A Light Causal Feature Selection Approach to High-Dimensional Data

Zhaolong Ling, Ying Li, Yiwen Zhang*, Kui Yu, Peng Zhou, Bo Li, and Xindong Wu, *Fellow, IEEE*

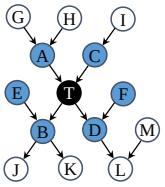
Abstract—Causal feature selection has received increasing attention in recent years. However, the state-of-the-art causal feature selection algorithms use the conditional independence tests, which require enumerating conditioning sets, leading to an exponential increase in computational complexity along with an increase in feature space. To address this problem, in this paper, we theoretically analyze the unique performance of causal features in mutual information, and propose a novel Causal Feature Selection algorithm using Mutual Information, called CFS-MI. Specifically, CFS-MI separately instantiates the pairwise comparison of mutual information in two stages to reduce computational complexity, and thus improves the efficiency on high-dimensional data. Extensive experiments on 5 benchmark Bayesian networks and 16 real-world datasets validate that CFS-MI has comparable accuracy compared to 7 state-of-the-art causal feature selection algorithms, while presenting more superior computational efficiency.

Index Terms—Causal Feature Selection, Mutual Information, Markov Blanket.

1 Introduction

Causal feature selection aims to identify the Markov blanket (MB) of a class variable to improve the robustness and explanatory capability of predictive models [1], [2], [3]. Under the faithfulness assumption, the MB of a target variable in a Bayesian network (BN) is unique and consists of its parents (direct causes), children (direct effects), and spouses (other parents of these children), which represents the local causal relationships of the variable, as illustrated in Fig. 1 (a). Since all other features are probabilistically independent of a class variable conditioned on its MB, the MB of the class variable is the optimal and minimal feature subset with maximum predictivity for classification [4], [5].

Existing causal feature selection algorithms can be divided into two categories: simultaneous and divide-and-conquer algorithms [6]. The simultaneous algorithms discover parents, children (PC), and spouses of a target variable simultaneously, using the MB as a conditioning set. The accuracy of simultaneous algorithms is relatively low for the insufficient data samples; thus, it is difficult to apply them in practice [7]. To improve the accuracy of MB discovery with limited data, the divide-and-conquer algorithms have divided the MB discovery into PC and spouses discovery which have been widely discussed. However, as shown in Fig. 1 (b), the existing divide-and-conquer algorithms use enumerating conditioning sets in PC discovery, requiring numerous conditional independence tests, leading to a generally exponential computational complexity [8]. Consequently, the state-of-the-art causal feature selection algorithms have high computational costs or even unacceptable time consumption when high-dimensional data are encountered [6]. However, feature selection is applied to high-dimensional data as a data dimensionality reduction technique to reduce computational costs [9], [10]. Although causal feature selection improves the interpretability of prediction models, the original intention of the state-of-the-art causal feature selection algorithms is contrary to the purpose that feature selection can reduce dimensionality on high-dimensional data. Thus, we hope to design a low-complexity causal feature selection algorithm to solve the time challenge when high-dimensional data are encountered.



(a) A simple Bayesian network

Conditioning Set		Dependency of T and D	Number of CI-tests
Size	Variables		
0	∅	T ⫫̸ D | ∅	Combination $\binom{0}{3}$
1	A	T ⫫̸ D | {A}	Combination $\binom{1}{3}$
1	B	T ⫫̸ D | {B}	
1	C	T ⫫̸ D | {C}	
2	A, B	T ⫫̸ D | {A, B}	Combination $\binom{2}{3}$
2	A, C	T ⫫̸ D | {A, C}	
2	B, C	T ⫫̸ D | {B, C}	
3	A, B, C	T ⫫̸ D | {A, B, C}	Combination $\binom{3}{3}$

2^3

(b) The conditional independence tests process of identifying D (child of T)

Fig. 1. (a) In a faithful Bayesian network, the Markov blanket (in blue) of the target variable T (in black) consists of its parents $\{A,\ C\}$, children $\{B,\ D\}$, and spouses $\{E,\ F\}$. (b) The conditional independence (CI) tests are used to identify a true child D in the PC discovery of existing algorithms. This process requires enumerating the conditioning sets $\{A,\ B,\ C\}$ from size 0 to 3, and the total number of CI tests is 2^3. Thus, if the size of the conditioning set reaches n, the CI tests will be conducted 2^n times.

Mutual information refers to the strength of the relationship between two variables [11], [12]. The closer the relationship between them, the larger the quantity of the mutual information, and the more likely an edge is to

- *Z. Ling, Y. Li, Y. Zhang, P. Zhou and B. Li are with the School of Computer Science and Technology, Anhui University, Hefei, Anhui, 230601, China. Email: zlling@ahu.edu.cn, ahu_liying@163.com, zhangyiwen@ahu.edu.cn, ahulibo@163.com and zhoupeng@ahu.edu.cn.*
- *K. Yu and X. Wu are with the Key Laboratory of Knowledge Engineering with Big Data (the Ministry of Education of China), and the School of Computer Science and Information Technology, Hefei University of Technology, Hefei, 230009, China. E-mail: yukui@hfut.edu.cn, xwu@hfut.edu.cn.*

*(*Corresponding author: Yiwen Zhang.)*

connect [9]. Because of the symmetry and probabilistic correlation of mutual information [9], it can not only measure the correlation between two variables, but also filter redundant relationships between the variables (i.e., non-causal features [10]) through pairwise comparison of mutual information regarding different variables without enumerating the conditioning set [13].

Thus, to address this challenging issue, we propose a light approach that uses mutual information for causal feature selection. The main contributions of this paper are summarized as follows:

- We theoretically analyze the mutual information relationships between the target variable and its PC of each variable in PC, and between the target variable and its spouses, and thus use pairwise comparisons instead of enumerating conditioning sets to measure the relationships between variables.
- We propose CFS-MI, the first Causal Feature Selection algorithm using Mutual Information. Based on our analysis, CFS-MI uses mutual information to identify the MB of the class variable, reducing the computational complexity from exponential to quadratic through pairwise comparisons.
- We conduct the experiments on five benchmark BNs and sixteen real-world datasets to show that CFS-MI has comparable accuracy, but significantly better efficiency than seven state-of-the-art causal feature selection algorithms.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 introduces the basic definitions and notations. Section 4 analyzes the mutual information and proposes our new algorithm. In section 5, we present and discuss the experimental results, and finally Section 6 concludes the paper.

2 Related work

The dimensional challenges brought by high-dimensional data exist in data mining and machine learning, and technologies at various stages are developed to work with high-dimensional data [14], [15]. For instance, to deal with the rapid data growth, the direct data processing technology POCLib (Processing On Compression Library) [16] is applied to compressed high-dimensional data to reduce the time and space pressure in text analytics and solve the complexity problem caused by large-scale data. However, the data dimensionality of a large corpus after text analytics is often high, and the most representative text features from the vector representation of text need to be identified [17]. The most effective method to solve this problem is by reducing data dimensionality through feature selection [15]. Feature selection is applied to high-dimensional data as a data dimensionality reduction technique [18], [19], which selects subsets from the original feature space without changing the original feature space [20]. Among them, causal feature selection utilizes BN and MB theory to identify potential causal relationships, which is more interpretable and robust to prediction models and attracts much more attention [6]. Many causal feature selection algorithms have been proposed. According to the framework of these algorithms, they are divided into two categories, i.e., simultaneous and divide-and-conquer algorithms.

The simultaneous algorithms appeared earlier, among which the Growth-Shrink algorithm (GS) [21] was an earlier version. GS searches for possible MB during the growth phase and removes false positive MB during the shrinkage phase. This approach became the fundamental framework of many later algorithms. Tsamardinos et al. proposed the Incremental Association MB (IAMB) [22] to improve the accuracy of GS by reordering the variables during each iteration. Based on the IAMB, its variants were proposed thereafter, including inter-IAMB [22], Fast-IAMB [23], and KIAMB [7]. Since these algorithms use the entire MB as a conditioning set, they are efficient but require exponential data samples to ensure high accuracy.

To improve the accuracy of small-sized sample data, divide-and-conquer algorithms have appeared. Min-Max MB (MMMB) [24] first used the divide-and-conquer strategy to address this problem. HITON-MB [25] introduced the interleave concept based on the former. However, neither of these two approaches can ensure the correct MB in a faithful BN. Thus, the Parents-and-Children-based MB algorithm (PCMB) [7] utilizes symmetry checking to ensure that the correct results can be obtained. Furthermore, Iterative Parent-Child-based search of MB (IPCMB) [26] improves the former checking stage and can identify spouses faster. Gao et al. recently proposed the Simultaneous MB (STMB) [27], which used the strategy of simultaneous algorithms in spouses discovery and thus has the disadvantages of simultaneous algorithms. By solving this problem, Balanced MB discovery (BAMB) [8] and Efficient and Effective MB discovery (EEMB) [28] improved the accuracy of small-sized sample data. However, all existing divide-and-conquer algorithms are based on enumerating the conditioning sets [29], thus require numerous conditional independence tests, unavoidably leading to exponential time consumption of computational resources.

In summary, the state-of-the-art causal feature selection algorithms face the dilemma of high computational complexity. Thus, to avoid unacceptable time consumption of these causal feature selection algorithms, this paper proposes a light algorithm using a divide-and-conquer strategy based on mutual information.

3 Notations and Definitions

In this section, we will introduce the relevant basic definitions and theorems. Table 1 summarizes the notations used in this paper.

Definition 1 (Faithfulness) [30]. A BN is presented by a directed acyclic graph $\mathbb{G}$ and a joint probability distribution $\mathbb{P}$ over a variable set $\mathbf{U}$. $\mathbb{G}$ is faithful to $\mathbb{P}$ iff every conditional independence present in $\mathbb{P}$ is entailed by $\mathbb{G}$ and the Markov condition. Further, $\mathbb{P}$ is faithful to $\mathbb{G}$ iff $\mathbb{G}$ is faithful to $\mathbb{P}$.

Definition 2 (D-separation) [3]. A path π between X and Y given $\mathbf{S} \subseteq \mathbf{U} \setminus \{X, Y\}$ is open, iff (1) every collider on π is in $\mathbf{S}$ or has a descendant in $\mathbf{S}$ and (2) no other non-colliders on π are in $\mathbf{S}$. Otherwise, the path π is blocked. X and Y are d-separated given $\mathbf{S}$ iff every path between X and Y is blocked by $\mathbf{S}$.

TABLE 1
Summary of notations

Symbol	Meaning
$\mathbf{U}$	a variable set
T, X, Y, Z	a variable
t, x, y, z	a discrete value that a variable may take
$\mathbf{S}$	a conditioning set within $\mathbf{U}$
$X \perp\!\!\!\perp Y \mid \mathbf{S}$	X is conditionally independent of Y given $\mathbf{S}$
$X \not\perp\!\!\!\perp Y \mid \mathbf{S}$	X is conditionally dependent on Y given $\mathbf{S}$
$\mathbf{PC}_T$	parents and children of T
$\mathbf{SP}_T$	spouses of T
$\lvert . \rvert$	the size of a set
δ	predefined threshold
$P(x)$	the probability of $X = x$
$H(X\|Y)$	the information entropy of X given Y
$I(X;Y\|Z)$	the mutual information between X and Y given Z
$SU(X;Y)$	a normalized form of $I(X;Y)$

Definition 3 (Information Entropy) [11]. Given a variable X, where x is a value that X may take, the information entropy of X is defined as follows:

$$H(X) = -\Sigma_x P(x) log P(x). \tag{1}$$

This value defines the uncertainty of the variable. The information entropy of X is defined as follows after observing the value of another variable Y:

$$H(X|Y) = -\Sigma_y P(y) \Sigma_x P(x|y) log P(x|y). \tag{2}$$

This value measures the degree of influence of one variable on another variable. However, it is asymmetrical because it is fixed for each variable and its corresponding condition. Thus we introduce the following definition of mutual information.

In Eqs. (1) and (2), $P(x)$ is the prior probability of $X = x$ (X takes value x) and $P(x|y)$ is the posterior probability of $X = x$ given that $Y = y$. According to Eqs. (1) and (2), the mutual information between X and Y is represented by $I(X; Y)$, defined as:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \Sigma x, y P(x,y) log \frac{P(x,y)}{P(x)P(y)}. \end{aligned} \tag{3}$$

As in Eq. (3), the mutual information is defined as the strength of correlation between two variables (i.e., the weakening degree of uncertainty for the other variable when one of the variables is known). It is only related to these two variables and thus has symmetry. The stronger the relationship between the two variables, the greater the mutual information, and this will measure the correlation between two variables.

According to Eq. (3), the conditional mutual information between X and Y given Z can be defined as follows:

$$\begin{aligned} &I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \\ =&\Sigma_{z \in Z} P(z) \Sigma_{x \in X, y \in Y} P(x,y|z) log \frac{P(x,y|z)}{P(x|z)P(y|z)}. \end{aligned} \tag{4}$$

Referring to the definition of conditional mutual information, this value can be understood as the strength of the relationship between the two variables after a given condition. Thus, we use conditional mutual information to measure the correlation between two variables given another variable.

Theorem 1 [31]. Under the faithfulness assumption, for $X, Y \in \mathbf{U}$, there is an edge between X and Y iff $X \not\perp\!\!\!\perp Y \mid \mathbf{S}$, for $\forall\ \mathbf{S} \subseteq \mathbf{U} \setminus \{X,\ Y\}$.

Definition 2 and Theorem 1 illustrate that the probability distributions and edge connections on the BN remain consistent under the faithfulness assumption. They are also the basis for applying mutual information theory to the BN.

Definition 4 (V-Structure) [3]. The triplet of variables X, Y, and Z form a V-structure if Z has two incoming edges from X and Y, forming $X \rightarrow Z \leftarrow Y$, and X is not adjacent to Y. Here, X and Y are each other spouses.

Definition 5 (Strongly relevant feature) [32]. For variable X, Y is its strongly relevant feature, iff $I(X; Y|\mathbf{U} \setminus \{Y\}) > 0$.

Definition 6 (Weakly relevant feature) [32]. For variable X, Y is its weakly relevant feature, iff $I(X; Y|\mathbf{U} \setminus \{Y\}) = 0$ and $\exists\ \mathbf{S} \subseteq \mathbf{U} \setminus \{Y\}$ such that $I(X; Y|\mathbf{S}) > 0$.

Definition 7 (Irrelevant feature) [32]. For variable X, Y is its irrelevant feature (is irrelevant to X), iff $\forall\ \mathbf{S} \subseteq \mathbf{U} \setminus \{Y\}$, $I(X; Y|\mathbf{S}) = 0$.

Theorem 2 [33]. Under the faithfulness assumption, the spouses of a target variable are included in the strongly relevant features of the target.

Theorem 2 indicates that, in a BN, the spouses maintain a strong relationship with the target that differs from other non-spouses, and the spouses will be identified, if we can capture this strong relationship.

Proposition 1 [34]. The joint mutual information measures information in two random variables X and Y that are shared with another random variable Z, defined as follows:

$$\begin{aligned} I((X,Y);Z) = &I(X;Z) + I(Y;Z) \\ &- I(X;Y) + I(X;Y|Z). \end{aligned} \tag{5}$$

Proposition 2 [34]. The interaction information can be viewed as a measure of the amount of information shared by multiple random variables. A three-way interaction information among X, Y, and Z is defined as follows:

$$\begin{aligned} I(X;Y;Z) &= I(X;Z) - I(X;Z|Y) \\ &= I(Y;Z) - I(Y;Z|X) \\ &= I(X;Y) - I(X;Y|Z). \end{aligned} \tag{6}$$

Propositions 1 and 2 are two expressions in the chain rule of mutual information. Furthermore, we use the chain rules to explore the mutual information form of this strong relationship.

Proposition 3 [33]. If Y is a spouse of T through X, i.e., X is a child of both Y and T, as shown in Fig. 3 (a), $I((X,Y);T)$ provides more information than $I(Y;T)$.

Proposition 3 shows that, although a spouse of T is not a direct cause or a direct effect of T, from the viewpoint of class-conditional relevancy, it still has a strong relationship with the target variable.

4 Our Algorithm

In this section, we first discuss spouses of the target variable and gradually analyze its other causal features in mutual information, and discover the unique property of the

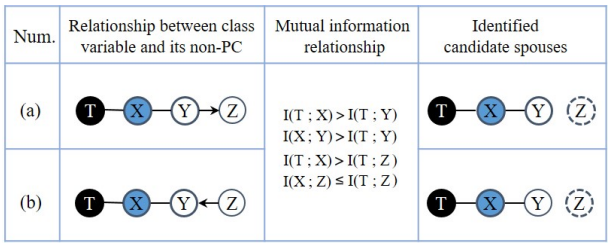


Fig. 2. Two cases of T-X-Y-Z according to the edge direction between Y and Z are shown in (a) and (b), where T is a target variable, X is the PC variable of T, Y is another PC variable of X (except T), and Z is the non-PC variable of T with no edge to X. According to Theorem 3, using the mutual information property between the target variable and its PC of a variable in PC, Z in (a) and (b) will be removed, and the T-X-Y skeleton will be constructed.

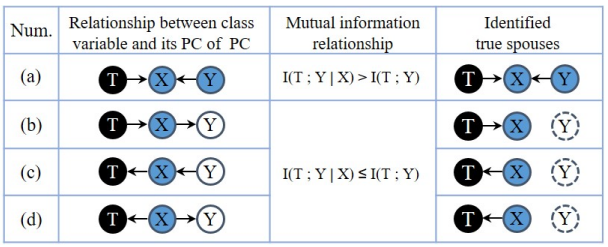


Fig. 3. Four cases of T-X-Y skeleton according to the edge direction between these three variables are shown in (a), (b), (c), and (d), where T is a target variable, X is PC variable of T, and Y is other PC variable of X except T. According to Theorem 4, using the mutual information property between the target variable and its spouses, the true spouse Y in (a) can be discovered and the false spouses Y in (b), (c), and (d) will be removed.

spouses that differs from other non-causal features (non-PC). Furthermore, we describe how to identify this property and introduce the specific implementation details of the proposed algorithm. Finally, we analyze the complexity of the proposed algorithm compared to that of the state-of-the-art algorithms.

4.1 Mutual Information Properties of Causal Features

In this section, we first formulate the process of building a PC skeleton for each variable in PC based on mutual information. Then, we analyze unique mutual information properties between class variable and its causal features that distinguished from other non-causal features.

As shown in Fig. 2, according to the direction between PC of each variable in PC and other non-PC, the structures can be simplified into two forms as presented in Figs. 2 (a) and (b). Both in Figs. 2 (a) and (b), the false PC of each variable in PC, Z, need to be removed to construct the skeleton T-X-Y. Thus, we propose Theorem 3 to illustrate the mutual information relationships between the target variable and its true PC of each variable in PC, to identify the true PC of each variable in PC.

Theorem 3. In a faithful BN, for T, X, and $Y \in \mathbf{U}$, T and Y are not adjacent but connected by X. The mutual information between T and X is larger than that between T and Y (i.e., $I(T;X) > I(T;Y)$), and that between X and Y is larger than that between T and Y (i.e., $I(X;Y) > I(T;Y)$).

Proof: For the true positive PC variable of each variable in PC (e.g., Y in Fig. 2), there is an edge between X and Y, and the correlation between X and Y is extremely close (i.e., $I(X;Y) > 0$). According to Definition 5, Y is a strongly relevant feature of X.

Similarly, the mutual information between T and X is greater than zero (i.e., $I(T;X) > 0$). According to Definition 5, X is a strongly relevant feature of T. Since no edge exists between T and Y, regardless of the direction of the edge between T, X, and Y in Fig. 2, the correlation between Y and T is not as related as the correlation between X and Y or between T and X, thus we can conclude that:

$$I(X;Y) > I(T;Y);\ I(T;X) > I(T;Y). \tag{7}$$

It is clear that, the true PC variable of each variable in PC has the property that $I(T;X) > I(T;Y)$ and $I(X;Y) > I(T;Y)$, using this property, the PC of each variable in PC is identified. ■

For Z in Figs. 2 (a) and (b), since Z does not have a direct edge with X, it cannot satisfy this relationship in Theorem 3. Therefore, we use pairwise comparisons to remove false PC of each variable in PC from non-PCs. Specifically, according to the mutual information relationships, Y will be identified as true PC of each variable in PC, $I(X;Y) > I(T;Y)$ && $I(T;X) > I(T;Y)$. Further, Z will be removed, $I(X;Z) \leq I(T;Z)$ && $I(T;X) \leq I(T;Z)$, to construct the skeleton T-X-Y.

After applying Theorem 3 to exclude false PC of each variable in PC, the true spouses can be identified subsequently. As shown in Fig. 3, the direction between the target variable and its PC of each variable in PC (i.e., the skeleton T-X-Y) has four cases in Figs. 3 (a), (b), (c), and (d), respectively. Among these four cases, only Fig. 3 (a) is the structure that we need to identify and save Y as a spouse. Thus, we propose Theorem 4 to illustrate the unique mutual information property between the target variable and its spouses, which distinguishes them from other non-spouses.

Theorem 4. In a faithful BN, for T, X, and $Y \in \mathbf{U}$, T and Y are not adjacent but are connected by X. X will increase the mutual information between T and Y (i.e., $I(T;Y|X) - I(T,Y) > 0$) if X is a common child of T and Y.

Proof: Given a dataset $\mathcal{D}$ comprising a variable set $\mathbf{U}$ and for a variable $T \in \mathbf{U}$, four structures between T and its PC of each variable within the PC are shown in Fig. 3 (where $Y \in \mathbf{U} \setminus \mathbf{PC}$, and X is one of the PC variables connecting T and Y).

In Fig. 3 (a), Y is a spouse of T with respect to X. According to Theorem 1, no edge exists between T and Y, and $T \perp\!\!\!\perp Y \mid \emptyset$. Under the faithfulness assumption, we can obtain the following equation:

$$P(t,y) = P(t)P(y). \tag{8}$$

Bringing equation (7) into the definition of mutual information (i.e., Eq. (3)), the mutual information between T and Y can be expressed as:

$$\begin{aligned} I(T;Y) &= \Sigma t, y P(t,y) log \frac{P(t,y)}{P(t)P(y)} \\ &= 0. \end{aligned} \tag{9}$$

After obtaining the result of Eq. (8), in Fig. 3 (a), the mutual information between T and Y is maintained

as 0. After the variable X (common child of T and Y) is given, the dependence relationship between T and Y changes according to the d-separation (i.e., $T \not\perp\!\!\!\perp Y \mid X$). According to the chain rule of mutual information (Eq. (6) in Proposition 2), $I(T;X;Y) = I(T;Y) - I(T;Y|X) = I(T;X) - I(T;X|Y) = I(X;Y) - I(X;Y|T)$, and we can conclude that:

$$I(T;Y|X) = I(X;Y|T) - I(X;Y) + I(T;Y). \tag{10}$$

From the above analysis, $I(T;Y) = 0$, and in terms of Eq. (5) in Proposition 1 , we can obtain $I(X;Y|T) - I(X;Y) = I((X,Y);T) - I(X;T)$. According to Proposition 3, $I((X,Y);T)$ provides more information than $I(X;T)$ (i.e., $I((X,Y);T) - I(X;T) > 0$). Therefore, the result of this equation is as follows:

$$\begin{aligned} &I(X;Y|T) - I(X;Y) \\ &= I((X,Y);T) - I(X;T) > 0. \end{aligned} \tag{11}$$

Combining the value of Eq. (10) and $I(T;Y) = 0$, $I(T;Y|X)$ is equivalent to the following form:

$$\begin{aligned} I(T;Y|X) &= I(X;Y|T) - I(X;Y) + I(T;Y) \\ &= I((X,Y);T) - I(X;T) > 0. \end{aligned} \tag{12}$$

In Fig. 3 (a), the conditional mutual information between T and Y changes after given X. Here, T and Y perform a closer relationship.

Combining the results of Eqs. (8) and (11), we obtain the following conclusion: when X is a common child of T and Y, the mutual information after a given X is greater than that given an empty set.

$$I(T;Y|X) > I(T;Y|\emptyset). \tag{13}$$

From Eq. (12), in Fig. 3 (a), it is clear that from the information-theoretic perspective, X provides more information for T and Y as the conditioning set.

Thus far, we have illustrated an unequal relationship for true positive spouses where other situations occur in other features. As shown in Fig. 3, three cases also exist between the target and its PC of each variable within the PC that are (b), (c), and (d), denoting the descendant, ancestor, and sibling node, respectively. All three situations are the same in the d-separation and thus are combined for illustration.

For Definition 2, after a given X, $T \perp\!\!\!\perp Y \mid X$, and combined with the faithfulness assumption, we can obtain $P(t,y|x) = P(t|x)P(y|x)$. According to the definition of mutual information, the mutual information between T and Y given X is as follows:

$$\begin{aligned} &I(T;Y|X) \\ &= \Sigma_{x\in X} P(x) \Sigma_{t\in T, y\in Y} P(t,y|x) log\ 1 = 0. \end{aligned} \tag{14}$$

In line with Proposition 2, $I(T;Y|\emptyset) - I(T;Y|X) = I(T;X;Y)$, therefore $I(T;Y|\emptyset) - I(T;X;Y) = 0$. We can determine from the definition of mutual information, $I(T;X;Y) \geq 0$. For the three cases in Figs. 3 (b), (c), and (d), X is a non-collider, we can conclude that:

$$I(T;Y|\emptyset) \geq I(T;Y|X). \tag{15}$$

Following Eq. (13), independency remains between T and Y after given X. From Eq. (14), in Figs. 3 (b), (c), and (d), from the information-theoretic perspective, X is a conditional set that makes the relationship between T and Y independent, that is, $T \not\perp\!\!\!\perp Y \mid \emptyset$, $T \perp\!\!\!\perp Y \mid X$, and $I(T;Y|X) \leq I(T;Y|\emptyset)$. In these three cases, the performance of mutual information is entirely different from that of true spouses.

Therefore, we analyze the mutual information between the target variable and its true positive spouses from causal features, and analyze the mutual information of the corresponding PC in an interspersed pattern. Finally, we conclude that the mutual information between the target variable and its spouses, given a common child, is greater than the given empty set. ■

In summary, Theorem 3 first removes the false PC of each variable in PC to construct the skeleton, and then Theorem 4 is used to identify the spouses from this skeleton. Based on Theorems 3 and 4, we use pairwise comparisons to replace enumerating conditional sets to identify causal features. The following section introduces our algorithm in detail.

4.2 Algorithm Implementation

Based on the above analysis, we demonstrated the theoretical correctness and feasibility of our concept. Thus, we propose a light algorithm, CFS-MI, that separately instantiates the pairwise comparison of mutual information in the PC and spouses discovery, thereby identifying the causal features without enumerating the conditioning sets. As shown in Algorithm 1, CFS-MI has two phases: Phase 1 (lines 2–15) uses the normalized form of mutual information to discover the PC, whereas Phase 2 (lines 17–26) uses mutual information to discover spouses. The specific phases of the CFS-MI algorithm are as follows:

Phase 1 (lines 2–15). It obtains all PC of the target variable. We first calculate SU (a normalized form of mutual information) between the target variable and all other variables(i.e., $SU(X;T)$, $X \in \mathbf{U} \setminus \{T\}$), and place them in descending order (line 5). Then, we remove the variables where the values of SU are below the predefined δ (SU of true positive PC is greater than δ). According to the property in which the mutual information between the PC of the target and their PC (except the target variable) is greater than that between the target variable and the PC of each variable within the PC (i.e., $SU(X;Y) > SU(T;Y)$), we compare these two numbers to identify the true positive PC variables (line 10).

Phase 2 (lines 17–26). It traverses the set of candidate spouses obtained in Phase 1. First, according to Theorem 3 (line 19), we confirm whether the SU between A and B is larger than that between T and B, and whether the SU between T and A is also larger than that between T and B (i.e., $SU(A;B) > SU(T;B)$ && $SU(T;A) > SU(T;B)$). Thus, we remove many false PC of each variable in PC in this manner. If the condition is satisfied, we continue calculating the mutual information relationships between T and B given A (the corresponding PC) and an empty set. If the former is larger, according to Theorem 4 (line 20), we can assert that B is a spouse of T with respect to A. Combining the results of Phases 2 and 1, we finally obtain the MB of the target variable.

Algorithm 1: CFS-MI

Input: T: target; $\mathcal{D}$: dataset ; δ: predefined threshold
Output: [$\mathbf{PC}_T$, $\mathbf{SP}_T$]: Markov blanket of T

/*$Phase\ 1\ :\ PC\ discovery\ in\ mutual\ information$*/
$\mathbf{S} = \emptyset$;
for each $X \in \mathbf{U} \setminus \{T\}$ **do**
  if $SU(X;T) > \delta$ **then**
    $\mathbf{S} = \mathbf{S} \cup \{X\}\ in\ descending\ order$;
  end
end
for $X \in \mathbf{S}$ **do**
  for $Y \in \mathbf{S} \setminus \{X\}$ **do**
    if $SU(X;Y) > SU(Y;T)$ **then**
      $\mathbf{S} = \mathbf{S} \setminus \{Y\}$;
    end
  end
end
$\mathbf{PC}_T = \mathbf{S}$;
/*$Phase\ 2\ :\ SP\ discovery\ in\ mutual\ information$*/
for each $A \in \mathbf{PC}_T$ **do**
  for each $B \in \mathbf{U} \setminus \mathbf{S}$ **do**
    if $SU(A;B) > SU(T;B)\quad \&\&\quad SU(T;A) > SU(T;B)$ **then**
      if $I(T;B|A) > I(T;B|\emptyset)$ **then**
        $\mathbf{SP}_T = \mathbf{SP}_T \cup \{B\}$;
      end
    end
  end
end
Return [$\mathbf{PC}_T,\ \mathbf{SP}_T$];

4.3 Algorithm Complexity Analysis

This section conducts a comprehensive evaluation of the complexity of the algorithm, including analyzing computational and space complexities.

4.3.1 Computational Complexity

We now analyze the computational complexity of CFS-MI before conducting an empirical study. The main time overhead is computing the relationship between two variables in the causal feature selection algorithms. Thus, we use the computations of mutual information as meta-computations to analyze the computational complexity of the proposed CFS-MI algorithm.

For the proposed CFS-MI, in the PC discovery phase, CFS-MI first adds all variables that their mutual information with the target variable are larger than the threshold to **S**. Then, it traverses **S** and identifies true PC variables by pairwise comparisons. During this process, traversing **S** requires two layers: the outer layer aims to identify true PC variables, $|\mathbf{PC}|$ times, and the inner layer aims to remove false PC variables, $|\mathbf{U}| - |\mathbf{PC}|$ times. Moreover, the mutual information (*SU* is the normalized form of mutual information) is only calculated twice and compared with each other. Thus, the computational complexity of the PC discovery phase can be expressed as O($2(|\mathbf{U}| - |\mathbf{PC}|)|\mathbf{PC}|$)=O($|\mathbf{U}||\mathbf{PC}|$). In the spouses discovery phase, CFS-MI identifies the spouses of the target variable from non-PC variables by pairwise comparisons given different conditioning sets. Identifying the spouses also requires two layers during this process. The out layer needs to traverse the PC variables, $|\mathbf{PC}|$ times. Meanwhile, the inner layer needs to traverse the non-PC variables, and the specific mutual information calculation times are separated between a variable in PC and that in non-PC, $|\mathbf{U}|$ times, between the target variable and a variable in non-PC, $|\mathbf{U}|$ times, between the target variable and a variable in PC, $|\mathbf{U}|$ times, and between the target variable and a variable in non-PC given a corresponding PC variable, $|\mathbf{U}|$ times. Thus, the computational complexity of the spouses discovery phase can be expressed as O($|\mathbf{PC}|(4|\mathbf{U}|)$)=O($|\mathbf{PC}||\mathbf{U}|$). In summary, the overall computational complexity of the CFS-MI algorithm is O($|\mathbf{PC}||\mathbf{U}| + |\mathbf{U}||\mathbf{PC}|$) = O($|\mathbf{U}||\mathbf{PC}|$).

TABLE 2
Computational complexity of CFS-MI and comparison algorithms

Algorithm	Computational Complexity
MMMB	O($\|\mathbf{U}\|\|\mathbf{PC}\|2^{\|\mathbf{PC}\|}$)
HITON-MB	O($\|\mathbf{U}\|\|\mathbf{PC}\|2^{\|\mathbf{PC}\|}$)
PCMB	O($\|\mathbf{U}\|\|\mathbf{PC}\|^2 2^{\|\mathbf{PC}\|}$)
IPCMB	O($\|\mathbf{U}\|\|\mathbf{PC}\|2^{\|\mathbf{U}\|}$)
STMB	O($\|\mathbf{U}\|2^{\|\mathbf{U}\|}$)
BAMB	O($\|\mathbf{U}\|2^{\|\mathbf{PC}\|}$)
EEMB	O($\|\mathbf{U}\|2^{\|\mathbf{PC}\|}$)
CFS-MI	O($\|\mathbf{U}\|\|\mathbf{PC}\|$)

TABLE 3
Space complexity of CFS-MI and comparison algorithms

Algorithm	Space Complexity
MMMB	O($\|\mathbf{PC}\| + \|\mathbf{Sep}\|\|\mathbf{U}\| + \|\mathbf{PC}\|^2 + \|\mathbf{MB}\|$)
HITON-MB	O($\|\mathbf{PC}\| + \|\mathbf{Sep}\|\|\mathbf{U}\| + \|\mathbf{PC}\|^2 + \|\mathbf{MB}\|$)
PCMB	O($\|\mathbf{PC}\| + \|\mathbf{Sep}\|\|\mathbf{U}\| + \|\mathbf{PC}\|^2 + \|\mathbf{MB}\|$)
IPCMB	O($\|\mathbf{PC}\| + \|\mathbf{Sep}\|\|\mathbf{U}\| + \|\mathbf{PC}\|^2 + \|\mathbf{MB}\|$)
STMB	O($\|\mathbf{PC}\| + \|\mathbf{Sep}\|\|\mathbf{U}\| + \|\mathbf{MB}\|$)
BAMB	O($\|\mathbf{PC}\| + \|\mathbf{Sep}\|\|\mathbf{U}\| + \|\mathbf{MB}\|$)
EEMB	O($\|\mathbf{PC}\| + \|\mathbf{Sep}\|\|\mathbf{U}\| + \|\mathbf{MB}\|$)
CFS-MI	O($\|\mathbf{U}\| + \|\mathbf{MB}\|$)

Table 2 presents the computational complexity of seven state-of-the-art causal feature selection algorithms and CFS-MI. We can observe that CFS-MI has the optimal computational complexity, whereas the other divide-and-conquer algorithms generally have an exponential computational complexity due to enumerating conditioning sets.

4.3.2 Space Complexity

In this section, we analyze the space complexity of the proposed algorithm. The space complexity analysis will take the maximum temporary memory space saved by the running algorithms as the unit space.

For the proposed algorithm CFS-MI, the size of $|\mathbf{U}|$ memory space is first needed to store the mutual information between the target and other variables. Then, that of $|\mathbf{PC}|$ memory space is required to store the true PC identified through pairwise comparisons in the PC discovery phase. Thus, the space complexity of CFS-MI in the PC discovery phase can be expressed as O($|\mathbf{U}| + |\mathbf{PC}|$). In the spouses discovery phase, CFS-MI identifies all true spouses for each variable in PC from non-PCs. During this process, the first round of pairwise comparisons aims to build the skeleton of

PC for each variable in the PC, and new memory space is not required. The second round of pairwise comparisons identifies the true spouses, which needs to calculate the mutual information and compare with each other, and new memory space is also not required. After two rounds of comparisons, when a non-PC variable is identified as a spouse, we add it to the spouse set and store it temporarily, and the size of $|\mathbf{SP}|$ memory space is needed. Thus, the space complexity of CFS-MI in the spouse discovery phase can be expressed as O($|\mathbf{SP}|$). In summary, the totally space complexity of CFS-MI can be expressed as O($|\mathbf{U}| + |\mathbf{PC}| + |\mathbf{SP}|$)= O($|\mathbf{U}| + |\mathbf{MB}|$).

Table 3 shows the space complexity of CFS-MI and its comparison algorithms, $|\mathbf{Sep}|$ is the size of memory space that need to save the separating sets of the target variable and other variables. From Table 3, we can observe that CFS-MI has evident advantages over other algorithms regarding space complexity due to it does not save the separating sets.

5 Experiments

To validate the accuracy and efficiency of the CFS-MI algorithm, we compared it with seven state-of-the-art algorithms on five benchmark BNs and sixteen real-world datasets. They are HITON-MB [25], MMMB [24], PCMB [7], IPCMB [26], STMB [27], BAMB [8], and EEMB [28]. The existing MATLAB package Causal Learner[1] has implemented all comparison algorithms, and thus we used MATLAB to implement CFS-MI and compared it with other comparison algorithms in this package [35]. In CFS-MI, the value of δ is predefined and taken as 0.05, in Section 5.3, we have analyzed this value. All experiments were conducted on Windows 10 running on a computer with an Intel Core i7-4790 CPU and 16 GB of RAM.

5.1 Benchmark BN Datasets

Since causal feature selection is to discover the MB of a class variable, and MB is a critical substructure of BN representing local causal relationships, the algorithms are first applied on the datasets generated based on benchmark BNs for MB discovery. The structures of these benchmark BNs are known, and the MB of a node in BN is also known as a substructure. Thus, we compared the MB selected by the algorithms with the known substructure MB of BN to evaluate the performance of the algorithms. We used two groups of data from the five benchmark BNs [36] as shown in Table 4. The two groups include ten datasets each with 500 and 1000 data instances. The conditional independence test that the other seven rivals use the G^2 test at the 0.01 significance level. The detailed results are shown in Table 1 in the Appendix, and best results are highlighted in bold face in the tables.

In the BN experiments, we evaluate the algorithms using the accuracy and efficiency. For accuracy, since the purpose of the causal feature selection algorithms aim to discover the MB of each variable on the BN datasets, we use the known substructure MB of BN as evaluation criteria. We can identify whether the variable is a true positive, false positive, true negative or false negative type. Thus, we can obtain

1. The codes of CFS-MI and all comparison algorithms in Causal Learner are available at http://bigdata.ahu.edu.cn/causal-learner.

TABLE 4
Summary of benchmark BN datasets

Network	Num. Vars	Num. Edges	Max In/Out-Degree	Min/Max |PCset|	Domain Range
Mildew	35	46	3/3	1/5	3-100
Barley	48	84	4/5	1/8	2-67
Pigs	441	592	2/39	1/41	3-3
Link	724	1125	3/14	0/17	2-4
Gene	801	972	4/1	0/11	3-5

the metric of Precision, Recall and the their comprehensive form, F1. For efficiency, we evaluate the algorithms using the critical metric, running time. The details of the metrics are described as follows:

- Accuracy. F1=2 * precision * recall / (precision + recall). Precision denotes the number of true positives in the output (i.e., the features in the output of an algorithm belonging to the true MB of a given target in a test DAG) divided by the total number of features in the output of the algorithm. Recall represents the number of true positives in the output divided by the number of true positives (the number of true MB of a given target) in a test DAG. The F1 score is the harmonic average of precision and recall, where F1 = 1 is the best case (perfect precision and recall) and F1 = 0 is the worst case.
- Efficiency. We measure the efficiency of an algorithm using runtime. Runtime is the average time in seconds to discover the MB for each variable in BN.

For each algorithm, we report the average F1, precision, recall, and runtime. In Table 1 from the Appendix, the results represent each network for different sample sizes and are shown in the format of $A \pm B$, where A represents the average F1, precision, recall, or runtime, and B is the standard deviation.

CFS-MI achieved the best time efficiency on 5/5 datasets, regardless of whether the dataset size was 500 or 1000. On the Mildew dataset, CFS-MI is the most accurate in terms of the F1 metric and runtime. On the Barley dataset, the accuracy of CFS-MI is second only to EEMB, and the gap is controlled within 0.02, but CFS-MI is much faster (over 48 times faster) than EEMB. PCMB cannot obtain the correct PC on the first two datasets, and its accuracy is zero, which can be considered invalid data. On the Pigs and Gene datasets, CFS-MI is also the fastest. Regarding accuracy, as these two datasets are relatively regular, PCMB and IPCMB can obtain the best result because of their symmetry check. Except for them, CFS-MI remains the most accurate algorithm. On the Link dataset, CFS-MI is the most accurate algorithm for 500 data samples and has accuracy comparable to that of HITON-MB, MMMB, and IPCMB. The number of edges of the Link network is over 1000; thus, the relationship between variables is the most complicated, as both PCMB and BAMB ran for over 24 h on the Link dataset, and we use '-' to present this invalid result.

It would be interesting to compare the two algorithms in terms of both time efficiency and accuracy simultaneously as one algorithm may choose to sacrifice structure learning

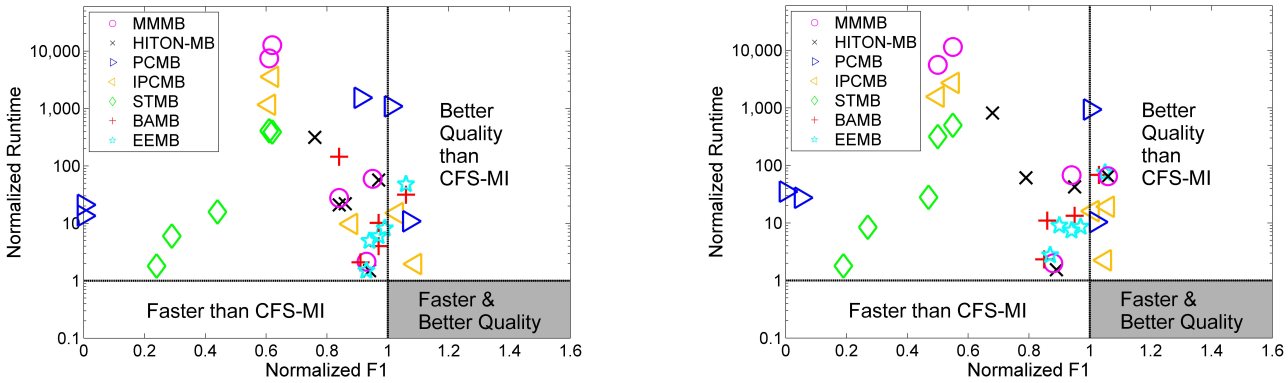


Fig. 4. Normalized runtime versus normalized F1 with 500 (left) and 1000 (right) data samples on benchmark BNs.

quality to enhance computational efficiency, while another algorithm may sacrifice time efficiency for the possibility of increasing structure learning quality. Therefore, to show the experimental results more intuitively, according to data instances, we have normalized the logarithm of F1 and running time in two figures, respectively (i.e., left and right of Fig. 4). Fig. 4 illustrates the results of the normalized time versus the normalized F1 for the two groups of datasets with sample sizes of 500 and 1000, respectively. Table 2 from the Appendix details the normalized running time and F1. Normalized F1 or running time is the value of F1 or running time of each algorithm for a particular data sample size and network divided by the F1 or running time of CFS-MI on the same sample size and network. In Table 2 from the Appendix, a normalized running time over 1.00 implies that an algorithm is slower than CFS-MI, while a normalized F1 below 1.00 implies that an algorithm is worse than CFS-MI.

Therefore, the values corresponding to CFS-MI in Table 2 from the Appendix are all 1.00. After the normalized calculation, the results of other algorithms are also shown. Considering F1, we have achieved better results on most datasets, only a few values are greater than 1.00, but all are less than 1.10. In terms of time, we can observe that the values of all other algorithms after normalization are greater than 1.00, the minimum is 1.50, and the maximum is 12681.00, demonstrating the excellent performance of our algorithm in time consumption.

Each point in Fig. 4 denotes the performance in terms of the two metrics of a given algorithm on learning one of the five benchmark BNs and one-to-one correspondence with each item in normalized Table 2 from the Appendix. Since we use a normalized method to evaluate the CFS-MI, the performance of CFS-MI always falls on point (1,1), and is therefore not indicated in the figures. Two horizontal and vertical lines in Fig. 4 represent the basis of CFS-MI in time consumption and F1.

As shown in Fig. 4, there are no points in the gray area, indicating that no algorithms outperform CFS-MI in terms of both running time and F1. In terms of F1, CFS-MI is extremely competitive with the other algorithms in most cases, and no algorithm is faster than CFS-MI. PCMB cannot obtain the correct result on the Mildew and Barley datasets, and thus two points fall in the position where F1 equals to 0.0. For most datasets, the other algorithms are slower and have lower accuracy. Although there are only a few points on the right of CFS-MI in Fig. 4, all of them are controlled below a factor of 1.10 compared to the results of CFS-MI, indicating that the accuracy of CFS-MI is comparable.

In summary, it can be seen that, compared to the other algorithms, the accuracy of the CFS-MI algorithm is competitive or at worst, the error margins remain within 10% of each other. However, the time consumption of CFS-MI is less than that of all the above algorithms. This is because we use pair comparison of mutual information to conduct MB discovery instead of enumerating the conditioning sets, and the original exponential computational complexity can be reduced to quadratic complexity. The results on these five BN datasets illustrate that CFS-MI is both time-efficient and accurate and is a light algorithm that outperforms all other comparison algorithms.

TABLE 5
Summary of real-world datasets

Dataset	Num.Features	Num.Samples
Wine	13	178
Congress	16	435
Soy	35	47
Krvskp	36	3186
Splice	60	3175
Seme	256	1593
Hiva	1617	3845
Colon	2000	62
Ovarian	2190	216
Lymph	4062	96
Gisette	5000	6000
Leukemia	7129	72
Lungcancer	12533	181
Breastcancer	17816	286
Dexter	20000	300
Dorothea	100000	800

5.2 Real-world Datasets

Since the MB of the class variable is the optimal and minimal feature subset with maximum predictivity for classification, causal feature selection uses the MB of a class variable

to perform causality-based feature selection. Moreover, the real-world classification datasets are used for classification tasks and do not have corresponding BN structures, and the MB of the class variable is unknown in advance. Thus, we apply the algorithms on the real-world datasets to identify the MB of class variable as causal features and use these features combined with classifiers for classification to evaluate the performance of the algorithms further. In this section, we test the algorithms on 16 real-world datasets in Table 5 with dimensions ranging from low to high. The Wine, Congress, Soy, Krvskp, Splice, Seme, Lymph, Gisette, Breastcancer, Dexter, and Dorothea are UCI machine learning databases [37], the Hiva is a WCCI2006 functional prediction challenge dataset, the Leukemia and Lungcancer are two frequently studied public microarray datasets [38], and the Colon, and Ovarian [39] are biological datasets.

In the real-world experiments, we also evaluate the algorithms using accuracy and efficiency. For accuracy, the MB of a class variable cannot be known in advance, thus, we use the selected MB as the causal features combined with two classifiers, KNN and SVM, for classification and use KNN and SVM to mark the classification accuracy as the evaluation metrics respectively. Moreover, fewer features selected can better reflect the result of causal feature selection with comparable accuracy; thus, an additional accuracy metric Compactness is used to mark the number of selected features. For efficiency, we also report the running time of algorithms, and Tests is used to mark the number of tests to further illustrate the efficiency of the algorithms. The details of the metrics are described as follows:

- Accuracy. Compactness is the number of selected features. Prediction accuracy is the percentage of the correctly classified test instances that were previously unseen. We report both compactness and prediction accuracy of the KNN and SVM classifiers as the accuracy measures of the various algorithms.
- Efficiency. We report conditional independence tests and runtime as the efficiency metric of the different algorithms. Since the meta-computations of CFS-MI are calculating the mutual information that varies from its rivals, we use tests to unified evaluate of all algorithms. Runtime is the average time in seconds to find the MB for class variable.

In Table 3 from the Appendix, we summarize the feature selection results by the different algorithms. The results are shown in the format of $A \pm B$, where A represents the accuracy, compactness, or runtime, and B is the corresponding standard deviation.

The experimental results show that the prediction accuracy of CFS-MI is better or comparable that of other algorithms using KNN and SVM. The lower accuracy of CFS-MI is generally 2% lower than that of the best algorithm and no more than 6% lower. Regarding time efficiency, CFS-MI is much faster (most results have a tenfold increase in time) than other algorithms on five datasets. On the Wine and Soy datasets that are low-dimensional, but large sample-to-feature ratios, among all algorithms CFS-MI selects the minimum number of features, but has comparable or even higher accuracy and greatly improves the efficiency. With the increasing of the data dimension, our algorithm significantly reduces the time consumption while ensuring the accuracy. On some datasets with small sample-to-feature ratios such as Ovarian and Leukemia, their number of features is much larger than the number of samples, the accuracy of our algorithm decreases slightly. A possible explanation is that because the probability distributions of the features that need to be counted for calculating mutual information cannot comprehensively summarize the state of features, and CFS-MI using pairwise comparisons typically selects more causal features than other algorithms. Among them, some bad impact features on the classification accuracy may be included in the result of CFS-MI, leading to accuracy fluctuations. However, for most high-dimensional datasets, the state-of-the-art algorithms cannot obtain results due to the exponential computational complexity.

On the Seme dataset, we take the experiment of HITON-MB as an example. Although the dimension of the Seme dataset is low, HITON-MB still selects 111 features in the PC discovery phase. Thus, when HITON-MB identifies whether a feature is the PC of the target variable, it needs to enumerate all subsets of these 111 features. In the spouses discovery phase, it is necessary to continue to discover the PC of each variable in these 111 features, which directly leads to the inability of HITON-MB to obtain experimental results in a reasonable time. Moreover, the other comparison algorithms also need to enumerate conditioning sets the same as HITON-MB, and quite many selected features lead to an exponential increase in time consumption. Consequently, they cannot obtain valid experimental results. Regarding the Hiva dataset, although its dimension is high, HITON-MB only selects seven features, which is completely acceptable. Thus, the comparison algorithms can obtain the results within a reasonable time range. On the Gisette and Dorothea datasets that are extremely high-dimensional, existing algorithms cannot obtain valid experimental results as quite many features are selected. The proposed CFS-MI, no matter how many features it selects, only consumes the quadratic time of several features, which is completely acceptable. Thus, we can achieve the best time consumption on all datasets. Similar to the results in BNs, we use '-' to represent the invalid results. On the Lungcancer dataset, CFS-MI is better than MMMB, HITON-MB, PCMB, IPCMB, STMB, BAMB, and EEMB on most datasets using both KNN and SVM.

On the Congress, Krvskp, and Splice datasets with low dimensions, all algorithms can get experimental results within a reasonable time range. CFS-MI has the highest classification accuracy using KNN and SVM and selects the minimum number of features among all algorithms. The running time of CFS-MI is also the lowest due to pairwise comparisons. Meanwhile, on high-dimensional datasets of Colon, Lymph, and Breastcancer, other algorithms have selected quite many features, making it difficult to obtain valid results. However, the number of selected features will not limit the algorithm. Thus, our algorithm can achieve effective experimental results on these high-dimensional datasets. On the Dexter dataset, the prediction accuracy of CFS-MI is better than other algorithms, except BAMB when using KNN, with a gap of only 0.1. Furthermore, in addition to STMB and BAMB, the prediction accuracy of CFS-MI outperforms that of the other algorithms when using SVM.

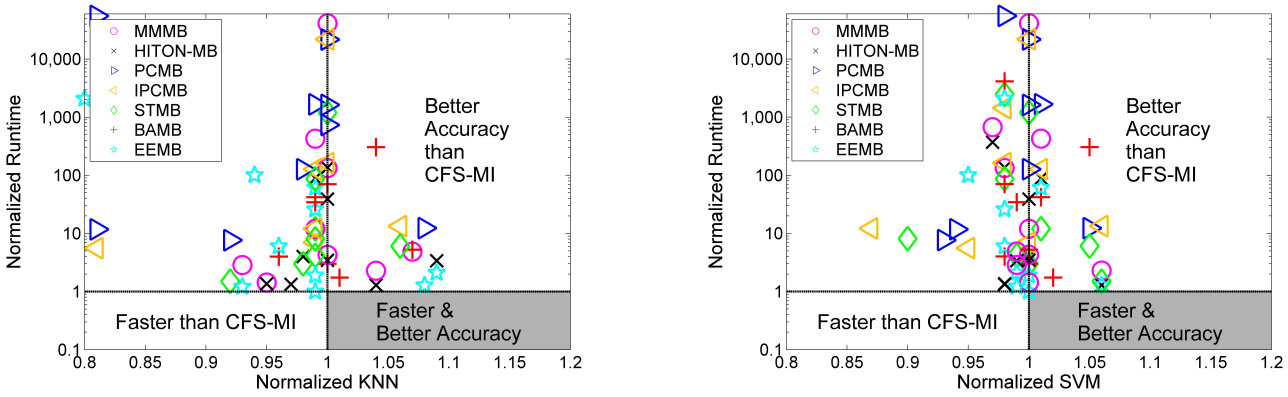


Fig. 5. Normalized runtime vs. normalized KNN (left) and SVM (right) on 16 real-world datasets.

Regardless of the datasets, CFS-MI has the lowest time consumption and comparable prediction accuracy. Regarding tests, the number of the Compactness affects the size of CIT. For the same number of the Compactness, the mutual information calculation performed by CFS-MI is the least.

The same as the results on BN datasets, we still separately normalize the logarithms of two metrics. The results on these 16 real-world datasets are presented in Table 4 from the Appendix, Fig. 5, respectively. Therefore, the values corresponding to CFS-MI in Table 4 from the Appendix are all 1.00. After the normalized calculation, the results of other algorithms are shown in Table 4 from the Appendix as well. Regarding KNN and SVM, we have achieved better results on most datasets, only a few values are greater than 1.00, but all are below 1.10. We have achieved results similar to those of other algorithms, and as datasets become larger, we can perform better for high-dimensional data. In terms of time, the values of all other algorithms after normalization are greater than 1.00, the minimum is 1.50, and the maximum is 21915.00, which shows the excellent performance of our algorithm in time.

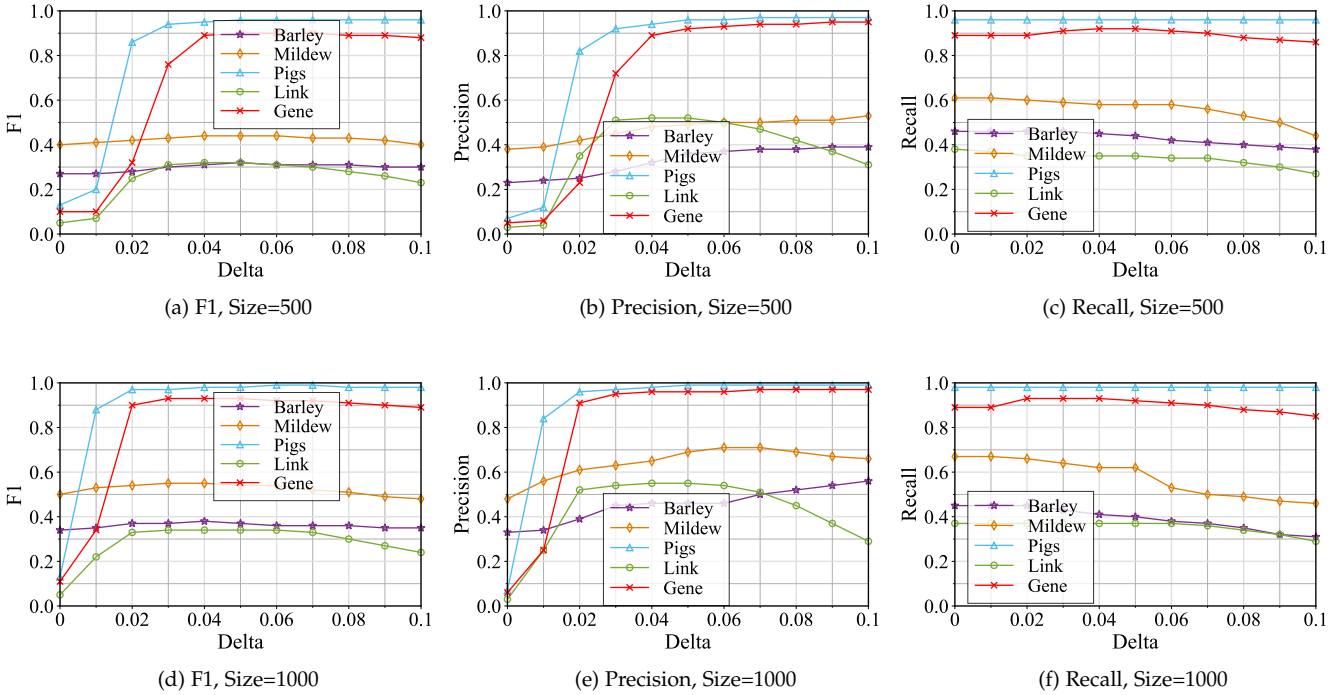
As shown in Fig. 5, in nearly 70 experimental results, no other algorithms are faster than the CFS-MI algorithm, illustrating the lower time consumption of CFS-MI in comparison with the other approaches. In terms of classification accuracy, although the results of some algorithms achieved better results than those of CFS-MI, the error margins remain within 10% of each other; that is, the accuracy of CFS-MI is acceptable and comparable. Regarding time efficiency, we can observe that the points that fall in about 10 times are the most numerous, and the highest point even exceeds 10^4 times. Many points also fall in places around 10^2 times, enumerating the conditioning sets caused by conditional independence tests being the key reason. This further illustrates the time efficiency of CFS-MI in time. No points fall in the gray area in both the left and right one of Fig. 5, indicating that no algorithms can perform better in terms of both time and accuracy than CFS-MI. Based on a symmetry check, PCMB requires over 10^4 times the time required to obtain the same accuracy as CFS-MI on the Ovarian dataset.

In summary, on 16 real-world datasets, CFS-MI can achieve the lowest time consumption and maintain a higher classification accuracy on KNN and SVM compared to other algorithms. After testing and comparing it on benchmark BNs and real-world datasets, we can conclude that CFS-MI performs well on various datasets owing to its pairwise comparison of mutual information and is a light algorithm capable of high-dimensional datasets.

5.3 Analysis of Hyperparameter δ for CFS-MI

In this section, we will discuss the influence of hyperparameter δ on the experimental performance of CFS-MI. We set δ to gradually increase from 0 to 0.1 on five BN datasets with data sizes of 500 and 1000. The influence of different δ values of CFS-MI on F1, Recall, and Precision (for the specific meaning, please refer to the accuracy metrics of Section 5.1) is recorded and shown in Fig. 6.

In Figs. 6 (a) and (d), we can observe that when the value of δ is close to 0, the value of F1 is lower. As the value of δ increases, F1 gradually increases and becomes more stable. In this regard, our explanation is that with the increase of value of δ within [0, 0.04], the false positive variables are removed, and the Precision of CFS-MI increases rapidly, leading to the CFS-MI being sensitive to the value of δ, as shown in Figs. 6 (b) and (e). When the value of δ is within [0.04, 0.08], the value of F1 stabilizes as the value of Precision stabilizes. Figs. 6 (c) and (f) also illustrate that the value of δ has little effect on the truly correct output of CFS-MI in this interval. Given that the target variable is much more closely related to its MB variables than the other non-MB variables, even if the value of δ becomes larger, they are not easily removed. Thus, CFS-MI is insensitive to the value of δ within [0.04, 0.08]. Only when the value of δ increases to greater than 0.08 some true positive variables will be removed to reduce the Recall of CFS-MI. As shown in Fig. 6 (c) and (f), the value of Recall of CFS-MI reducing leads to the fact that CFS-MI is again sensitive to the value of δ within [0.08, 0.1]. In summary, CFS-MI is insensitive to the value of δ within the range of [0.04, 0.08], and the performance of our algorithm tends to be stable. Within this interval, the value of F1 on each dataset almost reaches the highest value when that of δ equals 0.05. Thus, we choose 0.05 as the value of δ in the experiments.



(a) F1, Size=500 (b) Precision, Size=500 (c) Recall, Size=500

(d) F1, Size=1000 (e) Precision, Size=1000 (f) Recall, Size=1000

Fig. 6. F1, Precision, and Recall of CFS-MI on five BNs when varying the value of δ from 0 to 0.1.

6 Conclusion

In this paper, we analyzed the unique mutual information relationships between the class variable and its causal features that vary from others. Then, we proposed a light causal feature selection algorithm, CFS-MI, that uses pairwise comparisons of mutual information to reduce computational complexity from exponential to quadratic. Finally, extensive experiments on 5 benchmark BNs and 16 real-world datasets validated that CFS-MI has the lowest time consumption and high accuracy within error and is capable of high-dimensional data. However, the mutual information calculation that is used in CFS-MI can only handle discrete values, and continuous values need to be discretized in advance. In this process, the discretization method will also affect the performance of the algorithm. Thus, future research could focus on how to directly applying the mutual information calculation to continuous data and further mixed data.

Acknowledgments

This work was supported by the National Science Foundation of China [No. U1936220, 62272001, 61872002, 62176001, and 61806003], the National Key Research and Development Program of China [No. 2021ZD0111801 and 2020AAA0106100], and the Natural Science Foundation of Anhui Province of China [No. 2108085QF270].

References

[1] D. Koller and M. Sahami, "Toward optimal feature selection," Stanford InfoLab, Tech. Rep., 1996.

[2] I. Guyon, C. Aliferis *et al.*, "Causal feature selection," in *Computational methods of feature selection*. Chapman and Hall/CRC, 2007, pp. 75–97.

[3] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[4] 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," *Journal of Machine Learning Research*, vol. 11, no. Jan, pp. 171–234, 2010.

[5] ——, "Local causal and markov blanket induction for causal discovery and feature selection for classification part ii: Analysis and extensions," *Journal of Machine Learning Research*, vol. 11, no. Jan, pp. 235–284, 2010.

[6] K. Yu, X. Guo, L. Liu, J. Li, H. Wang, Z. Ling, and X. Wu, "Causality-based feature selection: Methods and evaluations," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–36, 2020.

[7] J. M. Pena, R. Nilsson, J. Björkegren, and J. Tegnér, "Towards scalable and data efficient learning of markov boundaries," *International Journal of Approximate Reasoning*, vol. 45, no. 2, pp. 211–232, 2007.

[8] Z. Ling, K. Yu, H. Wang, L. Liu, W. Ding, and X. Wu, "Bamb: A balanced markov blanket discovery approach to feature selection," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 5, pp. 1–25, 2019.

[9] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[10] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of machine learning research*, vol. 5, no. Oct, pp. 1205–1224, 2004.

[11] S. L. Salzberg, "C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993," 1994.

[12] N. X. Vinh, S. Zhou, J. Chan, and J. Bailey, "Can high-order dependencies improve mutual information based feature selection?" *Pattern Recognition*, vol. 53, pp. 46–58, 2016.

[13] P. E. Meyer, C. Schretter, and G. Bontempi, "Information-theoretic feature selection in microarray data using variable complementarity," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 261–274, 2008.

[14] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[15] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM computing surveys (CSUR)*, vol. 50, no. 6, pp. 1–45, 2017.

[16] F. Zhang, J. Zhai, X. Shen, O. Mutlu, and X. Du, "Poclib: A high-

performance framework for enabling near orthogonal processing on compression," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 2, pp. 459–475, 2022.
[17] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text classification using machine learning techniques." *WSEAS transactions on computers*, vol. 4, no. 8, pp. 966–974, 2005.
[18] X. Chang, F. Nie, S. Wang, Y. Yang, X. Zhou, and C. Zhang, "Compound rank- k projections for bilinear analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 7, pp. 1502–1513, 2016.
[19] C. Yan, X. Chang, M. Luo, Q. Zheng, X. Zhang, Z. Li, and F. Nie, "Self-weighted robust lda for multiclass classification with edge classes," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 1, pp. 1–19, 2020.
[20] I. K. Fodor, "A survey of dimension reduction techniques," Lawrence Livermore National Lab., CA (US), Tech. Rep., 2002.
[21] D. Margaritis and S. Thrun, "Bayesian network induction via local neighborhoods," in *Advances in neural information processing systems*, 2000, pp. 505–511.
[22] I. Tsamardinos, C. F. Aliferis, A. R. Statnikov, and E. Statnikov, "Algorithms for large scale markov blanket discovery." in *FLAIRS conference*, vol. 2, 2003, pp. 376–380.
[23] S. Yaramakala and D. Margaritis, "Speculative markov blanket discovery for optimal feature selection," in *Fifth IEEE International Conference on Data Mining (ICDM'05)*. IEEE, 2005, pp. 809–812.
[24] I. Tsamardinos, C. F. Aliferis, and A. Statnikov, "Time and sample efficient discovery of markov blankets and direct causal relations," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 673–678.
[25] C. F. Aliferis, I. Tsamardinos, and A. Statnikov, "Hiton: a novel markov blanket algorithm for optimal variable selection," in *AMIA annual symposium proceedings*, vol. 2003. American Medical Informatics Association, 2003, pp. 21–25.
[26] S. Fu and M. C. Desmarais, "Fast markov blanket discovery algorithm via local learning within single pass," in *Conference of the Canadian Society for Computational Studies of Intelligence*. Springer, 2008, pp. 96–107.
[27] T. Gao and Q. Ji, "Efficient markov blanket discovery and its application," *IEEE transactions on cybernetics*, vol. 47, no. 5, pp. 1169–1179, 2017.
[28] H. Wang, Z. Ling, K. Yu, and X. Wu, "Towards efficient and effective discovery of markov blankets for feature selection," *Information Sciences*, vol. 509, pp. 227–242, 2020.
[29] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt, "Feature selection via dependence maximization." *Journal of Machine Learning Research*, vol. 13, no. 5, 2012.
[30] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000.
[31] D. Koller, N. Friedman, and F. Bach, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
[32] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.
[33] K. Yu, L. Liu, and J. Li, "A unified view of causal and non-causal feature selection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 4, pp. 1–46, 2021.
[34] B. V. Bonnlander and A. S. Weigend, "Selecting input variables using mutual information and nonparametric density estimation," in *Proceedings of the 1994 International Symposium on Artificial Neural Networks (ISANN94)*. Citeseer, 1994, pp. 42–50.
[35] Z. Ling, K. Yu, Y. Zhang, L. Liu, and J. Li, "Causal learner: A toolbox for causal structure and markov blanket learning," *arXiv preprint arXiv:2103.06544*, 2021.
[36] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Machine learning*, vol. 65, no. 1, pp. 31–78, 2006.
[37] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml
[38] L. Yu, C. Ding, and S. Loscalzo, "Stable feature selection via dense feature groups," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 803–811.
[39] B. Hitt and P. Levine, "Multiple high-resolution serum proteomic features for ovarian cancer detection," Mar. 23 2006, uS Patent App. 11/093,018.

Zhaolong Ling received the the PhD degree in the School of of Computer and Information at the Hefei University of Technology, China, in 2020.

He is a Lecturer with the School of Computer Science and Technology, Anhui University, China. His research interests include feature selection, casual discovery, and data mining.

Ying Li received the B.S. degree from Qufu Normal University, China, in 2020.

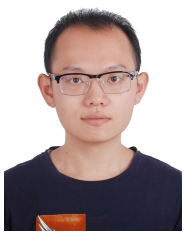
Currently, he is working toward the MSc degree in the School of Computer Science and Technology, Anhui University, China. His research interests include feature selection, casual discovery, and data mining.

Yiwen Zhang received the PhD degree in management science and engineering from the Hefei University of Technology, Anhui, China, in 2013.

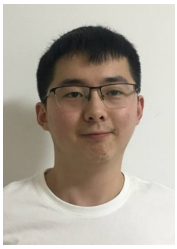
He is a Professor with the School of Computer Science and Technology, Anhui University, China. His current research interests include service computing, cloud computing, and big data.

Kui Yu received his PhD degree in Computer Science in 2013 from the Hefei University of Technology, China. From 2013 to 2015, he was a postdoctoral fellow at the School of Computing Science of Simon Fraser University, Canada. From 2015 to 2018, He was a research fellow at the University of South Australia, Australia.

He is a Professor with the School of of Computer and Information, Hefei University of Technology. His main research interests include causal discovery and machine learning.



Peng Zhou received the B.E. degree in computer science and technology from University of Science and Technology of China in 2011, and Ph.D degree in Computer Science from Institute of Software, Chinese Academy of Sciences in 2017. He is currently an Associate Professor with the School of Computer Science and Technology, Anhui University. He has published more than 20 papers in highly regarded conferences and journals, including IEEE TNNLS, IEEE T-CYB, Pattern Recognition, IJCAI, AAAI, SDM, ICDM, etc. His research interests include machine learning, data mining and artificial intelligence. More publications and codes can be found in his homepage: https://doctor-nobody.github.io/.



Bo Li received the B.S. degree from Henan University of Engineering, China, in 2020.

Currently, he is working toward the MSc degree in the School of Computer Science and Technology, Anhui University, China. His research interests include feature selection, casual discovery, and data mining.

Xindong Wu (F'11) is Director and Professor of the Key Laboratory of Knowledge Engineering with Big Data (the Ministry of Education of China), Hefei University of Technology, China. His research interests include big data analytics, data mining and knowledge engineering. He received his Bachelor's and Master's degrees in Computer Science from the Hefei University of Technology, China, and his Ph.D. degree in Artificial Intelligence from the University of Edinburgh, Britain, in 1993. He is a Foreign Member of the Russian Academy of Engineering, and a Fellow of IEEE and the AAAS (American Association for the Advancement of Science).