\boldsymbol{MB}_2 = \{F, G, H\}$. Then, taking $\boldsymbol{Z}_1 = \{C, D\}$, $\boldsymbol{Z}_2 = \{G, H\}$, we can obtain that $\boldsymbol{Z} = \{A, B\}$ is the solution of Eq. (1), which indicates that A and B are common MB variables of $\{T_1, T_2\}$. Additionally, some supersets of $\boldsymbol{Z}$ (e.g., $\{A, B, E, H\}$ in Fig. 4) also meet the requirement in Eq. (1), thus Definition 5 further restrains that any subset of $\boldsymbol{Z}$ does not satisfy Eq. (1).

The main difference between Definitions 4 and 5 is that, Definition 4 introduces the literal meaning of the common MB variable, i.e., $\forall T \in \boldsymbol{T}$, there exists an MB set of T including X. According to Definition 4, we need to obtain all MBs of each target (including multiple MBs) to find common MB variables. Inversely, Definition 5 only requires one MB previously known of each target, and Eq. (1) retrieves other MBs by replacing $\boldsymbol{Z}_i$ with $\boldsymbol{Z}$. In the following, we theoretically prove the equivalence of these two definitions.

Theorem 4. *Definitions 4 and 5 are equivalent.*

Proof. Definition 4 $\Rightarrow$ Definition 5: Assume X is a variable satisfying Definition 4. According to Definition 4, for any target T_i, there exists at least one MB containing X, denoted as $\boldsymbol{MB}_X$. Denote the known MB of T_i as $\boldsymbol{MB}_i$. If $\boldsymbol{MB}_X = \boldsymbol{MB}_i$, i.e., $\boldsymbol{MB}_X$ happens to be the known MB in Definition 5, then $\boldsymbol{Z} = \{X\}$, and $\boldsymbol{Z}_i = \boldsymbol{Z}$ $(\forall T_i)$. We have $I(\boldsymbol{MB}_i - \boldsymbol{Z}_i \cup \boldsymbol{Z}, T_i) = I(\boldsymbol{MB}_i, T_i)$ and affirm that $X \in \boldsymbol{Z}$. On the other hand, if $\boldsymbol{MB}_X \neq \boldsymbol{MB}_i$, then $\boldsymbol{Z} = \cup_i(\boldsymbol{MB}_X - \boldsymbol{MB}_i)$ and $\boldsymbol{Z}_i = \boldsymbol{MB}_i - \boldsymbol{MB}_X$. Hence, we obtain $I(\boldsymbol{MB}_i - \boldsymbol{Z}_i \cup \boldsymbol{Z}, T_i) = I(\boldsymbol{MB}_i, T_i)$ and can affirm that $X \in \boldsymbol{Z}$.

Definition 4 $\Leftarrow$ Definition 5: Assume $\boldsymbol{Z}$ is a subset satisfying Definition 5. Suppose $\exists X \in \boldsymbol{Z}$ is not a common MB variable, i.e., $\exists T_i$ such that $X \notin \boldsymbol{MB}_i$ $(\forall \boldsymbol{MB}_i)$. Then, by the chain rule of mutual information [25]:

$$\begin{aligned} &I(\boldsymbol{MB}_i - \boldsymbol{Z}_i \cup \boldsymbol{Z} - \{X\}, T_i) \\ &= I(\boldsymbol{MB}_i - \boldsymbol{Z}_i \cup \boldsymbol{Z}, T_i) \quad - I(X, T_i | \boldsymbol{MB}_i - \boldsymbol{Z}_i \cup \boldsymbol{Z} - \{X\}). \end{aligned} \tag{2}$$

According to Eq. (1), $\boldsymbol{MB}_i - \boldsymbol{Z}_i \cup \boldsymbol{Z}$ is a Markov blanket of T_i. Since $\forall \boldsymbol{MB}_i$, $X \notin \boldsymbol{MB}_i$, $\boldsymbol{MB}_i - \boldsymbol{Z}_i \cup \boldsymbol{Z} - \{X\}$ is a Markov blanket of T_i. Thus,

$$I(X, T_i | \boldsymbol{MB}_i - \boldsymbol{Z}_i \cup \boldsymbol{Z} - \{X\}) = 0 \tag{3}$$

Substituting Eqs. (3) into (2) and we obtain $I(\boldsymbol{MB}_i - \boldsymbol{Z}_i \cup \boldsymbol{Z} - \{X\}, T_i) = I(\boldsymbol{MB}_i, T_i)$. Thus, $\boldsymbol{Z} - \{X\} \subset \boldsymbol{Z}$ also satisfies Definition 5, contradicting the condition. Therefore, all variables in $\boldsymbol{Z}$ satisfy Definition 4. (Q.E.D.) □

Theorem 4 demonstrates the equivalence of Definitions 4 and 5. Hence, the proof of "Definition 4 $\Rightarrow$ Definition 5" should find a sound $\boldsymbol{Z}$ that contains a given common MB variable X satisfying the Definition 4. Correspondingly, the proof of "Definition 4 $\Leftarrow$ Definition 5" should prove that all variables in the $\boldsymbol{Z}$ satisfying the requirements in Definition 5 are included by the MB of any target. By Definition 5, the problem discussed in this paper can be described as: for non-target variable set $\boldsymbol{U}$ and target set $\boldsymbol{T}$, we need to search two types of MB variables from $\boldsymbol{U}$, i.e., (1) the common MB variables of $\boldsymbol{T}$, and (2) the target-specific MB variables of each single target $T_i \in \boldsymbol{T}$. And the designed algorithms should search the two types of MB variables for both $\boldsymbol{T}$ and subsets of $\boldsymbol{T}$. Different from the single-target problem, a special issue must be discussed in the multi-target case, i.e., the possible correlations in the target set. To simplify the problem, we first propose the Target-relation Assumption.

Assumption 2. *(Target-relation Assumption)* $\forall T_1, T_2 \in \mathbf{T}$, $T_1 \notin \mathbf{MB}(T_2)$ *and* $T_2 \notin \mathbf{MB}(T_1)$.

Target-relation Assumption considers the relationships among targets, and divides the discussion according to whether a target contains the critical information about another target. Note that this assumption allows the indirect dependence between targets, for which an eligible example is that a target influences another target through a non-target variable. In Sections 3.2 and 3.3, the property of the common MB variable is discussed in the cases where Target-relation Assumption is satisfied and violated, respectively.

3.2 Discussion Under the Target-Relation Assumption

According to Definition 5, the intersection of MB sets of targets in $\boldsymbol{T}$ is a common MB variable set of $\boldsymbol{T}$ since $\boldsymbol{Z} = \boldsymbol{Z}_i = \bigcap_{i=1}^{k} \boldsymbol{MB}_i$ is a constant solution of Eq. (1). For example in Fig. 3, $\boldsymbol{Z} = \{A, B\}$. Therefore, if all targets have a unique MB, then the intersection of these MBs is the intact common

MB variable set. However, the real-world applications always violate the unique MB assumption, and most targets have multiple MBs.

Directly finding the multiple MBs is time-consuming since the time complexity is exponential to the size of the variable set. It will also suffer from incorrect CI-tests due to the large conditioning sets in the process. Detecting whether the Intersection property is violated is a possible method according to Theorem 2, whereas it is infeasible to identify the strictly positive joint probability distribution. Another criterion for unique MB, by Theorem 3, is to detect the equivalent information, as mainly discussed in this section. As the multi-target problem has complex relationships, the equivalent information phenomenon has diversified forms. It can be roughly classified into two types, i.e., equivalent information about target variables or non-target variables. To narrow the discussion, we first prove that only the equivalent information about targets has influences on the identification of common MB variables.

Theorem 5. *For a target T, if for any disjoint variable subsets $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subset \mathbf{V} - \{T\}$, $\mathbf{X}$ and $\mathbf{Y}$ do not contain equivalent information about T conditioned on $\mathbf{Z}$, then T has a unique MB.*

Proof. Assuming that T has two MB sets $\boldsymbol{MB}_1$ and $\boldsymbol{MB}_2$, then we need to find two variable sets containing equivalent information about T. According to Definition 2, we have:

$$T \perp \boldsymbol{V} - \boldsymbol{MB}_1 - \{T\} | \boldsymbol{MB}_1, T \perp \boldsymbol{V} - \boldsymbol{MB}_2 - \{T\} | \boldsymbol{MB}_2. \tag{4}$$

According to the Decomposition property in Theorem 1, Eq. (4) indicates that:

$$T \perp \boldsymbol{MB}_2 - \boldsymbol{MB}_1 | \boldsymbol{MB}_1, T \perp \boldsymbol{MB}_1 - \boldsymbol{MB}_2 | \boldsymbol{MB}_2. \tag{5}$$

Now we prove that $\boldsymbol{MB}_1 - \boldsymbol{MB}_2$ and $\boldsymbol{MB}_2 - \boldsymbol{MB}_1$ contain equivalent information about T conditioned on $\boldsymbol{MB}_1 \cap \boldsymbol{MB}_2$. Assume that $T \perp (\boldsymbol{MB}_1 - \boldsymbol{MB}_2) | \boldsymbol{MB}_1 \cap \boldsymbol{MB}_2$. Considering with Eq. (4), we obtain the following relationship according to the Contraction property in Theorem 1:

$$T \perp (\boldsymbol{V} - \boldsymbol{MB}_1 - \{T\}) \cup (\boldsymbol{MB}_1 - \boldsymbol{MB}_2) | \boldsymbol{MB}_1 \cap \boldsymbol{MB}_2. \tag{6}$$

Simplify the Eq. (6), then:

$$T \perp (\boldsymbol{V} - \{T\} - \boldsymbol{MB}_1 \cap \boldsymbol{MB}_2) | \boldsymbol{MB}_1 \cap \boldsymbol{MB}_2. \tag{7}$$

We can conclude from Eq. (7) that $\boldsymbol{MB}_1 \cap \boldsymbol{MB}_2$ is an Mb of T according to Definition 2. However, $\boldsymbol{MB}_1 \cap \boldsymbol{MB}_2 \subset \boldsymbol{MB}_1$ and $\boldsymbol{MB}_1 \cap \boldsymbol{MB}_2 \subset \boldsymbol{MB}_2$, which leads to $\boldsymbol{MB}_1$ and $\boldsymbol{MB}_2$ are Mb instead of MB, contradicting the condition. Therefore,

$$T \not\perp (\boldsymbol{MB}_1 - \boldsymbol{MB}_2) | \boldsymbol{MB}_1 \cap \boldsymbol{MB}_2. \tag{8}$$

Similarly, we can prove that

$$T \not\perp (\boldsymbol{MB}_2 - \boldsymbol{MB}_1) | \boldsymbol{MB}_1 \cap \boldsymbol{MB}_2. \tag{9}$$

Combining Eqs. (8) and (9) with Eq. (5), we can conclude that, $\boldsymbol{MB}_1 - \boldsymbol{MB}_2$ and $\boldsymbol{MB}_2 - \boldsymbol{MB}_1$ contain equivalent information about T conditioned on $\boldsymbol{MB}_1 \cap \boldsymbol{MB}_2$, contradicting the condition. Hence, T has a unique MB. (Q.E.D.)□

Theorem 5 proves that multiple MBs of a target are brought by the equivalent information about the corresponding target, while equivalent information about non-target variables does not influence the uniqueness of MB, as well as common MB variables. Therefore, only equivalent information about each target needs to be considered for common MB variable identification. Theorem 6 is proposed below to describe this criterion.

Theorem 6. *Let $\mathbf{MB}_i$ denote the MB set of T_i ($i \in \{1, 2, \ldots, k\}$) in target set $\mathbf{T} = \{T_1, T_2, \ldots, T_k\}$. Under the Target-relation Assumption, $\mathbf{Z} \subset \mathbf{U}$ is a common MB variable set of targets in $\mathbf{T}$ if and only if $\exists \mathbf{Z}_i \subset \mathbf{MB}_i$ such that $\mathbf{Z}_i$ and $\mathbf{Z}$ contain equivalent information about T_i conditioned on $\mathbf{MB}_i - \mathbf{Z}_i$ for each $T_i \in \mathbf{T}$.*

Proof. "⇒": Since $\boldsymbol{Z}$ is a common MB variable set of labels in $\boldsymbol{T}$, then for each $T_i \in \boldsymbol{T}$, there exists an MB set $\boldsymbol{MB}_i$ and corresponding MB subset $\boldsymbol{Z}_i \subset \boldsymbol{MB}_i$ s.t.

$$I(\boldsymbol{MB}_i - \boldsymbol{Z}_i \cup \boldsymbol{Z}, T_i) = I(\boldsymbol{MB}_i, T_i), \tag{10}$$

where $\boldsymbol{MB}_i - \boldsymbol{Z}_i \cup \boldsymbol{Z}$ is also an MB of T_i. According to Definition 2, we have $\boldsymbol{Z} \perp T_i | \boldsymbol{MB}_i$ and $\boldsymbol{Z}_i \perp T_i | \boldsymbol{MB}_i - \boldsymbol{Z}_i \cup \boldsymbol{Z}$. According to the minimality of the MB, we have $\boldsymbol{Z}_i \not\perp T_i | \boldsymbol{MB}_i - \boldsymbol{Z}_i$ (corresponding MB: $\boldsymbol{MB}_i$) and $\boldsymbol{Z} \not\perp T_i | \boldsymbol{MB}_i - \boldsymbol{Z}_i$ (corresponding MB: $\boldsymbol{MB}_i - \boldsymbol{Z}_i \cup \boldsymbol{Z}$). According to the four conditional (in)dependence relations, we can conclude that $\exists \boldsymbol{Z}_i \subset \boldsymbol{MB}_i$ such that $\boldsymbol{Z}_i$ and $\boldsymbol{Z}$ contain equivalent information about T_i conditioned on $\boldsymbol{MB}_i - \boldsymbol{Z}_i$ for each $T_i \in \boldsymbol{T}$ according to Definition 3.

"⇐": To prove that $\boldsymbol{Z}$ is a common MB variable set for $\boldsymbol{T}$, we need to prove that $\boldsymbol{Z}$ satisfies Eq. (1) in Definition 5. By the chain rule of mutual information, we express $\boldsymbol{MB}_i \cup \boldsymbol{Z}$ as $(\boldsymbol{MB}_i \cup \boldsymbol{Z} - \boldsymbol{Z}_i) \cup \boldsymbol{Z}_i$ and obtain:

$$\begin{aligned} I(\boldsymbol{MB}_i \cup \boldsymbol{Z} - \boldsymbol{Z}_i, T_i) = {} & I(\boldsymbol{MB}_i \cup \boldsymbol{Z}, T_i) \\ & - I(\boldsymbol{Z}_i, T_i | \boldsymbol{MB}_i \cup \boldsymbol{Z} - \boldsymbol{Z}_i). \end{aligned} \tag{11}$$

Since $\boldsymbol{Z}$ and $\boldsymbol{Z}_i$ contain equivalent information about T_i, according to Definition 3, we have:

$$I(\boldsymbol{Z}_i, T_i | \boldsymbol{MB}_i \cup \boldsymbol{Z} - \boldsymbol{Z}_i) = 0. \tag{12}$$

Since $\boldsymbol{Z} \perp T_i | \boldsymbol{MB}_i$, we have:

$$I(\boldsymbol{MB}_i \cup \boldsymbol{Z}, T_i) = I(\boldsymbol{MB}_i, T_i). \tag{13}$$

Substitute Eqs. (12) and (13) into Eq. (11), thus,

$$I(\boldsymbol{MB}_i \cup \boldsymbol{Z} - \boldsymbol{Z}_i, T_i) = I(\boldsymbol{MB}_i, T_i). \tag{14}$$

According to Theorem 5, all common MB variables are considered in the theorem since the Target-relation Assumption is satisfied and thus any target is not an MB variable of another.

In conclusion, Theorem 6 is true. (Q.E.D.) □

Theorem 6 proves that equivalent information between the MB subset and another variable set can be used to detect common MB variables. For example in Fig. 4, $\{A, B\}$ and $\{C, D\}$ contain equivalent information about T_1 (conditioned on E), $\{A, B\}$ and $\{G, H\}$ contain equivalent information about T_2 (conditioned on E). Assume it has been known that $\{C, D, E\}$ is an MB set of T_1, and $\{F, G, H\}$ is an MB set of T_2. According to Theorem 6, $\{A, B\}$ can be detected as common MB variables of $\{T_1, T_2\}$ without mining other MB sets of T_1 and T_2. The conclusion can be obtained by following the proof of Theorem 6. Since $\{C, D, E\}$ is the MB of T_1, $I(T_1, \{C, D, E\}) = I(T_1, \{C, D, E\} \cup \{A, B\})$ according to property of MB. As $\{C, D\}$ and $\{A, B\}$ contain equivalent information about T_1 (conditioned on E), we immediately obtain $I(T_1, \{C, D\}|\{A, B, E\}) = 0$ due to the independence relation $T_1 \perp \{C, D\}|\{A, B, E\}$. And according to the chain rule of mutual information, $I(T_1, \{A, B, C, D, E\}) = I(T_1, \{A, B, E\}) + I(T_1, \{C, D\}|\{A, B, E\})$. Therefore, $I(T_1, \{A, B, E\}) = I(T_1, \{A, B, C, D, E\}) = I(T_1, \{C, D, E\})$ and $\{A, B, E\}$ is an MB of T_1.

3.3 Relax the Target-Relation Assumption

When the Target-relation Assumption is relaxed, there might exist more common MB variables undetected. Different from the case satisfying the assumption, the local relations around a target are represented with non-target variables as well as targets. Therefore, it is improper to make a difference between non-target variables and targets when mining the MBs of each target. Furthermore, the equivalent information about both targets and non-target variables needs to be considered so that some common MB variables are not ignored. Theorem 7 is proposed below to describe the case where common MB variables cannot be detected.

Theorem 7. *For targets $T_1, T_2 \in \mathbf{T}$, $T_1 \in \mathbf{MB}_2$ and $T_2 \in \mathbf{MB}_1$, variable subset $\mathbf{Z}$ is a common MB variable set of T_1 and T_2 but might not be detected if the following statements hold: (1) $\mathbf{Z} \subset \mathbf{MB}_2$. T_2 and $\mathbf{Z}$ contain equivalent information about T_1 conditioned on $\mathbf{MB}_1 - \{T_2\}$. (2) $\mathbf{Z} \subset \mathbf{MB}_1$ and $\mathbf{Z} \subset \mathbf{MB}_2$. T_1 and T_2 contain equivalent information about $\mathbf{Z}$.*

Proof. For (1): Since $\boldsymbol{Z} \subset \boldsymbol{MB}_2$, $\boldsymbol{Z}$ satisfies Eq. (1) in Definition 5 for T_2. We prove that $\boldsymbol{Z}$ satisfies Eq. (1) for T_1. According to the chain rule of mutual information, we have:

$$\begin{aligned} &I(T_1, (\boldsymbol{MB}_1 - \{T_2\}) \cup \boldsymbol{Z}) \\ &= I(T_1, \boldsymbol{Z}|\boldsymbol{MB}_1 - \{T_2\}) + I(T_1, \boldsymbol{MB}_1 - \{T_2\}), \end{aligned} \tag{15}$$

Also, $\boldsymbol{MB}_1$ can be split into $\boldsymbol{MB}_1 - \{T_2\}$ and $\{T_2\}$:

$$I(T_1, \boldsymbol{MB}_1) = I(T_1, T_2|\boldsymbol{MB}_1 - \{T_2\}) + I(T_1, \boldsymbol{MB}_1 - \{T_2\}). \tag{16}$$

Since T_2 and $\boldsymbol{Z}$ contain equivalent information about T_1, thus:

$$I(T_1, \boldsymbol{Z}|\boldsymbol{MB}_1 - \{T_2\}) = I(T_1, T_2|\boldsymbol{MB}_1 - \{T_2\}). \tag{17}$$

Then, substituting Eq. (17) into Eqs. (15) and (16), we obtain:

$$I(T_1, (\boldsymbol{MB}_1 - \{T_2\}) \cup \boldsymbol{Z}) = I(T_1, \boldsymbol{MB}_1). \tag{18}$$

Thus, $\boldsymbol{Z}$ is a common MB variable set of T_1 and T_2. Since $T_1 \perp \boldsymbol{Z}|\boldsymbol{MB}_1$ and $T_2 \in \boldsymbol{MB}_1$, if T_2 is selected by the MB discovery algorithm first, then $\boldsymbol{Z}$ will be excluded in the MB set according to the Decomposition property in Theorem 1.

For (2): It is readily justified that the variables in $\boldsymbol{Z}$ are common MB variables of T_1 and T_2 according to Definition 4. Since T_1 and T_2 contain equivalent information about $\boldsymbol{Z}$, then $T_1 \perp \boldsymbol{Z}|T_2$ and $T_2 \perp \boldsymbol{Z}|T_1$. If T_1 is selected by the MB discovery algorithm before $\boldsymbol{Z}$ when selecting MB of T_2 and T_2 is selected before $\boldsymbol{Z}$ when selecting MB of T_1, then $\boldsymbol{Z}$ can not be found. (Q.E.D.) □

We use an example to illustrate Theorem 7. If a target T_1 and the common MB variable A contain equivalent information about another target T_2, then A might be ignored since $A \perp T_2|T_1$, which describes the case in Theorem 7 (1). While the same risk does not exist under the case that two variables contain equivalent information about a target since these variables are found when detecting the equivalent information according to Theorem 6. By Theorem 7 (1), it is necessary to treat all of the targets and non-target variables as ordinary variables so that some common MB variables are not ignored due to the influence of targets. For Theorem 7 (2), also using the above example, when the two targets T_1 and T_2 contain equivalent information about common MB variable A, A might be discarded since it might be excluded when searching MB sets both of T_1 and T_2. To solve this problem, we can remove the target variable from the discovered MB first and continue to search the undetected variables.

3.4 Learn the Common & Target-Specific MB Variables

Based on the property of common MB variables, we propose the Common and Target-specific MB variable discovery (CTMB) algorithm. For the sake of preciseness, the above analyses provide the corresponding conditioning set where the equivalent information exists. While in the design of the algorithm, considering the complex conditioning set will introduce some time-consuming and unreliable processes. For comprehensive consideration of effectiveness and efficiency, we adopt a simplified strategy presented in [12], i.e., assume that all information equivalence relations are context-independent and there is no need to consider the conditioning sets [12]. CTMB consists of three phases:

Phase 1: Mine one MB for each target. Though each target could have multiple MBs, Phase 1 only needs to find one of them. According to Theorem 7 (1), CTMB equally treats targets and non-target variables and only focuses on the relations among them. A divide-and-conquer-based MB discovery algorithm $\mathbb{A}$ is used so that CTMB can distinguish the parent-child set $\boldsymbol{PC}_T$ and spouse set $\boldsymbol{SP}_T$) of T. Here, the child of the corresponding spouse in $\boldsymbol{SP}_T$ also needs to be recorded, which will be used in Phase 3.

Phase 2: To guarantee the accuracy when Target-relation Assumption is violated, Phase 2 retrieves the ignored variables whose information is equivalently included by two targets, which is the case described in Theorem 7 (2). For each

pair of targets where one is included by the MB of another (Line 6), Line 7 finds the $\boldsymbol{Z}$ satisfying the condition in Theorem 7 (2), which is retrieved in Lines 8-9.

Algorithm 1. The *CTMB* Algorithm

1: **Input:** Target set $\boldsymbol{T}$ and non-target variable set $\boldsymbol{U}$; A divide-and-conquer-based MB discovery algorithm $\mathbb{A}$ with significance level α.
{***Phase 1: Search an MB for each target.***}
2: **for** each $T \in \boldsymbol{T}$ **do**
3: $\boldsymbol{PC}_T, \boldsymbol{SP}_T, \boldsymbol{C}_T \leftarrow$ Learn the MB of T from $\boldsymbol{T} \cup \boldsymbol{U} - \{T\}$ using $\mathbb{A}$, and record the parents and children to $\boldsymbol{PC}_T$, and spouses to $\boldsymbol{SP}_T$ with corresponding child in $\boldsymbol{C}_T$.
{***Phase 2: Retrieve the ignored variables.***}
4: **for** each $T_i, T_j \in \boldsymbol{T}$ **do**
5: **if** $T_i \in \boldsymbol{PC}_j$ **do**
6: **for** each $\boldsymbol{Z}$ satisfying $\boldsymbol{Z} \not\perp T_i$, $\boldsymbol{Z} \not\perp T_j$, $\boldsymbol{Z} \perp T_i | T_j$ and $\boldsymbol{Z} \perp T_j | T_i$ **do**
7: $\boldsymbol{PC}_{T_j} \leftarrow \boldsymbol{PC}_{T_j} \cup \boldsymbol{Z}$ if $\forall \boldsymbol{S} \subset \boldsymbol{PC}_j - \{T_i\}$, $\boldsymbol{Z} \not\perp T_j | \boldsymbol{S}$.
{***Phase 3: Distinguishing process.***}
8: **for** $X \in \boldsymbol{T} \cup_{T \in \boldsymbol{T}} \boldsymbol{C}_T$ **do**
9: **for** each $\boldsymbol{Z} \subset \boldsymbol{U} - \boldsymbol{PC}_X$ and $\boldsymbol{Z} \not\perp X$ **do**
10: **if** $\exists \boldsymbol{S} \subset \boldsymbol{PC}_X$ *s.t.* $X \perp \boldsymbol{Z} | \boldsymbol{S}$ and $X \perp \boldsymbol{S} | \boldsymbol{Z}$ **then**
11: $\boldsymbol{EI}_X = \boldsymbol{EI}_X \cup \{ < \boldsymbol{S}, \boldsymbol{Z} > \}$
12: Common MB variables for any target (sub)sets $\boldsymbol{T_S} \subset \boldsymbol{T}$: $\boldsymbol{CCV_{T_S}} \leftarrow \{X | X \in \boldsymbol{Z}$ where $\Theta_{\boldsymbol{T_S}}(\boldsymbol{Z}) = 1\}$, and target -specific MB variables for each target $T \in \boldsymbol{T}$: $\boldsymbol{TCV}_T = \{X | X \in \boldsymbol{MB}_T$ and $X \notin \boldsymbol{CCV_{T_S}}$ for $\forall \boldsymbol{T_S}$ including $T\}$.

Phase 3: Find the variables containing equivalent information first and then discover the common and target-specific MB variables. Since CI-tests with large-scale variable sets will be involved if we directly find the equivalent subsets from MB of each target as described in Theorem 6, CTMB searches the common MB variables from parent-child set and spouse set, respectively. Thus, both of the equivalent MB variables of targets in $\boldsymbol{T}$ and variables in $\boldsymbol{C}$ are recorded to the $\boldsymbol{EI}$ of each corresponding variable (Lines 12-18) to make preparations for the discovery of common MB variables. This process considers the influence from multiple MBs by mining the equivalent information, instead of learning all MBs directly, where the strategy is different from multiple MB learning algorithms. According to Theorem 6, the variables in subset $\boldsymbol{Z}$ are common MB variables if at least one of the three conditions in Eq. (19) is satisfied for each target in $\boldsymbol{T}$, which can be formalized as the logical operation in Eq. (19).

$$\Theta_{\boldsymbol{T}}(\boldsymbol{Z}) = \bigwedge_{T \in \boldsymbol{T}} (\theta_1(\boldsymbol{Z}, T) \vee \theta_2(\boldsymbol{Z}, T) \vee \theta_3(\boldsymbol{Z}, T)) \tag{19}$$

- $\theta_1(\boldsymbol{Z}, T) = 1$ when $\exists \boldsymbol{Z}_T \subset \boldsymbol{MB}_T$ s.t. $\boldsymbol{Z} = \boldsymbol{Z}_T$, and 0 otherwise.
- $\theta_2(\boldsymbol{Z}, T) = 1$ when $\exists \boldsymbol{Z}_T \subset \boldsymbol{PC}_T$ s.t. $< \boldsymbol{Z}, \boldsymbol{Z}_T > \in \boldsymbol{EI}_T$, and 0 otherwise.
- $\theta_3(\boldsymbol{Z}, T) = 1$ when $\exists \boldsymbol{Z}_T \subset \boldsymbol{SP}_T$ s.t. $< \boldsymbol{Z}, \boldsymbol{Z}_T > \in \boldsymbol{EI}_C$, and 0 otherwise, where C is the common child of $\boldsymbol{Z}_T$ and T.

Specifically, for a target T and variable subset $\boldsymbol{Z}$, $\theta_1(\boldsymbol{Z}, T) = 1$ indicates that $\boldsymbol{Z}$ is a subset of the searched MB, and $\theta_2(\boldsymbol{Z}, T) = 1$ indicates that $\boldsymbol{Z}$ is equivalent with a subset of the searched PC set, and $\theta_3(\boldsymbol{Z}, T) = 1$ indicates that $\boldsymbol{Z}$ is equivalent with a subset of the searched SP set. $\boldsymbol{Z}$ satisfying one of the conditions contains critical information about this target, and thus variables in $\boldsymbol{Z}$ making $\Theta_{\boldsymbol{T}}(\boldsymbol{Z}) = 1$ are common MB variables of targets in $\boldsymbol{T}$. Thus, in Line 19, we obtain common MB variables $\boldsymbol{CCV_{T_S}}$ of any target subset $\boldsymbol{T_S}$ and target-specific MB variables $\boldsymbol{TCV}_T$ for each target T.

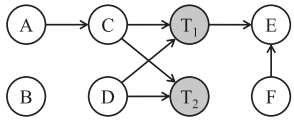


Fig. 5. A toy example to illustrate the efficiency of CTMB.

3.5 A Toy Example to Illustrate its Efficiency

In Fig. 5, a two-target example is provided, where equivalent information occurs once on each target. We revisit the trace of CTMB and explain the time efficiency against brute-force methods, which search the MB for different targets first and then take the intersection of MBs as common MB variables. Assume A and C contain equivalent information about T_1 and T_2. Sing-MB algorithm MMMB and multiple-MB algorithm TIE* are chosen as benchmarks. For MMMB, it is conducted on two targets respectively and outputs $MB[T_1] = \{C, D, E, F\}$ (152 CI-tests) and $MB[T_2] = \{C, D\}$ (106 CI-tests) with 258 CI-tests. Hence, the common MB variables found by MMMB are incomplete since A is ignored. For TIE*, it uses MMMB to search an initial MB ($MB[T_1]_1 = \{C, D, E, F\}$) first (152 CI-tests), then it removes an MB subset $\{C\}$ and searches the MB again, and obtains $MB[T_1]_2 = \{A, D, E, F\}$ with 116 CI-tests. The same process will be repeated when other MB subsets of T_1 ($\{D\}$, $\{E\}$, $\{F\}$, $\{A, C\}$) are removed (with 88, 88, 148, 62 CI-tests, respectively) until all MBs are discovered. Similarly, there are 74, 52, 52, 52, and 31 CI-tests when removing $\{C\}$, $\{D\}$, $\{E\}$, $\{F\}$, and $\{A, C\}$, respectively. Therefore, TIE* needs to use 1021 CI-tests. For CTMB, it also uses MMMB to search an initial MB first (258 CI-tests), then it constructs the $\boldsymbol{EI}$ for targets T_1, T_2 and child variable E, with 4, 7, and 20 CI-tests respectively. Then, we obtain $\boldsymbol{EI}_{T_1} = \{ < A, C > \}$, $\boldsymbol{EI}_{T_2} = \{ < A, C > \}$, $\boldsymbol{EI}_E = \varnothing$. Therefore, CTMB discovers complete common MB variable sets with 289 CI-tests.

Owing to Phase 3, CTMB does not need to discover all MBs, which not only improves the time efficiency, but also yields better results due to more small-scale conditioning sets in CI-tests compared with the brute-force methods. Detailed time complexity analysis and experiments are provided in Sections 4.3 and 5.1.

4 Applying CTMB to Multi-Label Feature Selection

To demonstrate the generality of CTMB proposed in Section 3.4, we apply CTMB to multi-label feature selection problem. Compared with the single-label feature selection, the additional introduced label correlations construct more complex relationships in multi-label data, including feature-feature, feature-label and label-label relationships. Although the label relationships are crucial, it is unreasonable to specially treat them as more important information.

Conversely, the ideal strategy is to consider the relationships of all variables in a unified framework. Therefore, in this section, we try to use CTMB to handle these various forms of complex relationships in consideration of its ability to map the complex relationships among features and labels to a BN model. And the process of constructing the skeleton of the graph on multi-label data naturally takes all types of relationships into consideration, which can be easily 'read' from the graph. In the following, the novel CTMB-driven multi-Label Feature Selection (CLFS) algorithm is present in Section 4.1. Subsequently, the relevance and redundancy of CLFS are analyzed in Section 4.2, and the time complexity of CLFS is analyzed in Section 4.3.

4.1 CLFS Algorithm

Pellet and Elisseeff have proved that MB is the optimal solution for single-label feature selection problem under the faithfulness condition [10], and the strongly relevant features are included in its MB set in terms of Kohavi-John feature relevance [9]. Thus, on each label, the independence property of MB indicates that the variable subset contains all of the predictive information about each corresponding label, while the minimality of it can guarantee the minimal redundancy in the variable set. As previously mentioned, in multi-label data, the union of MB sets can not be used directly as the selected feature subset due to the redundancy between MBs of different labels. While CTMB can be used to identify and select the common features simultaneously containing predictive information about several labels as many as possible to minimize the redundancy in the selected feature subset, which is just what Theorem 6 does, i.e., replacing the MB subset $\boldsymbol{Z}_i$ of multiple labels with a common equivalent feature subset $\boldsymbol{Z}$ and keeping the information constant.

It is worth mentioning that, using CTMB to directly search the common features of several labels might import some labels into the feature subset due to the case violating the Label-relation Assumption. Since labels are usually undetermined and thus can neither used as a factor to infer another label nor a feature to model a predictive learner (or classifier) in most real-world multi-label applications, it is necessary to remove them and find other predictive features. However, it does not mean that we can directly select features in the variable set $\boldsymbol{U}$ seeing that if a pair of labels are the parent or child of each other, the spouse will be ignored when searching MB in $\boldsymbol{U}$. To search the substitutes, we can remove the labels in the discovered MB set and continue to search the variables containing similar information until no label is included in the MB. Now, we present the CLFS in Algorithm 2.

CLFS finds an MB set of each label to detect the equivalent features for each label, so it inherits Phase 1 and Phase 2 of CTMB. We explain the additional components of Algorithm 2 below:

(1) Lines 3-8: Find the predictive common features shielded by label relations. Through removing the labels from the current $\boldsymbol{PC}_T$ set, the information loss about the label needs to be supplied with the features in the $\boldsymbol{PC}$ of each removed label. Thus, in Line 6, X is traversed from $\bigcup_{T_i \in \boldsymbol{PC}_T \cap \boldsymbol{T}} \boldsymbol{PC}_i - (\boldsymbol{PC}_T \cap \boldsymbol{T})$. Since the X could be a label, Lines 5-6 might be iterated several times.

(2) Lines 10-13: Search the common features and label-specific features. To minimize the redundancy as previously discussed, Line 11 finds the common feature subset $\boldsymbol{Z}$ containing information about as many labels as possible, which satisfies the three rules in Eq. (19). At the same time, CLFS can record the relationships between selected features and each label, i.e., "which labels does a selected feature in $\boldsymbol{Z}$ relate to". Then, the corresponding $\boldsymbol{Z}_T$ needs to be removed from the $\boldsymbol{PC}_T$ or $\boldsymbol{SP}_T$ to guarantee no redundancy about the same label. The above process is iterated until there are no features containing information about multiple labels ($|\boldsymbol{T}_S| > 1$). The remaining features in $\boldsymbol{PC}_T$ and $\boldsymbol{SP}_T$ are label-specific features of their corresponding label T.

Algorithm 2. The *CLFS* Algorithm

1: **Input:** Label set $\boldsymbol{T}$ and features set $\boldsymbol{U}$; A divide-and-conquer-based MB discovery algorithm $\mathbb{A}$ with significance level α.
2: CTMB (Phase 1, Phase 2)
3: **for** each $T \in \boldsymbol{T}$ **do**
4: **repeat**
5: $\boldsymbol{PC}_T \leftarrow \boldsymbol{PC}_T - \boldsymbol{T}$
6: $\boldsymbol{PC}_T \leftarrow \boldsymbol{PC}_T \cup \{X | X \in \bigcup_{T_i \in \boldsymbol{PC}_T \cap \boldsymbol{T}} \boldsymbol{PC}_i - (\boldsymbol{PC}_T \cap \boldsymbol{T})$ and $X \not\perp T | \boldsymbol{Z}$ for $\forall \boldsymbol{Z} \subset \boldsymbol{PC}_T\}$
7: **until** $\boldsymbol{PC}_T \cap \boldsymbol{T} = \varnothing$
8: CTMB (Phase 3: Lines 12 - 18)
9: **repeat**
10: Select $\boldsymbol{Z}$ to $\boldsymbol{CF}$ where $\Theta_{\boldsymbol{T}_S}(\boldsymbol{Z}) = 1$ for the most large-scale $|\boldsymbol{T}_S|$ ($\boldsymbol{T}_S \subset \boldsymbol{T}$).
11: $\boldsymbol{PC}_T \leftarrow \boldsymbol{PC}_T - \boldsymbol{Z}_T$, $\boldsymbol{SP}_T \leftarrow \boldsymbol{SP}_T - \boldsymbol{Z}_T$ for each T.
12: **until** for $\forall \boldsymbol{Z}$, $\Theta_{\boldsymbol{T}_S}(\boldsymbol{Z}) \neq 1$ for all $|\boldsymbol{T}_S| > 1$.
13: **Output:** Common features $\boldsymbol{CF}$, and label-specific features $\boldsymbol{PC}_T \cup \boldsymbol{SP}_T$ for each T.

Compared with traditional multi-label feature selection algorithms, the superiority of CLFS is reflected in three aspects: (1) Interpretability: CLFS not only selects predictive features but also interprets which labels a select feature influences, i.e., identifies the common features and label-specific features; (2) Practicability: CLFS automatically determines the number of selected features without training an additional classifier to achieve the optimal accuracy; (3) Theoretical Reliability: It can be proved that CLFS achieves maximum relevance and minimal redundancy.

4.2 Analyses of Relevance and Redundancy

In this subsection, we will give the theoretical analyses of relevance and redundancy.

4.2.1 Relevance

We prove that, by replacing $\boldsymbol{Z}_i$ with $\boldsymbol{Z}$ for each $T_i \in \boldsymbol{T}$, the obtained feature subset $\bigcup_{T_i \in \boldsymbol{T}} (\boldsymbol{MB}_i - \boldsymbol{Z}_i) \cup \boldsymbol{Z} - \boldsymbol{T}$[5] contains the same information as $\boldsymbol{U}$ about $\boldsymbol{T}$. Mathematically in other words, all features excluded by $\bigcup_{T_i \in \boldsymbol{T}} (\boldsymbol{MB}_i - \boldsymbol{Z}_i) \cup \boldsymbol{Z} - \boldsymbol{T}$ are independent of $\boldsymbol{T}$ conditioned on $\bigcup_{T_i \in \boldsymbol{T}} (\boldsymbol{MB}_i - \boldsymbol{Z}_i) \cup \boldsymbol{Z} - \boldsymbol{T}$. It is sufficient to prove the case with $\boldsymbol{T} =$

5. We use $\bigcup_{T_i \in \boldsymbol{T}} (\boldsymbol{MB}_i - \boldsymbol{Z}_i) \cup \boldsymbol{Z} - \boldsymbol{T}$ instead of $\bigcup_{T_i \in \boldsymbol{T}} (\boldsymbol{MB}_i - \boldsymbol{Z}_i) \cup \boldsymbol{Z}$ as in Theorem 6 since any label cannot be used to predict another label in the feature selection problem.

$\{T_i, T_j\}$ since any multi-label case is a direct consequence of the two-label case using induction on the number of variables involved in $\boldsymbol{T}$. According to Eq. (14), $\boldsymbol{MB}_i - \boldsymbol{Z}_i \cup \boldsymbol{Z}$ is a Markov blanket of T_i, denoted as $\boldsymbol{M}_i$. Thus,

$$T_i \perp \boldsymbol{U} - \boldsymbol{M}_i \cup \{T_j\} | \boldsymbol{M}_i. \tag{20}$$

Decompose the $\boldsymbol{U} - \boldsymbol{M}_i \cup \{T_j\}$ in Eq. (20) as:

$$\begin{aligned} \boldsymbol{U} - \boldsymbol{M}_i \cup \{T_j\} =& (\boldsymbol{U} - \boldsymbol{M}_i - \boldsymbol{M}_j) \cup \\ & (\boldsymbol{M}_j - \boldsymbol{M}_i - \{T_i\} \cup \{T_j\}). \end{aligned} \tag{21}$$

According to the Weak union property in Theorem 1, we have:

$$T_i \perp (\boldsymbol{U} - \boldsymbol{M}_i - \boldsymbol{M}_j) | (\boldsymbol{M}_j - \{T_i\} \cup \{T_j\}). \tag{22}$$

Due to the symmetry between T_i and T_j, a similar relationship will exist:

$$T_j \perp (\boldsymbol{U} - \boldsymbol{M}_i - \boldsymbol{M}_j) | (\boldsymbol{M}_i - \{T_j\} \cup \{T_i\}). \tag{23}$$

According to Theorem 1, we combine Eqs. (22) and (23), and obtain:

$$\begin{aligned} & \boldsymbol{U} - (\boldsymbol{M}_i \cup \boldsymbol{M}_j - \{T_i, T_j\}) \perp \{T_i, T_j\} \\ & | \boldsymbol{M}_i \cup \boldsymbol{M}_j - \{T_i, T_j\}. \end{aligned} \tag{24}$$

Thus, $I(\boldsymbol{T}, \boldsymbol{U}) = I(\boldsymbol{T}, \boldsymbol{M}_i \cup \boldsymbol{M}_j - \{T_i, T_j\})$, which means that the selected feature subset of CLFS contains all information about the labels and achieves the maximum relevance among the subsets of feature sets.

4.2.2 Redundancy

We continue to prove under the case $\boldsymbol{T} = \{T_i, T_j\}$. Assume that there exists a subset of $\boldsymbol{S} \subset \bigcup_{T_i \in \boldsymbol{T}} (\boldsymbol{MB}_i - \boldsymbol{Z}_i) \cup \boldsymbol{Z} - \boldsymbol{T}$ such that it also contains the same information as $\boldsymbol{U}$ about $\boldsymbol{T}$, then:

$$T_i \perp \boldsymbol{U} - \boldsymbol{S} | \boldsymbol{S} \tag{25}$$

We construct a subset of $\boldsymbol{M}_i$, $\boldsymbol{A} = \boldsymbol{M}_i \cap \{T_j\} \cup \boldsymbol{S}$, to assist the analysis. $\boldsymbol{M}_i$ can be written as the union of two sets $(\boldsymbol{M}_i - \boldsymbol{A}) \cup (\boldsymbol{M}_i \cap \boldsymbol{A})$. Thus, we have:

$$T_i \perp \boldsymbol{U} - \boldsymbol{M}_i - \{T_i\} | (\boldsymbol{M}_i - \boldsymbol{A}) \cup (\boldsymbol{M}_i \cap \boldsymbol{A}). \tag{26}$$

According to Eq. (25), extend the $\boldsymbol{S}$ as a more large-scale Mb $\boldsymbol{S} \cup \{T_j\} \cup (\boldsymbol{U} - \boldsymbol{M}_i - \{T_i\})$, which is equivalent to $(\boldsymbol{M}_i \cap \boldsymbol{A}) \cup (\boldsymbol{U} - \boldsymbol{M}_i - \{T_i\})$. Then, we have

$$T_i \perp \boldsymbol{M}_i - \boldsymbol{A} | (\boldsymbol{M}_i \cap \boldsymbol{A}) \cup (\boldsymbol{U} - \boldsymbol{M}_i - \{T_i\}). \tag{27}$$

If the Intersection property in Theorem 1 is satisfied here, then Eqs. (26) and (27) indicate that:

$$\begin{aligned} & T_i \perp (\boldsymbol{M}_i - \boldsymbol{A}) \cup (\boldsymbol{U} - \boldsymbol{M}_i - \{T_i\}) | \boldsymbol{M}_i \cap \boldsymbol{A} \\ \Rightarrow\ & T_i \perp \boldsymbol{U} - (\boldsymbol{M}_i \cap \boldsymbol{A}) - \{T_i\} | \boldsymbol{M}_i \cap \boldsymbol{A}. \end{aligned} \tag{28}$$

Thus, $\boldsymbol{M}_i \cap \boldsymbol{A}$ is an Mb of T_i. However, $\boldsymbol{M}_i$ is an MB of T_i, thus, $\boldsymbol{M}_i \cap \boldsymbol{A} = \boldsymbol{M}_i$. Also, $\boldsymbol{M}_i \subset \boldsymbol{M}_i \cap \boldsymbol{A}$, i.e., $(\boldsymbol{M}_i - \{T_j\}) \cup (\boldsymbol{M}_i \cap \{T_j\}) \subset \boldsymbol{S} \cup (\boldsymbol{M}_i \cap \{T_j\})$. Hence, $\boldsymbol{M}_i - \{T_j\} \subset \boldsymbol{S}$. Similarly, $\boldsymbol{M}_j - \{T_i\} \subset \boldsymbol{S}$. Since $\boldsymbol{S}$ is a subset of $(\boldsymbol{M}_i - \{T_j\}) \cup (\boldsymbol{M}_j - \{T_i\})$, the above three equations indicate that $\boldsymbol{S} = (\boldsymbol{M}_i - \{T_j\}) \cup (\boldsymbol{M}_j - \{T_i\})$.

In conclusion, if the Intersection property is satisfied, no redundancy exists in the $\bigcup_{T_i \in \boldsymbol{T}} (\boldsymbol{MB}_i - \boldsymbol{Z}_i) \cup \boldsymbol{Z} - \boldsymbol{T}$. While if the Intersection property is violated for Eqs. (26) and (27), then we can assert that $\boldsymbol{U} - \boldsymbol{M}_i - \{T_i\}$ and $\boldsymbol{M}_i - \boldsymbol{A}$ contain equivalent information about T_i and there might exist redundancy in $\bigcup_{T_i \in \boldsymbol{T}} (\boldsymbol{MB}_i - \boldsymbol{Z}_i) \cup \boldsymbol{Z} - \boldsymbol{T}$. We give an example to explain the redundancy brought by equivalent information. Assume that feature subsets $\{A, B\}$ and $\{C, D\}$ are equally effective to predict label T_1 since they contain equivalent information about T_1, but only $\{C, D\}$ can be used to predict T_2. Then, in $\{A, B, C, D\}$, there exists redundancy between $\{A, B\}$ and $\{C, D\}$, which could be reduced by removing $\{A, B\}$. The proposed CLFS algorithm tries to detect the features containing equivalent information, so the minimal redundancy is guaranteed in the selected features.

4.3 Time Complexity Analysis

Finally, we provide time complexity analysis as follows. The computational cost of the MB-based algorithms is measured via the number of CI-tests. Let $|*|$ denote the scale of variable set $*$ and p denote the largest scale of the parent-child set of any target. For Phase 1 in CTMB, the time complexity of the MB discovery process of any target is less than $O(2^p p|\boldsymbol{U}|)$, and thus the time complexity of Phase 1 is $O(2^p p|\boldsymbol{U}||\boldsymbol{T}|)$. For Phase 2, there are fewer than $C^2_{|\boldsymbol{T}|}$ pairs of targets connecting with each other and the actual operation for each pair is to traverse the pairwise dependence. Thus, the time complexity of Phase 2 is $O(2^p|\boldsymbol{U}||\boldsymbol{T}|^2)$. Let the scale of the child set of targets be c and the largest scale of $\boldsymbol{Z}$ in Phase 3 (Line 13) be z, then the computational cost is $O(2^p|\boldsymbol{U}|^z(|\boldsymbol{T}| + c))$. Normally, if only the pairwise dependencies are considered, z is set to 1, as followed by existing MB-based methods. The extra processes in CLFS possess lower time complexity. Let $m = max\{|\boldsymbol{T}|p, |\boldsymbol{T}|^2, |\boldsymbol{T}| + c\}$, then the time complexity of CTMB and CLFS is $O(2^p|\boldsymbol{U}|m)$. For better performance, z could be set higher so that multivariate dependence could be considered. Under these circumstances, the increase in running time is not obvious. The main reason is that the test results with large-scale $\boldsymbol{Z}$ and small-scale $\boldsymbol{Z}$ could be used to derive each other. For example, if $\boldsymbol{Z} \perp X$, then any subsets $\boldsymbol{Z}' \subset \boldsymbol{Z}$ satisfy $\boldsymbol{Z}' \perp X$, and the converse proposition could also simplify the computational process.

4.4 Difference With MB-MCF

As an MB-based multi-label feature selection algorithm, it is necessary to state the main difference between CLFS and MB-MCF, another MB-based method presented in our conference paper [35]: (1) From the aspect of discussed issues: Wu et al. [35] study multi-label feature selection problem and design the MB-MCF based on empirical knowledge without reliable theoretical guarantee. While in this paper, we discuss the MB discovery problem for variable set and present a complete theoretical framework, which provides the theoretical guarantee for CLFS. (2) From the aspect of the proposed algorithms: CLFS selects more relevant features than MB-MCF since it additionally considers the spouse variables, which could enhance the predictive power

of child features [7]. Moreover, due to the analyses of label relations, CLFS better shields its negative influence on the feature selection process, so that more relevant features can be identified. Based on CTMB, CLFS can discover most of the common MB variables, which help CLFS remove more redundant features than MB-MCF.

5 Experiments

We first verify the effectiveness of CTMB on synthetic data sets with foregone (in)dependence relationships in Section 5.1 by comparing precision, recall, and time efficiency. Afterwards, the multi-label feature selection experiments are conducted on real-world data sets to demonstrate the superiority of the CLFS against traditional algorithms, MB-MCF (in Section 5.2), and SHAP-based methods (in Section 5.3). We further present the relationships between labels and selected features on the Emotions data set in Section 5.4 to demonstrate the interpretability of CLFS.

5.1 Learn Common and Target-Specific MB Variables: Precision, Recall, and Time Efficiency

In this section, we present an evaluation of CTMB for the identification of common and target-specific variables in simulated data. The data sets are sampled from synthetic Bayesian networks with the simulation method presented in [36]. Simulated data allow us to evaluate methods in a controlled setting where the underlying mechanism and all MB variables of each target are exactly known. Detailed experiment settings are presented below.

Experiment Parameters on Synthetic Data. To validate CTMB and corresponding theory in this paper, each data set is set with different controlled parameters: (1) percentage of the targets that have direct relationships with each other (p_c); (2) percentage of the targets that have multiple MBs (p_m). The remaining settings to simulate a Bayesian network, are the same in all experiment groups, which are given in Table 1. For each target, we randomly choose 5-10 non-target variables and targets (their proportions are determined by p_c) as the MB. Among these targets, p_m of them have 5-10 equivalent MBs, which are induced by the probability distribution with equivalent information. Specifically, if variables X and Y contain equivalent information about T, then (a) for each combination of values of X and T such that $P(T = t|X = x) = p$, there exists a value y of variable Y such that $P(T = t|Y = y) = p$, and (b) for every combination of values of Y and T such that $P(T = t|Y = y) = p$, there exists a value x of variable X such that $P(T = t|\, X = x) = p$.

Comparing Algorithms [6]: Since there are no algorithms for the identification of common and target-specific MB variables, we deploy existing MB discovery algorithms to search the MB variables for different targets first and then take the intersection of MB sets of different targets as the common MB variables, and the remaining variables as the target-specific MB variables. Among extensive MB learning algorithms, we choose several representative algorithms from each type, including three single MB discovery algorithms (a simultaneous MB learning algorithm IAMB [13], two divide-and-conquer MB learning algorithms HITON-MB [15] and CCMB [18]) and two multiple MB discovery algorithms (KIAMB [16] and TIE* [12]). The characteristics of these types of algorithms are detailed in Section 2. The value of k in KIAMB is set to 10 (the average number of MBs). The MB discovery algorithm in CTMB is HITON-MB [15] and the parameter in its G^2-test [3] is set to 0.05.

TABLE 1
Experiment Parameters

Parameters	Settings
The number of targets	50
The number of non-target variables	1000
The number of training samples	5000
The number of MBs of each target	$\in$[1, 15]
The size of an MB of each target	$\in$[5, 15]

Metrics for Evaluation. The frequently used metrics $Precision$ and $Recall$ are adapted to measure the accuracy of the searched common and target-specific MB variables. $Precision$ is the fraction of retrieved true positives over the total amount of retrieved variables, and $Recall$ is the fraction of retrieved true positives over the total amount of true positives. Mathematically

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \tag{29}$$

where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively. The two metrics are calculated on each type of MB variable, and the average results of the two types are taken as the performance. The results are shown in Tables 2 and 3. Furthermore, the logarithmic CPU time is recorded to compute the time efficiency.

Performance Comparison. Tables 2 and 3 provide the average precision and recall of identification of common and target-specific MB variables. Each group keeps one of the p_c and p_m invariant and changes the other, to show the performance of CTMB and other comparing algorithms in different cases. We conclude from these results that CTMB constantly performs better than others in the cases satisfying ($p_c = 0$) and violating ($p_c \neq 0$) the Target-relation Assumption. Specifically for different comparing algorithms: (1) Single MB discovery algorithms IAMB, HITON-MB, and CCMB achieve lower recall but relatively higher precision with $p_m > 0$, which means that they fail to identify the two types of variables in the data sets due to the inability to solve the case with multiple MBs for some targets. (2) KIAMB uses a randomized strategy to discover multiple MBs, and thus it has unstable performance on both precision and recall, and can only capture part of these two types of variables although it is efficient. (3) TIE* has better precision and recall compared with other algorithms. However, it is computationally expensive. Like CTMB, TIE* also considers the common MB variables with equivalent information, whereas it tries to retrieve all MBs of each target at the first step, resulting in high computation time and statically low reliability. Hence, CTMB can be considered as the first algorithm targeting to distinguish between common and target-specific MB variables with reasonable time complexity.

6. Codes are collected in: http://home.ustc.edu.cn/~xingyuwu/MB.html

TABLE 2
Average Precision and Recall of Searched Common and Target-Specific MB Variables With Respect to the Percentage of the Targets That Have Multiple MBs

Metric	p_c	p_m	∩IAMB	∩HITON-MB	∩CCMB	∩KIAMB	∩TIE*	CTMB
Precision	$p_c = 0.5$	$p_m = 0$	0.745	0.919	0.792	0.415	0.746	**0.915**
		$p_m = 0.5$	0.413	0.579	0.567	0.612	0.759	**0.909**
		$p_m = 1$	0.192	0.315	0.287	0.697	0.787	**0.906**
Recall	$p_c = 0.5$	$p_m = 0$	0.659	0.958	0.979	0.625	0.912	**0.979**
		$p_m = 0.5$	0.216	0.305	0.312	0.679	0.915	**0.973**
		$p_m = 1$	0.113	0.152	0.198	0.713	0.903	**0.981**
Average Time ($lg(Time)$)			**0.473**	2.295	2.874	1.629	5.672	2.871

TABLE 3
Average Precision and Recall of Searched Common and Target-Specific MB Variables With Respect to the Percentage of the Targets That Have Direct Relationships With Each Other

Metric	p_c	p_m	∩IAMB	∩HITON-MB	∩CCMB	∩KIAMB	∩TIE*	CTMB
Precision	$p_c = 0$	$p_m = 0.5$	0.452	0.583	0.581	0.672	0.771	**0.915**
	$p_c = 0.5$		0.413	0.579	0.567	0.612	0.759	**0.909**
	$p_c = 1$		0.394	0.560	0.551	0.654	0.715	**0.910**
Recall	$p_c = 0$	$p_m = 0.5$	0.237	0.325	0.346	0.631	0.923	**0.970**
	$p_c = 0.5$		0.216	0.305	0.312	0.679	0.915	**0.973**
	$p_c = 1$		0.191	0.286	0.307	0.677	0.877	**0.965**
Average Time ($lg(Time)$)			**0.462**	2.131	2.559	1.503	5.379	2.812

5.2 Multi-Label Feature Selection: Accuracy

To demonstrate the performance of the extended CLFS for the multi-label feature selection problem, in this subsection, five state-of-the-art multi-label feature selection algorithms are compared with four frequently-used metrics. Details of these experiments are given as follows.

Multi-Label Data Sets.[7] The six data sets are taken from diverse application domains. The domains and standard statistics are provided in Table 4. $Cardinality$ denotes the average number of labels per instance, and $density$ normalizes the label cardinality by the number of labels.

Comparing Algorithms.[8] To validate the performance of CLFS, six state-of-the-art multi-label feature selection algorithms are compared, including SFUS [37], CSFS [38], MIFS [39], CMFS [40], MCLS [41], and the previously proposed MB-MCF [35]. These recently proposed algorithms reflect the effectiveness of multi-label feature selection from different perspectives (or metrics). To evaluate the effectiveness of proposed methods, we use the binary classifier SVM cooperating with the multi-label classification model BR [42] to decompose a multi-label problem into several independent binary problems first and compute their classification accuracies archived by selected features. The main consideration is that BR does not involve the label correlations, which could more clearly demonstrate the strengths of these compared algorithms in terms of addressing complex relationships in multi-label data. Additionally, since CLFS and MB-MCF measure the importance of features through uncovering the mechanisms rather than calculating the correlations, CLFS and MB-MCF do not need to predetermine the number of selected features, as shown in Fig. 6.

Metrics for Evaluation. For a fair comparison, we choose two example-based metrics $HammingLoss$ and $Ranking\ Loss$, and two label-based metrics F_{Macro} and F_{Micro} (macro-averaging and micro-averaging of F1-measure) [34], to measure the performances of multi-label classification results with selected features of each comparing algorithm. $HammingLoss$ evaluates the ratio of false outputting labels, including the missed relevant labels and the predicted irrelevant labels:

$$HammingLoss = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}|\hat{Y}_i \Delta Y_i|, \tag{30}$$

where m is the number of samples and n is the number of labels. $\hat{Y}_i$ and Y_i represent the predicted and real label set of the i-th sample, respectively. Δ denotes the symmetric difference between them.

TABLE 4
Details of the Multi-Label Data Sets

Data set	domain	#Features	#Labels	$cardinality$	$density$
Birds	audio	260	19	1.014	0.053
CAL500	music	68	174	26.044	0.150
Emotions	music	72	6	1.869	0.311
EUR-Lex	text	5000	201	2.213	0.011
Mediamill	video	120	101	4.376	0.043
NUS-WIDE	images	500	81	1.869	0.023

7. Data Source: http://mulan.sourceforge.net/datasets-mlc.html
8. Codes are collected in: http://home.ustc.edu.cn/~xingyuwu/Traditional-Multi-label-FS.zip

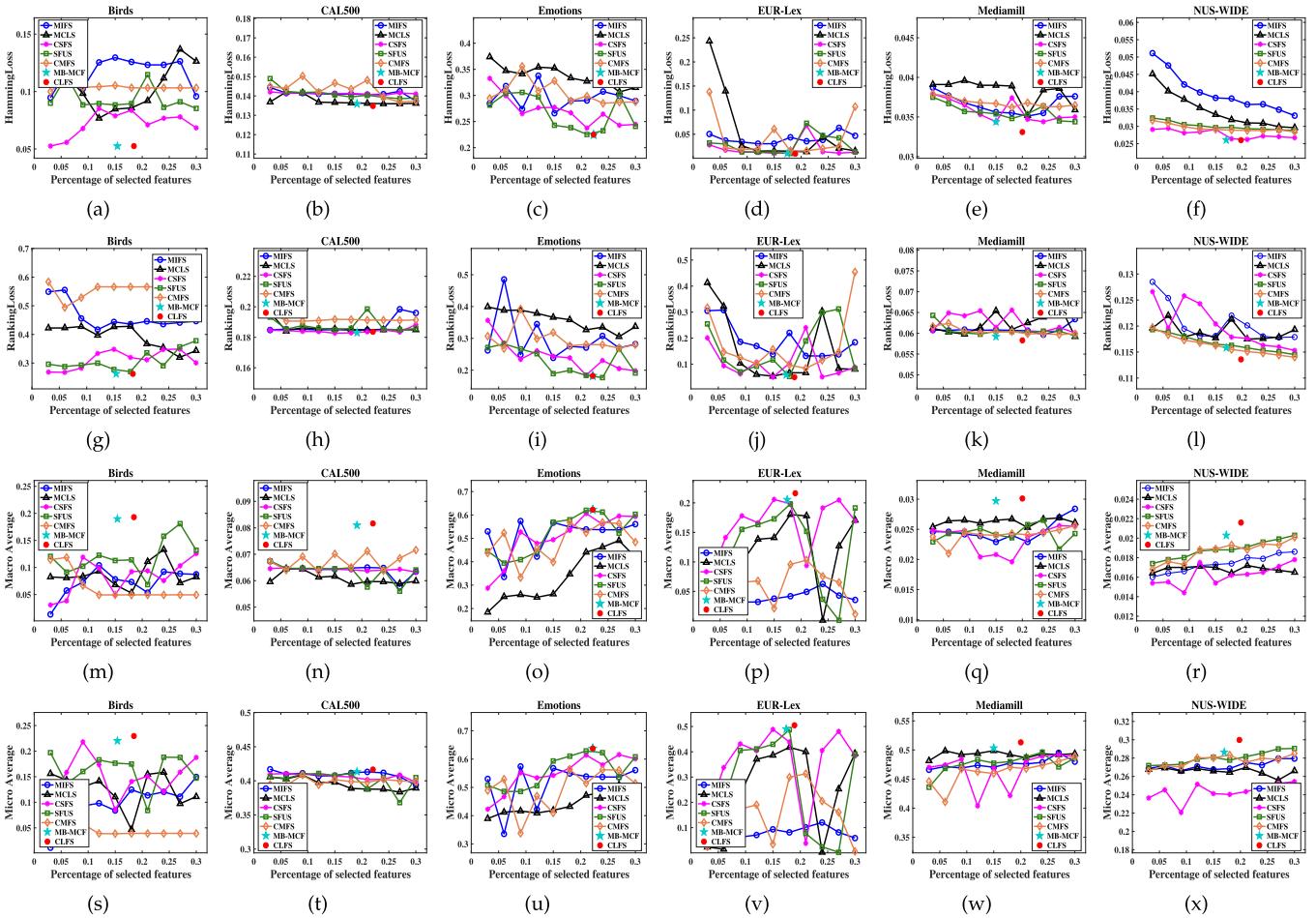


Fig. 6. The $HammingLoss$, $RankingLoss$, F_{Macro}, and F_{Micro} of CLFS and other state-of-the-art algorithms on six real-world data sets.

$RankingLoss$ evaluates the ratio of reversely ordered label pairs, i.e., an irrelevant label is ranked higher than a relevant label, which is calculated by:

$$RankingLoss = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|\hat{Y}_i||Y_i|}|\{(y_1, y_2)|f(x_i, y_1) \leq f(x_i, y_2), y_1 \in Y_i, y_2 \in \hat{Y}_i\}|, \quad (31)$$

where f denotes the intermediate real-valued function.

F_{Micro} is the weighted average arithmetic average of F_1-score (harmonic mean of $Precision$ and $Recall$) over all m samples, whereas F_{Macro} is an arithmetic average F_1-score of all n labels. Mathematically,

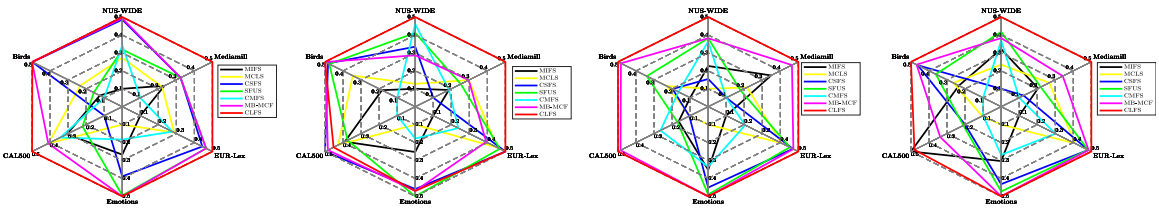
$$F_{Micro} = \frac{1}{n}\sum_{i=1}^{n}\frac{2TP_i}{2TP_i + FP_i + FN_i}. \quad (32)$$

$$F_{Macro} = \frac{\sum_{i=1}^{n} 2TP_i}{\sum_{i=1}^{n}(2TP_i + FP_i + FN_i)}. \quad (33)$$

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

Performance Comparison. We employ CLFS and other algorithms in comparison to select features first and then train the BR-SVM with these selected features. Each experiment is repeated 10 times with different training and test data, and we report the average performances, i.e., $HammingLoss$, $RankingLoss$, F_{Macro}, and F_{Micro}. As previously mentioned, traditional feature selection algorithms need to predetermine the number of features while CLFS and MB-MCF do not need to, therefore the percentage of the selected features is gradually turned in $\{0.03, 0.06, \ldots, 0.27, 0.3\}$ for these traditional algorithms. Similarly, the regularization parameters for all algorithms are searched from $\{0.01, 0.1, 0.3, \ldots, 0.9, 1\}$ by grid search. The MB discovery algorithm in CLFS and MB-MCF is HITON-MB [15] and the parameter in its G^2-test [3] is set as 0.05. Fig. 6 shows the average $HammingLoss$, $RankingLoss$, F_{Macro}, and F_{Micro} variation curves of different multi-label feature selection algorithms with respect to the percentage of selected features.

(1) *Comparison with traditional algorithms:* As mentioned previously, CLFS could automatically determine the number of selected features, and thus its performance trend is a red dot, instead of a curve. Based on the experimental results in Fig. 6, we make the following observations:(i) Under the same ratio of selected features, CLFS achieves the best performance in terms of four metrics compared with the traditional feature selection algorithms, as shown in Fig. 6. (ii) For $HammingLoss$, F_{Macro}, and F_{Micro}, CLFS consistently outperforms the best performance of these traditional algorithms. Especially on the large-scale data set (EUR-Lex, Mediamill, and NUS-WIDE), CLFS achieves significantly higher performance compared with traditional methods, which validates the practicability in the real-



(a) $Hamming\ Loss$ (b) $Ranking\ Loss$ (c) F_{Macro} (d) F_{Micro}

Fig. 7. Spider web diagrams showing the stability obtained on six multi-label data sets with $HammingLoss$, $RankingLoss$, F_{Macro}, and F_{Micro} of CLFS and other state-of-the-art multi-label feature selection algorithms.

world large-scale problem. (iii) For $RankingLoss$, CLFS is slightly worse than the CSFS algorithm on Birds data set, and SFUS algorithms on Emotions data set. It is reasonable since few algorithms could obtain the best performance on all these metrics simultaneously. Nevertheless, the performance of CLFS is also competitive compared with other algorithms. (iv) CLFS could automatically find the optimal number of features. The size of the feature set selected by CLFS exactly falls nearby the optimal point, and accordingly, CLFS achieves the best performance.

Overall, we can observe from Fig. 6 that, the superiority compared with traditional algorithms reflects on two sides: (i) CLFS could automatically determine the relatively optimal number of selected features; and (ii) CLFS achieves the best or very competitive performances, especially on large-scale data sets.

(2) *Comparison with MB-MCF:* On most data sets, CLFS selects more features than MB-MCF, and accordingly, CLFS achieves better performance than MB-MCF, which indicates that CLFS could better determine the relative optimal number of selected features. The additional features selected by CLFS generally include two types. Some of these features are the spouse features of labels, which could enhance the predictive ability of the direct effects (child features) of labels. Others are the relevant features shielded by label relations in the feature selection process, which are retrieved by CLFS to make up the predictive information loss contained by some labels. These improvements suggest that CLFS achieves better performance compared with MB-MCF.

We also note that MB-MCF and CLFS select the same features on the Emotions data set. Nonetheless, the relationships between labels and features mined by these algorithms are not exactly the same. In Section 5.4, we demonstrate the relationships retrieved by CLFS on Emotions, which is slightly different as compared with Fig. 4 in [35].

(3) *Stability Analysis:* To verify the stability of different methods, we draw four spider web diagrams in terms of each evaluation metric, i.e., $HammingLoss$, $RankingLoss$, F_{Macro}, and F_{Micro}. The best performance of each algorithm is normalized to a universal standard [0,0.5] so that the differences between the classification performances on different data sets do not influence the demonstration. Then, we present the stability index according to the value after normalization. The stability with different metrics is shown in Fig. 7, where the red line denotes the stability value of the proposed CLFS algorithm. We conclude from Fig. 7 that: (i) For $HammingLoss$, F_{Macro}, and F_{Micro}, the shapes of CLFS are regular hexagons, which means that CLFS obtains the most stable solution on each metric. For $RankingLoss$, CLFS is close to a regular hexagon. Nonetheless, CLFS more comes into contact with the regular hexagon than other algorithms. (ii) Compared with MB-MCF, the performance and stability of CLFS are significantly improved due to the theoretical guarantee proposed in this paper.

(4) *Experiment Time:* We recorded the CPU time for each algorithm on each data set in the above experiments. Table 5 provides the average running time under the same number of selected features with CLFS (except MB-MCF since it could determine the selected feature size).

We conclude from Table 5 that, the CPU time of CLFS is similar to MB-MCF but slightly higher than the traditional multi-label feature selection methods. Note that, MB-based feature selection methods usually have higher time complexity than traditional methods since they possess interpretability and theoretical guarantee [43]. Hence, the loss of time efficiency is unsurprising. However, these traditional methods need to execute many times to determine the optimal number of selected features, and the process using classifiers to obtain the predictability of a feature subset is far more time-consuming than the feature selection process, whereas CLFS could predetermine the number of the selected features. Therefore, the cost of time is reasonable coupled with many benefits of CLFS.

5.3 Compared With SHAP-Based Methods

SHAP (SHapley Additive exPlanation) Value [44] has become popular in the Explainable AI literature, which could also be applied for feature selection [45]. We compare the CLFS with the simplest general SHAP value feature selection procedure, which consists of three steps: (1) Choose a learning model; (2) Compute the SHAP value for each feature; (3) Select several highest-ranking features. Step (1) in the SHAP procedure chooses multi-label KNN and SVM as the classifiers, and Step (3) determines the number of selected features same as CLFS. Other experimental

TABLE 5
Experiment Time ($lg(Time)$) of Each Algorithm

Algorithm	MIFS	MCLS	CSFS	SFUS	CMFS	MB-MCF	CLFS
Birds	1.924	1.646	**1.409**	1.546	1.763	1.795	1.857
CAL500	4.793	4.788	4.802	**4.782**	4.792	4.841	4.880
Emotions	0.983	1.002	0.999	0.986	1.048	**0.957**	1.015
EUR-Lex	4.674	4.655	4.715	**4.649**	4.666	4.760	4.766
Mediamill	4.318	4.334	**4.296**	4.342	4.329	4.375	4.442
NUS-WIDE	4.166	4.176	**4.006**	4.181	4.184	4.233	4.274

TABLE 6
Performance of CLFS and Comparing SHAP Methods

Metric	Algorithm	Birds	CAL500	Emotions	EUR-Lex	Mediamill	NUS-WIDE
$HammingLoss \downarrow$	KNN+SHAP	0.1142	0.1825	0.2971	0.0129	0.0372	0.0334
	SVM+SHAP	0.0764	0.1793	0.2389	0.0825	0.0354	0.0293
	CLFS	0.0526	0.1348	0.2246	0.0093	0.0331	0.0260
$RankingLoss \downarrow$	KNN+SHAP	0.4195	0.2241	0.2997	0.1364	0.0632	0.1297
	SVM+SHAP	0.3247	0.1889	0.1846	0.0725	0.0612	0.1138
	CLFS	0.2625	0.1836	0.1825	0.0498	0.0583	0.1136
$F_{Macro} \uparrow$	KNN+SHAP	0.0742	0.0973	0.3974	0.1123	0.0220	0.1735
	SVM+SHAP	0.1258	0.1248	0.6012	0.1652	0.0285	0.1894
	CLFS	0.1932	0.1592	0.6230	0.2164	0.0301	0.2160
$F_{Micro} \uparrow$	KNN+SHAP	0.0346	0.3520	0.5281	0.4105	0.0469	0.2765
	SVM+SHAP	0.1688	0.3885	0.6105	0.4937	0.0473	0.2841
	CLFS	0.2297	0.4165	0.6374	0.5051	0.5132	0.2998

settings are the same as Section 5.2. Table 6 provides the average performance with different metrics. From Table 6, we can conclude that the performance of CLFS is consistently better than the SHAP-based methods. Furthermore, we find that SVM+SHAP achieves better performance than KNN+SHAP since the classifier used to evaluate the selected features is SVM. This phenomenon indicates that the SHAP-based method is learner-dependency while CLFS does not need to choose an extra learner.

5.4 An Example of the Interpretability of CLFS

Inherited from CTMB, CLFS naturally possesses the capacity of distinguishing the common features and label-specific features, and thus it can explicitly interpret which labels a feature influences. In this subsection, we further choose the Emotions data set as an example to demonstrate the interpretability of CLFS, because: (i) Emotions data set is derived from psychology and some known conclusion in the psychological study [46] can be used to validate our experimental results; (ii) Emotions data set contains 6 labels and 72 features, convenient to demonstrate in a figure.

These 72 features are extracted from musical context. And the 6 labels are derived from the Tellegen-Watson-Clark model of mood [46] as shown in Fig. 8, including 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). From Fig. 8, we can discover that label pairs (L_1, L_4), (L_2, L_5), and (L_3, L_6) are opposite emotions, which means that the labels in each pair should be influenced by similar features.

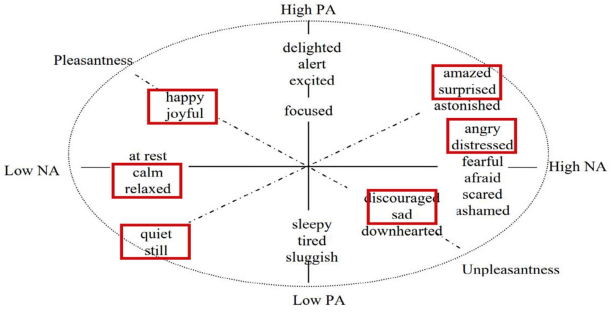


Fig. 8. The graphical representation of the Tellegen-Watson-Clark model [46], where the six labels in Emotions data set are amazed-surprised, happy-pleased, relaxing-calm, quiet-still, sad-lonely, and angry-aggressive.

The identified relationships between labels and selected features are provided in Fig. 9, where a dyed cell indicates that the corresponding feature has an effect on the corresponding label. We can observe from Fig. 9 that common features F_{21}, F_{40}, and F_{48} 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) share similar common features, which is consistent with the Tellegen-Watson-Clark model in the previous study [46].

Note that, although CLFS selects the same features as MB-MCF, the mined dependence relationships between labels and features are not exactly the same, which could be observed in the slight difference between Fig. 9 in this paper and Fig. 4 in [35]. Therefore, there exist dependence relationships unidentified when using MB-MCF, whereas the relationships between the feature and other labels are discovered. Under these circumstances, the two algorithms achieve similar performance.

6 Conclusion and Future Work

The identification of common MB variables and target-specific MB variables is an interesting topic due to their

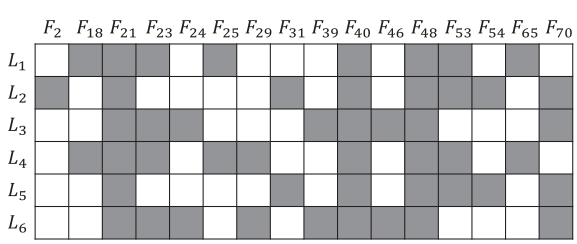


Fig. 9. The Identified relationship between each selected feature and each label in the Emotions data set. In the grid, each column corresponds to a feature selected by CLFS, and each row corresponds to a label (L_1: amazed-surprised, L_2: happy-pleased, L_3: relaxing-calm, L_4: quiet-still, L_5: sad-lonely, L_6: angry-aggressive). The shaded cell indicates that the corresponding feature affects the corresponding label.

imperative role in the underlying mechanism. In this paper, we investigate the theoretical property of common MB variables of multiple targets and find that the common MB variables are determined by equivalent information following different mechanisms with or without the existence of direct target dependence. Based on extensive analyses, the discovery and distinguishing algorithm CTMB is proposed to identify these two types of variables without mining all of the multiple MBs. Furthermore, we apply CTMB to the multi-label feature selection problem to improve the accuracy and interpretability. Experiments on synthetic and real-world data demonstrate the efficacy of these proposed methods. To our knowledge, it is the first study focusing on the common and target-specific MB variable discovery for a variable set.

The proposed concept of common MB variables, is frequently used and considered, however, it has not been formally discussed before in literature. We believe that some research could benefit from this work, which is presented below to prompt the possible future work.

- Common features mining. Some of the real-world applications can use CTMB to mine the common features of multiple labels from the data. These results would instruct data-analysts about the underlying knowledge and information and induce more potential solutions.
- MB-based learning tasks. Previous literature on multi-label learning [47] has pointed that label-specific features can facilitate the prediction of its corresponding label. However, these learning methods exploit extra steps to identify these features. With the CLFS, the predictor or classifier modeled on MB can distinguish the common and label-specific features before training.
- Improve computational efficiency of CTMB and CLFS. We have previously designed an acceleration strategy of MB discovery algorithms, called Pipeline Machine [18]. At face value, this suggests that organizing the computations with a data structure and modifying the algorithm with a parallel architecture can help decrease the run time.

ACKNOWLEDGMENTS

We appreciate the comments from anonymous reviewers, which helped to improve the paper.

REFERENCES

[1] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: Algorithms and empirical evaluation," *J. Mach. Learn. Res.*, vol. 11, no. 1, pp. 171–234, 2010.

[2] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "Local causal and Markov blanket induction for causal discovery and feature selection for classification part II: Analysis and extensions," *J. Mach. Learn. Res.*, vol. 11, no. 1, pp. 235–284, 2010.

[3] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA, USA: Morgan kaufmann, 1988.

[4] T. Gao and Q. Ji, "Local causal discovery of direct causes and effects," in *Proc. 29th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2512–2520.

[5] J.-P. Pellet and A. Elisseeff, "Finding latent causes in causal networks: An efficient approach based on Markov blankets," in *Proc. 23rd Int. Conf. Neural Inf. Process. Syst.*, 2009, pp. 1249–1256.

[6] R. Ram and M. Chetty, "A Markov-blanket-based model for gene regulatory network inference," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 2, pp. 353–367, Mar./Apr. 2011.

[7] I. Guyon, C. Aliferis, and A. Elisseeff, "Causal feature selection," in *Computational Methods of Feature Selection*. London, U.K./Boca Raton, FL, USA: Chapman and Hall/CRC, 2007, pp. 79–102.

[8] R. Cai, Z. Zhang, and Z. Hao, "BASSUM: A Bayesian semi-supervised method for classification feature selection," *Pattern Recognit.*, vol. 44, no. 4, pp. 811–820, 2011.

[9] I. Tsamardinos and C. F. Aliferis, "Towards principled feature selection: Relevancy, filters and wrappers," in *Proc. 9th Int. Workshop Artif. Intell. Statist.*, 2003, pp. 300–307.

[10] J.-P. Pellet and A. Elisseeff, "Using Markov blankets for causal structure learning," *J. Mach. Learn. Res.*, vol. 9, no. 7, pp. 1295–1342, 2008.

[11] A. R. Masegosa and S. Moral, "A Bayesian stochastic search method for discovering Markov boundaries," *Knowl. Based Syst.*, vol. 35, no. 11, pp. 211–223, 2012.

[12] A. Statnikov, N. I. Lytkin, J. Lemeire, and C. F. Aliferis, "Algorithms for discovery of multiple Markov boundaries," *J. Mach. Learn. Res.*, vol. 14, no. 2, pp. 499–566, 2013.

[13] I. Tsamardinos, C. F. Aliferis, A. R. Statnikov, and E. Statnikov, "Algorithms for large scale Markov blanket discovery," in *Proc. Florida Artif. Intell. Res. Soc. Conf.*, 2003, pp. 376–380.

[14] I. Tsamardinos, C. F. Aliferis, and A. Statnikov, "Time and sample efficient discovery of Markov blankets and direct causal relations," in *Proc. 9th ACM Int. Conf. Knowl. Discov. Data Mining*, 2003, pp. 673–678.

[15] C. F. Aliferis, I. Tsamardinos, and A. Statnikov, "HITON: A novel Markov blanket algorithm for optimal variable selection," in *Proc. Amer. Med. Informat. Assoc. Annu. Symp.*, 2003, pp. 21–25.

[16] J. M. Pena, R. Nilsson, J. Björkegren, and J. Tegnér, "Towards scalable and data efficient learning of Markov boundaries," *Int. J. Approx. Reasoning*, vol. 45, no. 2, pp. 211–232, 2007.

[17] X. Wu, B. Jiang, K. Yu, and H. Chen, "Separation and recovery Markov boundary discovery and its application in EEG-based emotion recognition," *Inf. Sci.*, vol. 571, no. 9, pp. 262–278, 2021.

[18] X. Wu, B. Jiang, K. Yu, C. Miao, and H. Chen, "Accurate Markov boundary discovery for causal feature selection," *IEEE Trans. Cybern.*, vol. 50, no. 12, pp. 4983–4996, Dec. 2020.

[19] X. Wu, B. Jiang, Y. Zhong, and H. Chen, "Tolerant Markov boundary discovery for feature selection," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, 2020, pp. 2261–2264.

[20] X. Liu and X. Liu, "Swamping and masking in Markov boundary discovery," *Mach. Learn.*, vol. 104, no. 1, pp. 25–54, 2016.

[21] B. Huang, K. Zhang, P. Xie, M. Gong, E. P. Xing, and C. Glymour, "Specific and shared causal relation modeling and mechanism-based clustering," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 13 510–13 521.

[22] Y. Lin, Q. Hu, J. Liu, and J. Duan, "Multi-label feature selection based on max-dependency and min-redundancy," *Neurocomputing*, vol. 168, no. 11, pp. 92–103, 2015.

[23] K. Yu et al., "Tornado forecasting with multiple Markov boundaries," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2015, pp. 2237–2246.

[24] J. M. Pena, R. Nilsson, J. Björkegren, and J. Tegnér, "Towards scalable and data efficient learning of Markov boundaries," *Int. J. Approx. Reasoning*, vol. 45, no. 2, pp. 211–232, 2007.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

[26] J. H. McDonald, *Handbook of Biological Statistics*. Baltimore, MD, USA: Sparky House Publishing, 2009.

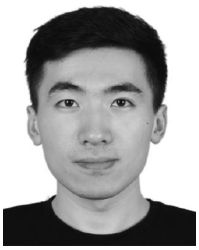
[27] J. M. Pena, "Learning Gaussian graphical models of gene networks with false discovery rate control," in *Proc. Eur. Conf. Evol. Comput. Mach. Learn. Data Mining Bioinf.*, 2008, pp. 165–176.

[28] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, "Kernel-based conditional independence test and application in causal discovery," 2012, *arXiv:1202.3775*.

[29] K. Yu et al., "Causality-based feature selection: Methods and evaluations," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–36, 2020.

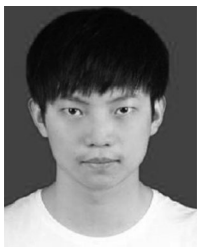
[30] S. Yaramakala and D. Margaritis, "Speculative Markov blanket discovery for optimal feature selection," in *Proc. 5th IEEE Int. Conf. Data Mining*, 2005, pp. 809–812.

[31] R. E. Neapolitan, *Learning Bayesian Networks*. Upper Saddle River, NJ, USA: Prentice Hall, 2004.
[32] H. Liu, L. Liu, and H. Zhang, "Ensemble gene selection by grouping for microarray data classification," *J. Biomed. Informat.*, vol. 43, no. 1, pp. 81–87, 2010.
[33] A. Statnikov and C. F. Aliferis, "Analysis and computational dissection of molecular signature multiplicity," *PLoS Comput. Biol.*, vol. 6, no. 5, 2010, Art. no. e1000790.
[34] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
[35] X. Wu, B. Jiang, K. Yu, H. Chen, and C. Miao, "Multi-label causal feature selection," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 6430–6437.
[36] A. Statnikov and C. F. Aliferis, "TIED: An artificially simulated dataset with multiple Markov boundaries," in *Proc. Workshop Causality: Objectives Assessment NIPS*, 2010, pp. 249–256.
[37] Z. Ma, F. Nie, Y. Yang, J. R. R. Uijlings, and N. Sebe, "Web image annotation via subspace-sparsity collaborated feature selection," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1021–1030, Aug. 2012.
[38] X. Chang, F. Nie, Y. Yang, and H. Huang, "A convex formulation for semi-supervised multi-label feature selection," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1171–1177.
[39] L. Jian, J. Li, K. Shu, and H. Liu, "Multi-label informed feature selection," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1627–1633.
[40] A. Braytee, W. Liu, D. R. Catchpoole, and P. J. Kennedy, "Multi-label feature selection using correlation information," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2017, pp. 1649–1656.
[41] R. Huang, W. Jiang, and G. Sun, "Manifold-based constraint Laplacian score for multi-label feature selection," *Pattern Recognit. Lett.*, vol. 112, no. 9, pp. 346–352, 2018.
[42] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, 2007.
[43] K. Yu, J. Li, W. Ding, and T. D. Le, "Multi-source causal feature selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2240–2256, Sep. 2020.
[44] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, *arXiv: 1705.07874*.
[45] R. Guha, A. H. Khan, P. K. Singh, R. Sarkar, and D. Bhattacharjee, "CGA: A new feature selection model for visual human action recognition," *Neural Comput. Appl.*, vol. 33, no. 10, pp. 5267–5286, 2021.
[46] A. Tellegen, D. Watson, and L. A. Clark, "On the dimensional and hierarchical structure of affect," *Psychol. Sci.*, vol. 10, no. 4, pp. 297–303, 1999.
[47] M.-L. Zhang and L. Wu, "LIFT: Multi-label learning with label-specific features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 107–120, Jan. 2015.



Xingyu Wu received the BSc degree from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2018. He is currently working toward the PhD degree with the School of Computer Science and Technology, University of Science and Technology of China (USTC), Hefei, China. His research interests include causality-based machine learning, causal learning, and causal inference. He has authored or coauthored some scientific papers in prestigious journals and conferences, and was a reviewer for the *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Knowledge and Data Engineering*, *IEEE/CAA Journal of Automatica Sinica*, and PC member of AAAI and EMNLP.

Bingbing Jiang received the BSc degree from the Chongqing University of Posts and Telecommunications (CQUPT), Chongqing, China, in 2014, and the PhD degree from the School of Computer Science and Technology, University of Science and Technology of China (USTC), Hefei, China, in 2019. He is currently a lecturer with the Hangzhou Normal University, Hangzhou, China. His research interests include Bayesian learning, feature selection, semi-supervised learning, and multi-view learning. He has published more than 10 scientific papers like the *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Cybernetics*, *IEEE Transactions on Emerging Topics in Computational Intelligence*, AAAI and CIKM. He also served as a reviewer of the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE Transactions on Neural Networks and Learning Systems, *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Image Processing* and *IEEE Transactions on Emerging Topics in Computational Intelligence*.



Yan Zhong received the BSc degree from the University of Northwestern Polytechnical University, Xi'an, China, in 2019, and the MSc degree from the University of Science and Technology of China, Hefei, China, in 2022. He is currently working toward the PhD degree with the School of Mathematical Sciences, Peking University (PKU), Beijing, China. His research interests include multi-label learning, feature selection, data mining, deep learning, and machine learning.

Huanhuan Chen (Senior Member, IEEE) received the BSc degree from the University of Science and Technology of China (USTC), Hefei, China, in 2004, and the PhD degree in computer science from the University of Birmingham, Birmingham, U.K., in 2008. He is currently a full professor with the School of Computer Science and Technology, USTC. His research interests include neural networks, Bayesian inference, and evolutionary computation. He received the 2015 International Neural Network Society Young Investigator Award, the 2012 IEEE Computational Intelligence Society Outstanding PhD Dissertation Award, the *IEEE Transactions on Neural Networks* Outstanding Paper Award (bestowed in 2011 and only one paper in 2009), and the 2009 British Computer Society Distinguished Dissertations Award. He is an associate editor of the *IEEE Transactions on Neural Networks and Learning Systems*, and the *IEEE Transactions on Emerging Topics in Computational Intelligence*.