A Unified View of Causal and Non-causal Feature Selection

KUI YU, Hefei University of Technology
LIN LIU and JIUYONG LI, University of South Australia

In this article, we aim to develop a unified view of causal and non-causal feature selection methods. The unified view will fill in the gap in the research of the relation between the two types of methods. Based on the Bayesian network framework and information theory, we first show that causal and non-causal feature selection methods share the same objective. That is to find the Markov blanket of a class attribute, the theoretically optimal feature set for classification. We then examine the assumptions made by causal and non-causal feature selection methods when searching for the optimal feature set, and unify the assumptions by mapping them to the restrictions on the structure of the Bayesian network model of the studied problem. We further analyze in detail how the structural assumptions lead to the different levels of approximations employed by the methods in their search, which then result in the approximations in the feature sets found by the methods with respect to the optimal feature set. With the unified view, we can interpret the output of non-causal methods from a causal perspective and derive the error bounds of both types of methods. Finally, we present practical understanding of the relation between causal and non-causal methods using extensive experiments with synthetic data and various types of real-world data.

CCS Concepts: • **Computing methodologies** → **Feature selection**;

Additional Key Words and Phrases: Causal feature selection, non-causal feature selection, mutual information, Markov blanket, Bayesian network

ACM Reference format:
Kui Yu, Lin Liu, and Jiuyong Li. 2021. A Unified View of Causal and Non-causal Feature Selection. *ACM Trans. Knowl. Discov. Data* 15, 4, Article 63 (April 2021), 46 pages.
https://doi.org/10.1145/3436891

1 INTRODUCTION

Feature selection is to identify a subset of features (predictor variables) from the original features for model building or data understanding [29, 56, 60]. In the big data era, feature selection is more pressing than ever, since high-dimensional datasets have become ubiquitous in various

This work is partially supported by the National Key Research and Development Program of China (under grant 2020AAA0106100), Australian Research Council Discovery (ARC) Projects (under grant DP170101306), National Natural Science Foundation of China (under Grant 61876206), Open Project Foundation of Intelligent Information Processing Key Laboratory of Shanxi Province (under grant CICIP2020003), and Anhui Province Key Research and Development Plan (under grant 201904a05020073).

Authors' addresses: K. Yu, Key Laboratory of Knowledge Engineering with Big Data of Ministry of Education (Hefei University of Technology), and School of Computer Science and Information Engineering, Hefei University of Technology, 485 Daxia Road, Shushan District, Hefei, 230601, China; email: ykui713@gmail.com; L. Liu and J. Li, UniSA STEM, University of South Australia, Mawson Lakes, Adelaide, 5095, SA, Australia; emails: {Lin.Liu, Jiuyong.Li}@unisa.edu.au.

1556-4681/2021/04-ART63 $15.00
https://doi.org/10.1145/3436891

applications [63]. For example, in cancer genomics, a gene expression dataset can contain tens of thousands of features (genes). For another example, the Webb Spam Corpus 2011 has a collection of approximately 16 million features for web spam detection [53]. The high dimensionality not only incurs high computational cost and memory usage, but also deteriorates the generalization ability of prediction models [11]. Therefore, in the last two decades, feature selection has been well studied and has achieved great successes in building high quality classification models [2, 37, 43, 64]. Many feature selection methods have been proposed, and they fall into three main categories, filter, wrapper, and embedded methods. While filter feature selection methods are classifier or prediction model agnostic, the other two types of methods are classifier dependent. With the rapid increase of high-dimensional data, filter feature selection methods are attracting more attentions than ever, because of their fast processing speed and independence of prediction models (i.e., no bias on specific prediction models). In this article, we focus on filter methods, and in the rest of this article, feature selection refers to filter feature selection, unless otherwise mentioned.

In classical feature selection, an input feature is considered as a strongly relevant feature, a weakly relevant feature, or an irrelevant feature with respect to a class attribute [25], and the feature selection methods aim to find the strongly relevant features of the class attribute. To achieve this goal, typically, a classical feature selection method will rank the features according to their relevance to the class attribute, and then iteratively selects for inclusion the most relevant features [46].

Among those feature selection methods, an emerging feature selection approach is to identify a Markov blanket (MB) of the class attribute [21, 26]. The notion of MB was invented by Pearl [35] in the context of a Bayesian network (BN). The MB of a variable in a BN consists of its parents (direct causes), children (direct effects), and spouses (other parents of this variable's children) (for an exemplar MB, please see Figure 1 in Section 3). Thus identifying the MB of a class attribute explicitly induces local causal relations between the class attribute and the features, while classical feature selection methods do not. The MB discovery approach to feature selection ties feature predictive power and causality together, and thus it is able to achieve more interpretable and robust prediction models than classical feature selection methods [2]. Based on the discussion above, in this article, we call the MB discovery approach causal feature selection while the classical (filter) feature selection approach non-causal feature selection [2, 21].

A series of causal feature selection algorithms have been developed [10, 32, 36, 46]. These developed causal feature selection algorithms provide a new and complementary algorithmic methodology to enrich feature selection. To connect causal feature selection with non-causal feature selection, Tsamardinos et al. [46] were the first to build the connection between local causal discovery and feature selection, which opened the way to study the relation of causal and non-causal feature selection methods. Guyon et al. [21] conducted a comparison of the motivations and pros/cons of causal and non-causal feature selection approaches. However, the analysis was at conceptual and general discussion level. Brown et al. [11] unified information theoretic feature selection methods. These pioneer work provides a basis of studying causal and non-causal feature selection methods. However, regarding the relations between the two major approaches to feature selection, the following fundamental questions are yet to be investigated:

- First, what is the relation between the objectives of causal feature selection and non-causal feature selection, i.e., What is the relation between the MB of a class attribute and the set of all features strongly relevant to the class attribute?
- Second, driven by their respective objectives, how are the search strategies employed by the two types of feature selection methods different?
- Third, what are the underlying assumptions leading to the different search strategies?

To answer these questions, in this article, we develop a unified view to systematically study the relation of causal and non-causal feature selection from the perspectives of their objective functions, assumptions, search strategies, and the error bounds by employing the BN framework and information theory. Specifically, we have made the following contributions in this article:

— We derive a mutual information-based representation of the optimal feature set for classification. Based on the representation, we develop a unified representation of the objective function of causal and non-causal feature selection by showing that both types of methods share the same objective.
— We analyze the assumptions made by the major causal and non-causal feature selection methods in their search for the feature set specified by the objective function. Our findings show that these assumptions can be unified under the BN framework, and the assumptions can be represented as different levels of restrictions on the structure of the BN model of the problem under consideration.
— We analyze the search strategies of the causal and non-causal feature selection methods, and discover that as a result of the different levels of assumptions, different search strategies have been taken by these methods, which then result in different levels of approximations of the optimal feature set.
— We analyze the output of non-causal feature selection methods from a causal perspective and derive the error bounds of the two major approaches to feature selection.
— We conduct extensive experiments using synthetic and real-world datasets to validate the relationship between the assumptions and approximations made by causal and non-causal feature selection methods, the causal interpretations of non-causal feature selection, and the derived error bounds of both types of feature selection methods.

In summary, we propose a unified view to bridge the gap in current understanding of the relation between causal and non-causal feature selection methods. With the unified view, we are able to understand the mechanisms of the two major feature selection approaches, and thus to connect causality to predictive feature selection and interpret the output of non-causal methods using a causal framework. Moreover, by filling in the gap, we hope to leverage the cross-pollination between causal and non-causal feature selection to develop new methodologies promising to deliver more robust data analysis than each field could individually do.

The rest of the article is organized as follows. Section 2 discusses the related work, and Section 3 presents the key notations and definitions. Section 4 analyzes the objective functions and the rationale of causal and non-causal feature selection methods. Section 5 identifies and examines the assumptions made by causal and non-causal feature selection methods and their corresponding search strategies. Section 6 discusses the error bounds of causal and non-causal feature selection methods. Section 7 presents the experiments and demonstrates how the developed unified view provides practical understanding the relations between causal and non-causal feature selection methods, and Section 8 concludes the article.

2 RELATED WORK

In this section, we will review causal and non-causal (filter) feature selection methods. Excellent reviews of non-causal feature selection (i.e., filter, embedded, and wrapper) algorithms can be found in [11, 29] and the reference therein.

2.1 Non-causal Feature Selection

A general filter feature selection method consists of two elements: a search strategy for feature subset generation and an evaluation criterion for measuring relevance of the features. This evaluation criterion is to estimate how useful a feature or a feature subset may be when used in a learning algorithm (e.g., a classifier). As the feature selection by a filter method is carried out separately from the process of learning a model, an effective evaluation criterion plays a key role in filter methods. In the past decades, different evaluation criteria have been proposed, such as those based on distance [38], mutual information [41], dependency [42], and consistency [14]. Since mutual information is a general measure of feature relevance with several unique properties [13], there has been a significant amount of work on mutual information based feature selection methods developed in the past two decades (see [11, 50] for an exhaustive list).

In this article, we use mutual information as a basic tool to develop the unified view, so in this section, we focus on non-causal feature selection methods which are based on mutual information. Many advances in the field have been reported since the pioneer work of Lewis [28] and Battiti [6]. Lewis proposed the Mutual Information Maximization (MIM) criterion. MIM simply ranks the features in order of their MIM scores (i.e., the value of mutual information between a feature and the class attribute) and selects the top ψ most relevant features from the original feature set. However, MIM only considers feature relevance. Then Battiti proposed the Mutual Information Feature Selection (MIFS) criterion which not only considers feature relevance, but also adds a penalty for feature redundancy. MIFS uses a greedy search to select features sequentially (i.e., a single feature at a time), and iteratively constructs the final feature subset, as an alternative to the evaluation of the combinatorial explosion of all subsets of features.

Based on the MIFS criterion, many variants have been proposed. The representative algorithms include mRMR [37], CIFE [30], FCBF [62], mIMR [9], and MRI [55]. Yang and Moody proposed the Joint Mutual Information (JMI) criterion [57]. Compared to the MIFS criterion, the JMI criterion considers complementary information between features by evaluating class-conditional relevance, that is, whether a feature would provide more predictive information or not when it is used jointly with other features in the prediction compared with the case when the feature is used alone. The IF [51], DISR [34], CMIM [16], and RelaxMRMR [52] methods can be considered as the variants of the JMI criterion. Brown et al. [11] unified almost two decades of research on commonly used heuristics of mutual information based feature selection methods into the framework of conditional likelihood maximisation.

Owing to the difficulty of estimating mutual information with high-dimensional data, most existing mutual information based methods use various low-order approximations for estimating mutual information. While those approximations have been successful in certain applications, they are heuristic in nature and lack theoretical guarantees. Thus, the main problems with the majority of mutual information based methods are that in most cases it is unknown what makes up an optimal feature selection solution independent of the type of models fitted, and under which conditions a filter method will output an optimal feature set for classification [2, 21].

2.2 Causal Feature Selection

As an emerging type of filter methods, causal feature selection has attracted much attention in recent years [32, 54, 61]. By bringing causality into play, causal feature selection naturally provides causal interpretation about the relationships between features and the class attribute, enabling a better understanding of the mechanisms behind data [39]. Compared to non-causal feature selection, causal feature selection has been shown to be theoretically optimal, thus answers the questions of what makes up an optimal feature selection solution and under which conditions a filter method will output an optimal feature set for classification [2].

Causal feature selection is to find the MB of the class attribute in a causal Bayesian network (CBN), where an edge $X \rightarrow Y$ indicates that X is a direct cause (parent) of Y, and Y is a direct effect (child) of X. Then the MB of a variable of interest, such as the class attribute, consists of the parents, children, and the spouses (i.e. other parents of the children) of the class attribute. Therefore, the MB of the class attribute provides a complete picture of the local causal structure around it and the MB is a minimal set of features which renders the class attribute statistically independent from all the remaining features conditioned on the MB [35]. Theoretically, the MB of the class attribute is the optimal feature subset for classification [26, 46]. Accordingly, the discovery of the MB of a class attribute is actually a procedure of feature selection [2].

Koller and Sahami [26] were the first to introduce MBs to feature selection and proposed the Koller–Sahami (KS) algorithm. However, the KS algorithm is not guaranteed to find the actual MB. Margaritis and Thrun [33] invented the first sound MB discovery algorithm, GS (Growing-Shrinking) for BN structure learning.

Tsamardinos and Aliferis [46] improved the GS algorithm and proposed a series of MB discovery algorithms for optimal feature selection, which led to the Incremental Association-based MB (IAMB) family of algorithms, such as IAMB, inter-IAMB, IAMBnPC [48], and Fast-IAMB [59].

Given a variable of interest, IAMB and its variants learn the parents and children (PC) and spouses simultaneously and do not distinguish PC from spouses during the MB discovery procedure. And these algorithms require a large number of data samples exponential to the size of the MB of the variable, thus they would not be effective when a dataset has thousands of variables with a small-sized data samples.

Then a divide-and-conquer approach was proposed to mitigate the problem. The representative algorithms include HITION-MB [3], MMMB [47], PCMB [36], IPC-MB [18], and STMB [20]. The ideas behind these algorithms are as follows. They first find the PC of a variable of interest, then learn the variable's spouses. Thus these methods can return both the PC and MB sets of the variable. How to efficiently and effectively find the PC set of a variable is the key to this type of approach. The representative PC learning algorithms include PC-simple [12], MMPC [49], and HITON-PC [3].

3 BAYESIAN NETWORK, MARKOV BLANKET, AND FEATURE SELECTION

In this section, we introduce the notation used in this article (summarized in Table 1). Then, we present the basic definitions and theorems for readers to understand causal and non-causal feature selection.

Let C be the class attribute of interest, and C has φ distinct values (class labels), denoted as $c = \{c_1, c_2, \ldots, c_\varphi\}$ and $F = \{F_1, F_2, \ldots, F_n\}$ be the set of all n distinct features. Assuming that a training dataset D is defined by $D = \{(d_i, c_i), 1 \leq i \leq m,\ c_i \in c\}$, where m is the number of data instances, d_i is the ith data instance which is a n-dimensional vector defined on F, and c_i is a class label associated with d_i. For the convenience of presentation, we use V to represent the set of all variables under consideration, i.e., $V = F \cup \{C\} = \{V_1, V_2, \ldots, V_{n+1}\}$, where $V_i = F_i$ $(1 \leq i \leq n)$, and $V_{n+1} = C$. For $\forall V_i \in V$, let $V \setminus V_i$ indicate the set $V \setminus \{V_i\}$, that is, all features excluding V_i. We use $V_i \perp\!\!\!\perp V_j | S$, where $i \neq j$ and $S \subseteq V \setminus \{V_i, V_j\}$, to denote that V_i is conditionally independent of V_j given S, and $V_i \not\!\perp\!\!\!\perp V_j | S$ to represent that V_i is conditionally dependent on V_j given S. The definition of conditional independence (and dependence) is given as follows.

Definition 3.1 (Conditional Independence). Two distinct variables $V_i, V_j \in V$ are said to be conditionally independent given a subset of variables $S \subseteq V \setminus \{V_i, V_j\}$ (i.e., $V_i \perp\!\!\!\perp V_j | S$), if and only if $P(V_i, V_j | S) = P(V_i | S) P(V_j | S)$. Otherwise, V_i and V_j are conditionally dependent given S, i.e., $V_i \not\!\perp\!\!\!\perp V_j | S$.

Table 1. A Summary of Notations

Notation	**Meaning**
DAG	directed acyclic graph
V, F	a set of random variables
E	the edge set in a DAG
$P(V)$	joint probability distribution over V
G	a DAG
n	number of variables in F
m	number of data samples in F
C	the class attribute
$V_i, F_i,$	a single variable in V and F respectively
X, Y, Z	random variables
S	a subset of F
S^*	the set of features that leads to the minimal Bayes error rate
$Pa(F_i)$	the set of true parents of F_i
PC	the set of parents and children
$PC(C)$	the parent and child set of C
$ch(C)$	the child set of C
$SP(C)$	the set of spouses of C
$MB(C)$	the MB of C
$F_i \not\perp\!\!\!\perp F_j \mid S$	F_i and F_j are conditionally dependent given S
$F_i \perp\!\!\!\perp F_j \mid S$	F_i and F_j are conditionally independent given S
$F \setminus F_i$	all features in F excluding F_i
$H(X)$	the entropy of X
$H(X \mid Y)$	the entropy of X after observing values of Y
$I(X; Y \mid Z)$	conditional mutual information between X and Y given Z
$I(X; Y)$	mutual information between X and Y
P_{err}	the Bayes error rate
ψ	the user-defined parameter (the numbers of features)
$\lvert . \rvert$	cardinality of a set
α	significance level for an independence test

3.1 Bayesian Network, Markov Blanket, and Causal Feature Selection

In this section, we introduce the background knowledge related to causal feature selection, including the basics of BN, MB, and the aim of causal feature selection. Let $P(V)$ be the joint probability distribution over the set of all variables V, and $G = (V, E)$ represent a directed acyclic graph (DAG) with nodes V and edges E, where an edge represents the direct dependence relationship between two variables. In a DAG, $V_i \rightarrow V_j$ denotes that V_i is a parent of V_j and V_j is a child of V_i.

Definition 3.2 (BN). [35] The triplet $\langle V, G, P(V) \rangle$ is called a BN if the Markov condition as defined in Definition 3.3 holds.

Definition 3.3 (Markov Condition). [35] For a DAG G, the Markov condition holds in G if and only if every node of G is independent of any subset of its non-descendants conditioned on its parents.

A BN encodes the joint probability over a set of variables V and decomposes $P(V)$ into the product of the conditional probability distributions of the variables given their parents in G. Let

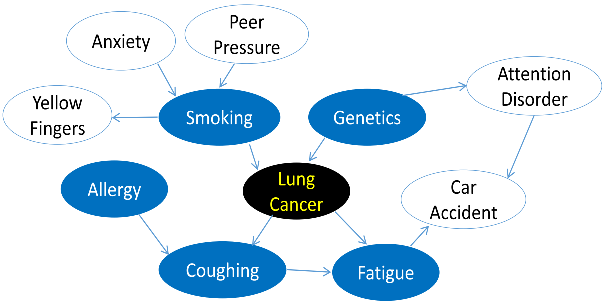


Fig. 1. An example of an MB in a lung-cancer BN.

$Pa(V_i)$ be the set of parents of V_i in G. Then, $P(V)$ can be written as

$$P(V_1, V_2, \ldots, V_{n+1}) = \prod_{i=1}^{n+1} P(V_i | Pa(V_i)). \tag{1}$$

In this article, we consider a CBN, a BN in which an edge $V_i \rightarrow V_j$ indicates that V_i is a direct cause of V_j [35, 44]. For simple presentation, however, we use the term BN instead of CBN. In the following, we introduce the key concepts and assumptions related to BNs and MBs.

Definition 3.4 (Faithfulness). [35] Given a BN $<V, G, P(V)>$, G is faithful to $P(V)$ if and only if every conditional independence present in P is entailed by G and the Markov condition. $P(V)$ is faithful if and only if G is faithful to $P(V)$.

Definition 3.5 (Causal Sufficiency). [35] Causal sufficiency assumes that any common cause of two or more variables in V is also in V.

Definition 3.6 (d-Separation). [35] In a path π of a DAG G, V_i and V_j are said to be blocked by a set of nodes $S \subset V$ if and only if (1) π contains a chain $V_i \rightarrow V_\omega \rightarrow V_j$ ($V_i \leftarrow V_\omega \leftarrow V_j$) or a fork $V_i \leftarrow V_\omega \rightarrow V_j$ such that the middle node V_ω is in S, or (2) π contains a v-structure $V_i \rightarrow V_\omega \leftarrow V_j$ such that $V_\omega \notin S$ holds and no descendants of V_ω are in S. A set S is said to d-separate V_i from V_j if and only if S blocks every path from V_i to V_j.

Theorem 3.7 [35, 44]. *Given a BN $<V, G, P(V)>$, under the faithfulness assumption, d-separation captures all conditional independence relations that are encoded in G, which implies that V_i and V_j in G are d-separated by $S \subset V \backslash \{V_i, V_j\}$, if and only if V_i and V_j are conditionally independent given S in $P(V)$.*

Theorem 3.7 concludes that under the assumption of faithfulness, conditional independence in a data distribution and d-separation in the corresponding DAG are equivalent.

Definition 3.8 (MB). [35] Under the faithfulness assumption, the MB of a variable in a BN is unique and consists of its parents (direct causes), children (direct effects), and spouses (other parents of the variable's children).

Figure 1 gives an example of an MB in the BN of lung cancer [21]. The MB of the variable *lung cancer* comprises: *Smoking* and *Gentics* (parents), *Coughing* and *Fatigue* (children), and *Allergy* (spouse). Given a dataset D defined on $F \cup \{C\}$, causal feature selection aims to find the MB of the class attribute C (denoted as $MB(C)$) from D [2]. In the following, Proposition 3.9 illustrates the relation between PC in a BN, and Proposition 3.10 presents the idea of how to identify spouses. The two propositions are the basis of designing causal feature selection algorithms.

Proposition 3.9 [44]. *In a BN, there is an edge between the pair of nodes V_i and V_j, if and only if $V_i \not\perp\!\!\!\perp V_j | S$, for all $S \subseteq V \setminus \{V_i, V_j\}$.*

Proposition 3.10 [44]. *In a BN, assuming that V_i is adjacent to V_j, V_j is adjacent to V_ω, and V_i is not adjacent to V_ω (e.g., $V_i \rightarrow V_j \leftarrow V_\omega$), if $\forall S \subseteq V \setminus \{V_i, V_j, V_\omega\}$, $V_i \perp\!\!\!\perp V_\omega | S$ and $V_i \not\perp\!\!\!\perp V_\omega | S \cup \{V_j\}$ hold, then V_i is a spouse of V_ω.*

3.2 Feature Relevancy and Non-causal Feature Selection

Non-causal feature selection categorizes a feature as strongly relevant, weakly relevant, or irrelevant to C [25] based on the following definitions in terms of conditional independence.

Definition 3.11 (Strongly Relevant Feature). [25] $F_i \in F$ is strongly relevant to C, if and only if there exists an assignment $F = f = (f_1, \ldots, f_{i-1}, f_i, f_{i+1}, \ldots, f_n)$ and $C = c_i,\ c_i \in c$, such that $P(F = f) > 0$ and $P(C = c_i | F = f) \neq P(C = c_i | F \setminus F_i = (f_1, \ldots, f_{i-1}, f_{i+1}, \ldots, f_n))$.

Definition 3.12 (Weakly Relevant Feature). [25] $F_i \in F$ is weakly relevant to C, if and only if F_i is not a strongly relevant feature and there exist $S \subset F \setminus F_i$, and an assignment $F_i = f_i$, $C = c_i$ and $S = s$ such that $P(S = s, F_i = f_i) > 0$ and $P(C = c_i | S = s, F_i = f_i) \neq P(C = c_i | S = s)$.

Definition 3.13 (Irrelevant Feature). [25] $F_i \in F$ is irrelevant to C, if and only if for any $S \subseteq F \setminus F_i$, for any assignment of F_i, S and C, denoted as f_i, s, and c_i, such that $P(C = c_i | S = s, F_i = f_i) = P(C = c_i | S = s)$.

A strongly relevant feature affects the conditional class distribution, and provides unique information about C, i.e., it cannot be replaced by other features. A weakly relevant feature is informative but redundant since it can be replaced by other features without losing information about C. An irrelevant feature does not bring any information about C and should be discarded.

Given a dataset D defined on $F \cup \{C\}$, non-causal (filter) feature selection aims to select all features that are strongly relevant to C [46]. In addition to the above conditional probability based definitions, recently, an explanation of feature relevance based on mutual information was proposed [8, 11, 50]. Before discussing the explanation, we first introduce the concepts about mutual information below. Given variable X, the entropy of X is defined as [13]

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

The entropy of X after observing the values of another variable Y is defined as

$$H(X|Y) = -\Sigma_y P(y) \Sigma_x P(x|y) \log P(x|y). \tag{3}$$

In Equations (2) and (3), $P(x)$ is the prior probability of $X = x$ (i.e., the value x that X takes), and $P(x|y)$ is the posterior probability of $X = x$ given $Y = y$. According to Equations (2) and (3), the mutual information between X and Y, denoted as $I(X, Y)$, is defined as

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

From Equation (4), the conditional mutual information between X and Y given another feature Z is defined as

$$\begin{aligned} I(X; Y|Z) &= H(X|Z) - H(X|YZ) \\ &= \Sigma_{z \in Z} P(z) \Sigma_{x \in X, y \in Y} P(x, y|z) \log \tfrac{P(x,y|z)}{P(x|z)P(y|z)}. \end{aligned} \tag{5}$$

Based on the above definitions about mutual information, we have the following propositions.

Proposition 3.14 [11]. *F_i is strongly relevant to C if and only if $I(F_i; C | F \setminus F_i) > 0$.*

Proposition 3.15 [11]. *F_i is weakly relevant to C if and only if $I(F_i;C|F \setminus F_i) = 0$ and $\exists S \subset F \setminus F_i$ such that $I(F_i;C|S) > 0$.*

Proposition 3.16 [11]. *F_i is irrelevant to C, if and only if $\forall S \subseteq F \setminus F_i, I(F_i;C|S) = 0$.*

4 CAUSAL AND NON-CAUSAL FEATURE SELECTION HAVE THE SAME OBJECTIVE

To develop a unified view of causal feature selection and non-causal feature selection, in this section, we will show that the two types of feature selection, although originating from different fields, share the same objective. In order to derive this conclusion (in Section 4.2), firstly in Section 4.1, inspired by the work in [11], we propose a mutual information based description of the optimal feature set for classification (i.e., Equation (12)), and then link the description to Bayes error rate of classification.

4.1 A Mutual Information Based Representation of the Objective Function of Optimal Feature Selection

Given a dataset D containing C and F, (filter) feature selection can be formulated as the problem of finding a subset $S^* \subset F$ such that

$$S^* = \underset{S \subset F}{\arg\max}\, P(C|S), \tag{6}$$

i.e., finding a subset S^* given which the conditional probability of C is maximized [11, 23].

Let $F = S \cup \overline{S}$ where S denotes the selected feature set and $\overline{S}$ represents the remaining features, i.e., $F \setminus S$. Given a dataset D of m instances, let $p(C|S)$ denote the true class distribution and $q(C|S)$ represent the predicted class distribution given S, then the conditional likelihood of C is $L(C|S,D) = \prod_{i=1}^{m} q(c_i|s_i)$, where $c_i \in c$ ($c = \{c_1, c_2, \ldots, c_\varphi\}$) represents the value of C in the i-th data instance and s_i denotes the value of feature set S in the i-th data instance. The (scaled) conditional log-likelihood of $L(C|S,D)$ is calculated by

$$\ell(C|S,D) = \frac{1}{m}\sum_{i=1}^{m} \log q(c_i|s_i). \tag{7}$$

Equation (7) can be re-written as Equation (8) below [11][1]

$$\ell(C|S,D) = \frac{1}{m}\sum_{i=1}^{m} \log \frac{q(c_i|s_i)}{p(c_i|s_i)} + \frac{1}{m}\sum_{i=1}^{m} \log \frac{p(c_i|s_i)}{p(c_i|f)} + \frac{1}{m}\sum_{i=1}^{m} \log p(c_i|f). \tag{8}$$

By negating Equation (8) and using E to represent statistical expectation, we have

$$-\ell(C|S,D) = E\left\{\log \frac{p(c|s)}{q(c|s)}\right\} + E\left\{\log \frac{p(c|f)}{p(c|s)}\right\} - E\left\{\log p(c|f)\right\}. \tag{9}$$

On the right hand side of Equation (9), the first term is the likelihood ratio between the true and predicted class distributions given S, averaged over the input data space. The second term equals to $I(C;\overline{S}|S)$, that is, the conditional mutual information between C and $\overline{S}$ given S [11]. The final term is $H(C|F)$ by Equation (3), the conditional entropy of C given all features, and is an irreducible constant.

Definition 4.1 (Kullback Leibler Divergence). [27] The Kullback Leibler divergence between two probability distributions $P(S)$ and $Q(S)$ is defined as $KL(P(S)||Q(S)) = \Sigma_s P(s) \log \frac{P(s)}{Q(s)} = E_s \log \{\frac{P(S)}{Q(S)}\}$.

[1] Please refer to Section 3.1 of [11] for the details on how to obtain Equations (7) and (8).

By Definition 4.1 and Equation (9), we have

$$\lim_{m\to\infty} -\ell(C|S,D) = KL(p(C|S)||q(C|S)) + I(C;\overline{S}|S) + H(C|F). \tag{10}$$

Since in Equation (10), $KL(p(C|S)||q(C|S))$ will approach zero with a large m. Based on Equation (10), we see that with a large value of m, minimizing $I(C;\overline{S}|S)$ maximizes $L(C|S,D)$. By the chain rule of mutual information, Equation (11) below holds.

$$\begin{aligned} I(C;F) &= I(C;\{S,\overline{S}\}) \\ &= I(C;S) + I(C;\overline{S}|S). \end{aligned} \tag{11}$$

Given the feature set F and the class attribute C, if $I(C;F)$ is fixed, then in Equation (11), minimizing $I(C;\overline{S}|S)$ is equivalent to maximizing $I(C;S)$. If $I(C;\overline{S}|S) = 0$ holds, $I(C;S)$ is maximized. Accordingly, by Equations (10) and (11), maximizing $I(C;S)$ is equivalent to maximizing the conditional likelihood of C (i.e., equivalent to maximizing $P(C|S)$). Thus, using mutual information, the objective function of feature selection of Equation (6) can be re-formulated as Equation (12) below.

$$S^* = \arg\max_{S\subset F} I(C;S). \tag{12}$$

In the following, we will show that the feature set S^* defined in Equation (12) is the set of features that leads to the minimal Bayes error rate. For a given classification problem, the minimum achievable classification error by any classifier is called its Bayes error rate [19]. We choose the Bayes error rate for justifying Equation (12) since it is the tightest possible classifier-independent lower-bound by depending on predictor features and the class attribute alone. Fano et al. [15, 24, 45] proposed the lower and upper bounds on the Bayes error rate, which connect the Shannon conditional entropy [40] to the Bayes error rate.

Let P_{err} represent the Bayes error rate, and the entropy $H(P_{err})$ is defined as

$$H(P_{err}) = -P_{err}\log P_{err} - (1-P_{err})\log(1-P_{err}). \tag{13}$$

Then given C and S, Fano's lower bound of the Bayes error rate [15] is defined as Equation (14) below

$$H(C|S) \leq H(P_{err}) + P_{err}\log(K-1). \tag{14}$$

Let $H(P_{err})^{-1}$ be the inverse of $H(P_{err})$, the upper bound of the Bayes error rate for a binary classification problem (K=2) is given as Equation (15) below [24, 45]

$$H(P_{err})^{-1} \leq P_{err} \leq 1/2H(C|S). \tag{15}$$

Meanwhile, considering $H(C|F) = H(C) - I(C;F)$ and $I(C;\overline{S}|S) = I(C;F) - I(C;S)$, Equation (10) is re-written as Equation (16) below

$$\lim_{m\to\infty} -\ell(C|S,D) = KL(p(C|S)||q(C|S)) + H(C|S). \tag{16}$$

In Equation (16), with a large m, $KL(p(C|S)||q(C|S))$ will approach zero. Thus we conclude that minimizing $H(C|S)$, that is, the conditional entropy of the class attribute C conditioning on S, is equivalent to maximizing the conditional likelihood of C or minimizing the Bayes error rate (from Equation (15)). Since $H(C|S) = H(C) - I(C;S)$, maximizing $I(C;S)$ in Equation (12) equals to minimizing the upper bound of $H(C|S)$, i.e., the upper bound of P_{err}. This thus justifies that the feature set selected by Equation (12) for classification will best facilitate minimizing the Bayes error rate. Equation (17) illustrates the relationships among $I(C;S)$, P_{err}, and $L(C|S,D)$ where both "<=>" denote "equivalent to," respectively.

$$\arg\min_{S\subset F} P_{err}(S) \;\; <=> \;\; \arg\max_{S\subset F} I(C;S) \;\; <=> \;\; \arg\max_{S\subset F} L(C|S,D). \tag{17}$$

4.2 The Objectives of Causal and Non-causal Feature Selection are the Same

In this section, we will demonstrate that the MB of C ($MB(C)$) is the feature set that maximizes Equation (12), and is the same as the set of strongly relevant features defined by non-causal feature selection.

Lemma 4.2 [35]. $\forall S \subset F \setminus MB(C),\ P(C|MB(C), S) = P(C|MB(C))$.

Lemma 4.3. $I(X;Y) \geq 0$ *with equality if and only if* $P(X,Y) = P(X)P(Y)$.

Lemma 4.4. $I(X;Y|Z) \geq 0$ *with equality if and only if* $P(X,Y|Z) = P(X|Z)P(Y|Z)$.

Clearly, by Equations (4) and (5), Lemmas 4.3 and 4.4 hold. Then according to Lemmas 4.2–4.4, Theorem 4.5 below illustrates that $MB(C)$ is the solution to Equation (12).

Theorem 4.5. $\forall S \subset F, I(C;MB(C)) \geq I(C;S)$ *with equality if and only if* $MB(C) = S$.

Proof. In the proof, we use MB to represent $MB(C)$.

Case 1: $\forall S \subseteq F \setminus MB$, by Equation (5), we have:

$$I(C;S|MB) = E_{\{C,S,MB\}} \log \frac{P(C,S|MB)}{P(C|MB)P(S|MB)}.$$

As $P(C,S|MB) = P(C|MB)P(S|MB)$, $I(C;S|MB) = 0$. By the chain rule, $I((S,MB);C) = I(C;MB) + I(C;S|MB) = I(C;S) + I(C;MB|S)$. Since $I(C;S|MB) = 0, I(C;MB) = I(C;S) + I(C;MB|S)$. By Lemmas 4.3 and 4.4, we get that $\forall S \subseteq F \setminus MB, I(C;MB) > I(C;S)$.

Case 2: $\forall S \subseteq MB$ and let $S' = MB \setminus S$, by $I(C;MB) - I(C;S) = I(C;S \cup S') - I(C;S) = I(C;S) + I(C;S'|S) - I(C;S) = I(C;S'|S)$, then $I(C;MB) \geq I(C;S)$ holds with equality if S equals to MB.

Case 3: Let $S' \subset MB$ and $S'' \subset F \setminus MB$, and $S = S' \cup S''$, by Equation (18) below, $I(C;S|MB) = 0$. Then by $I(C;MB) + I(C;S|MB) = I(C;S) + I(C;MB|S)$, in the case, $I(C;MB) > I(C;S)$.

$$\frac{P(C,S|MB)}{P(C|MB)P(S|MB)} = \frac{P(C,S'',MB)}{P(C|MB)P(S'',MB)} = \frac{P(C|S'',MB)P(S'',MB)}{P(C|MB)P(S'',MB))} = 1. \tag{18}$$

By Cases 1 to 3, $I(C;MB) \geq I(C;S)$ with equality holds if S equals to MB. □

Corollary 4.6. *Under the faithfulness assumption,* $\forall F_i \in F$, F_i *belongs to* $MB(C)$, *if and only if* F_i *is a strongly relevant feature.*

Proof. In the proof, we use MB to represent $MB(C)$. $PC(C)$ denotes parents and children of C and $SP(C)$ represents spouses of C.

We firstly prove that if $F_i \in MB$, F_i is a strongly relevant feature. Since $MB = PC(C) \cup SP(C)$ and $PC(C) \cap SP(C) = \emptyset$, then (1) $\forall F_i \in PC(C)$ and $\forall S \subseteq F \setminus F_i$, by Proposition 3.9, $I(F_i;C|S) > 0$, and thus, $I(F_i;C|F \setminus F_i) > 0$ holds; (2) $\forall F_i \in SP(C)$ via child $F_\omega \in PC(C)$, by Proposition 3.10, there exists a $S \subset F \setminus F_i$ such that $I(F_i;C|S) = 0$ but $I(F_i;C|S \cup \{F_\omega\}) > 0$. Then, $\forall F_j \in F \setminus \{F_\omega, F_i\}$, $I(F_i;C|F \setminus F_j) = I(F_i;C|\{S \cup F_\omega \cup F \setminus \{S \cup F_\omega \cup F_i\}\})$. So if $F_i \in SP(C), I(F_i;C|F \setminus F_j) > 0$ holds. By Proposition 3.14, F_i is a strongly relevant feature.

We now prove that a strongly relevant feature of C must be in MB. If F_i is a strongly relevant feature, by Proposition 3.14, $I(F_i;C|F \setminus F_i) > 0$. Assume $F_i \notin MB$, $S' = F \setminus \{F_i\} \cup MB$, and $S = F \setminus F_i = MB \cup S'$, we have:

$$\begin{aligned} I(F_i;C|F \setminus F_i) &= I(F_i;C|S) \\ &= E_{\{C,S,F_i\}} \log \frac{P(C,F_i|S)}{P(C|S)P(F_i|S)} \\ &= E_{\{C,S,F_i\}} \log \frac{P(C,F_i,S)}{P(C|S)P(F_i|S)P(S)} \\ &= E_{\{C,S,F_i\}} \log \frac{P(C|F_i,S)P(F_i|S)}{P(C|S)P(F_i|S)} \\ &= E_{\{C,S,F_i\}} \log \frac{P(C|F_i,S)}{P(C|S)} \\ &= E_{\{C,S',MB,F_i\}} \log \frac{P(C|F_i,S',MB)}{P(C|S',MB)} \\ &= 0. \end{aligned} \tag{19}$$

This makes a contrary, and thus $F_i \in MB(C)$. □

Accordingly, given a dataset D defined on $F \cup \{C\}$, by the analysis above, we show that $MB(C)$ maximizes the objective function in Equation (12) and it is the same as the set of strongly relevant features.

5 CAUSAL AND NON-CAUSAL FEATURE SELECTION: ASSUMPTIONS AND APPROXIMATIONS

For both causal and non-causal feature selection methods, finding a subset S that maximizes $I(C;S)$ (i.e., solving the objective function in Equation (12)) is a challenging combinatorial optimization problem. An exhaustive search will be of $O(2^n)$ time complexity. Although restricting the maximum size of S to ς ($\varsigma < n$) will reduce the time complexity to $O(\varsigma^n)$ where ς^n is the number of all subsets of F containing ς or less features, the computational cost will still be high. Therefore, both causal and non-causal feature selection methods have adopted a greedy strategy by considering features one by one to optimize Equation (12) [2, 5, 11]. That is, at each iteration, given the set S currently selected, choose $X^* \in F \setminus S$ such that

$$\begin{aligned} X^* &= \arg\max\nolimits_{X \in F \setminus S} I(S \cup X; C) \\ &= \arg\max\nolimits_{X \in F \setminus S} \{I(S;C) + I(X;C|S)\}. \end{aligned} \tag{20}$$

As for all $X \in F \setminus S$, the first item in Equation (20) is the same, finding X^* becomes solving the following optimization problem:

$$X^* = \underset{X \in F \setminus S}{\arg\max}\, I(X;C|S). \tag{21}$$

However, in Equation (21), when the size of S increases, computing the multidimensional mutual information becomes impractical because it demands a large number of training samples, exponential in the number of features in S. To tackle this challenge, different feature selection methods make different assumptions on the interactions (or dependency) between features in the underlying data distributions for the calculation of $I(X;C|S)$.

As described previously, a BN provides a representation of the probabilistic dependence among a set of variables under consideration. This provides us the opportunity to unify the dependence assumptions made by the feature selection methods under the BN framework. In this article, we propose a structure assumption approach to understanding the assumptions made by causal and non-causal feature selection methods and how these different levels of structural assumptions lead to the different approximations in their search for the solutions to Equation (21).

In the following, firstly Section 5.1 provides a summary of our findings on the structural assumptions and how they are related to the approximations, then in Sections 5.2 and 5.3, we discuss the findings in detail by analyzing the assumptions and approximations made by the commonly used non-causal and causal feature selection methods.

5.1 Summary of Findings

5.1.1 Structural Assumptions and Search Strategies. As illustrated in Figure 2, we have found that the dependence/independence relationships among features assumed by both causal and non-causal feature selection methods can be represented as different restrictions to the structure of the BN model of the set of variables under study. Based on the assumed BN structures, causal and non-causal methods select the subset of features, $S \subset F$, with the conditional likelihood of the class attribute C given S, $P(C|S)$ as close to $P(C|F)$ as possible.

Figure 3 summarizes the BN structure assumptions and search strategies used by causal and non-causal feature selection methods for the calculation of $I(X;C|S)$. The number after each equation in Figure 3 is the same as the equation number given in Sections 5.2 and 5.3. From Figure 3,

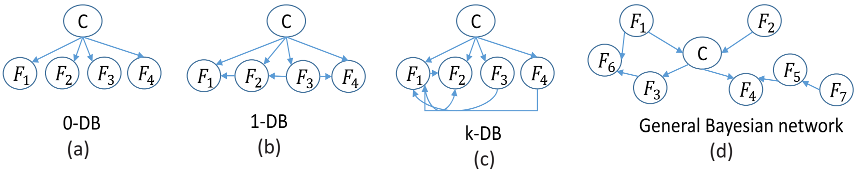


Fig. 2. An illustration of the BN structures corresponding to the structural assumptions made by the non-causal and causal feature selection methods.

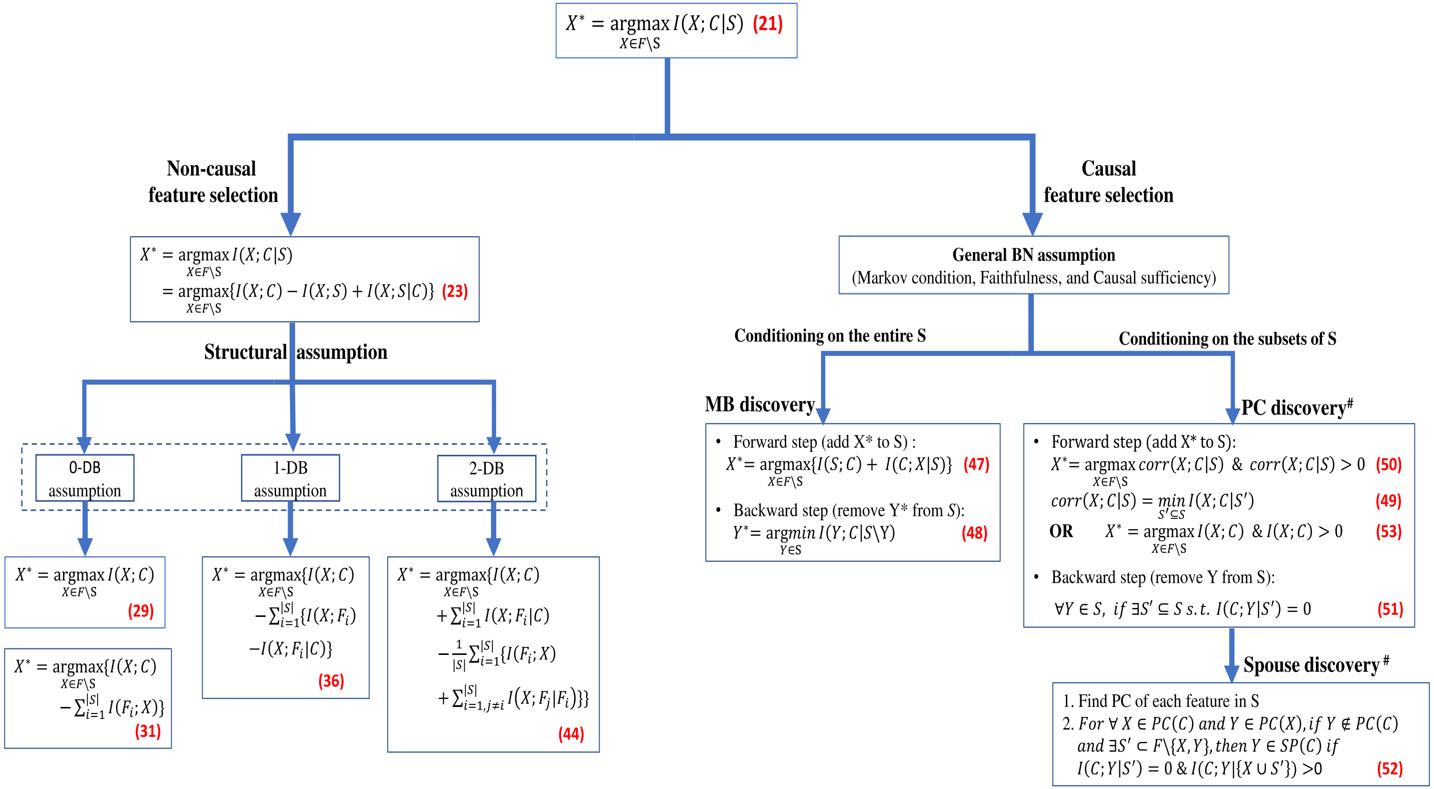
we see that a non-causal feature selection method firstly decomposes the multidimensional mutual information $I(X;C|S)$ into three terms $\{I(X;C) - I(S;X) + I(S;X|C)\}$ (See Equation (22)), then calculates the multidimensional mutual information $\{-I(S;X) + I(S;X|C)\}$ using linear combination of low-order mutual information terms based on the respective naive BN assumption made on the dependence/independence between features. We call the assumptions made by non-causal feature selection methods the series of naive BN assumptions, because the assumptions can be represented by the family of BNs with the restricted structures as illustrated in Figure 2(a), (b), and (c). For these naive BN structures, the class attribute has no parents while all the features each can only have a fixed number of parents, denoted as k-dependency (or k-DB) assumptions, where each feature can have at most other k features as its parents (details in Section 5.2).

Causal feature selection methods assume that one can learn from the given dataset a (general) BN without structural restrictions (as the example in Figure 2(d)), and in the learnt BN, X^* in Equation (21) is a feature in the MB of the class attribute. Therefore, as shown in Figure 3, causal feature selection does not decompose $I(X;C|S)$ for the use of any structural assumptions, and the assumptions made by causal feature selection are only those for a general BN and its learning, i.e., the Markov condition (Definition 3.3), the faithfulness (Definition 3.4), and causal sufficiency (Definition 3.5) assumptions. Unlike the non-causal feature selection methods, these assumptions do not pose any structural restrictions on a BN learnt from data (thus called the general BN assumptions in this article).

5.1.2 Linking the Assumptions with Approximations. We use the pyramid in Figure 4(a) to visualize the difference in the strictness of the structural assumptions made by the different feature selection methods. We see that causal feature selection methods make the weakest assumptions (no restrictions on the structures of the BN), while the non-causal feature selection methods make assumptions with different levels of strictness in terms of the maximum number of parents that a feature can have in addition to the class attribute (the value of k in Figure 4(a)).

As a result of the differences in the strictness of the structural assumptions, the degree of the corresponding approximations taken by the feature selection methods in their calculation of the multidimensional mutual information ($I(X;C|S)$) are different, and they can be visualized using an upside down pyramid (Figure 4(b)). Causal feature selection methods, since having had no structural restrictions, take fewer approximations by calculating higher order mutual information between X and C conditioning on all or a subset of the already selected features S (details of the conditioning sets are to be discussed in Section 5.3). Referring back to Figure 3, the non-causal feature selection methods eventually only look at the pairwise mutual information between X and C without conditioning on other features.

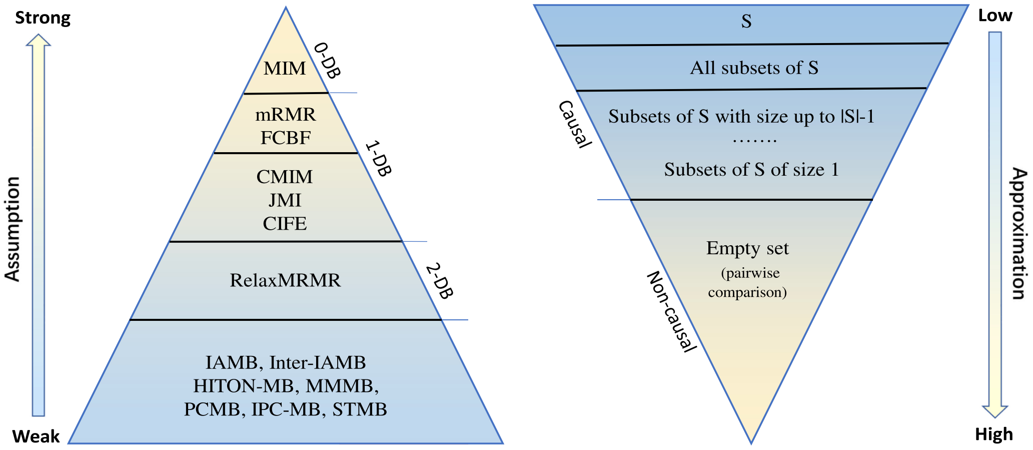
Therefore, in theory, a non-causal feature selection method is often more efficient than a causal feature selection method (see the analysis of time complexity in Sections 5.2.4 and 5.3), while the feature set obtained by a causal feature selection method is closer to the optimal feature set (i.e., the MB of the class attribute) than that of a non-causal feature selection method (see detailed analysis in Sections 5.2 and 5.3). However, as we will see in later sections, in practice, causal feature selection



$X^* = \underset{X\in F\backslash S}{\operatorname{argmax}} I(X;C|S)$ (21)
Non-causal feature selection
Causal feature selection
$X^* = \underset{X\in F\backslash S}{\operatorname{argmax}} I(X;C|S)$
$= \underset{X\in F\backslash S}{\operatorname{argmax}}\{I(X;C) - I(X;S) + I(X;S|C)\}$ (23)
Structural assumption
0-DB assumption
1-DB assumption
2-DB assumption
$X^* = \underset{X\in F\backslash S}{\operatorname{argmax}} I(X;C)$ (29)
$X^* = \underset{X\in F\backslash S}{\operatorname{argmax}}\{I(X;C) - \sum_{i=1}^{|S|} I(F_i;X)\}$ (31)
$X^* = \underset{X\in F\backslash S}{\operatorname{argmax}}\{I(X;C) - \sum_{i=1}^{|S|}\{I(X;F_i) - I(X;F_i|C)\}$ (36)
$X^* = \underset{X\in F\backslash S}{\operatorname{argmax}}\{I(X;C) + \sum_{i=1}^{|S|} I(X;F_i|C) - \frac{1}{|S|}\sum_{i=1}^{|S|}\{I(F_i;X) + \sum_{i=1,j\neq i}^{|S|} I(X;F_j|F_i)\}\}$ (44)
General BN assumption
(Markov condition, Faithfulness, and Causal sufficiency)
Conditioning on the entire S
Conditioning on the subsets of S
MB discovery
PC discovery#
• Forward step (add X* to S) :
$X^* = \underset{X\in F\backslash S}{\operatorname{argmax}}\{I(S;C) + I(C;X|S)\}$ (47)
• Backward step (remove Y* from S):
$Y^* = \underset{Y\in S}{argmin}\, I(Y;C|S\backslash Y)$ (48)
• Forward step (add X* to S):
$X^* = \underset{X\in F\backslash S}{\operatorname{argmax}}\, corr(X;C|S)\ \&\ corr(X;C|S) > 0$ (50)
$corr(X;C|S) = \underset{S'\subseteq S}{min}\, I(X;C|S')$ (49)
OR $X^* = \underset{X\in F\backslash S}{\operatorname{argmax}} I(X;C)\ \&\ I(X;C) > 0$ (53)
• Backward step (remove Y from S):
$\forall Y \in S,\ if\ \exists S' \subseteq S\ s.t.\ I(C;Y|S') = 0$ (51)
Spouse discovery#
1. Find PC of each feature in S
2. $For\ \forall\ X \in PC(C)\ and\ Y \in PC(X), if\ Y \notin PC(C)$ $and\ \exists S' \subset F\backslash\{X,Y\}, then\ Y \in SP(C)\ if$ $I(C;Y|S') = 0\ \&\ I(C;Y|\{X \cup S'\}) > 0$ (52)

#: With some algorithms, such as HITON-PC, the PC discovery and Spouse discovery steps are done interleavingly. Details see Section 5.3.2

Fig. 3. A road map of how causal and non-causal feature selection searches for X^* in Equation (21).



(a) Strictness of assumptions

(b) Level of approximations

Fig. 4. Strictness of structural assumptions and the corresponding level of approximations taken by causal and non-causal feature selection methods when calculating $I(X;C|S)$. (a) the strictness of structural assumptions in terms of maximum number of parents a feature can have (excluding the class attribute). Names of typical methods are shown. (b) the level of approximations in terms of the size of conditioning set used in the calculation.

does not always outperform non-causal feature selection when a dataset has high dimensionality and small-sized data samples, because the number of samples required by causal feature selection can be exponential in the number of features in S.

5.1.3 Causal Interpretation and Non-causal Feature Selection. By representing the dependency among predictive features and the class attribute using BN structures, we present a causal interpretation of the features selected by non-causal methods (see the detailed analysis in Section 5.2.5) and achieve the following findings. We have found that the non-causal feature selection methods prefer features within $MB(C)$ to the features not in $MB(C)$, which confirms that strongly relevant features belong to $MB(C)$ (i.e., Corollary 4.6). This finding provides a causal interpretation of the output of the non-causal feature selection methods and explains why non-causal feature selection also can achieve excellent classification results. This also provides a novel perspective to understand the relations between the two types of feature selection methods, and may motivate researchers to use the cross-pollination between causal and non-causal feature selection methods to develop novel methodologies promising to scalable local-to-global causal structure learning and feature selection with theoretical guarantees.

5.2 Non-causal Feature Selection: Assumptions and Approximations

In this section, we will explore in detail the assumptions made by non-causal feature selection under the naive BN framework, and under the assumptions how the major existing non-causal feature selection algorithms produce the same result as in Equation (21).

By $I(X;S;C) = I(X;S) - I(X;S|C) = I(X;C) - I(X;C|S)$, we have

$$I(X;C|S) = I(X;C) - I(X;S) + I(X;S|C). \quad (22)$$

The three terms on the right side of Equation (22) have the following interpretation:

- $I(X;C)$ corresponds to the relevancy of X to C.
- $I(X;S)$ represents the redundancy of X with respect to S.
- $I(X;S|C)$ indicates the class-conditional relevance, which considers the situation where a feature provides more predictive information jointly with another feature than by itself with respect to C. Since $I((S,X);C) = I(S;C) + I(X;C|S)$, $I(X;C|S) = I((S,X);C) - I(S;C)$. In Equation (22), when $I(X;S) > I(X;S|C)$, $I(X;C|S) < I(X;C)$ holds, and thus $I((S,X);C) < I(S;C) + I(X;C)$. This means that X contains redundant information about C when we add X to S. When $I(X;S) < I(X;S|C)$, $I(X;C|S) > I(X;C)$ holds, and thus $I((S,X);C) > I(S;C) + I(X;C)$. This indicates that X and S have a positive interaction and $I((S,X);C)$ provides more information than $I(S;C) + I(C;X)$.

By Equation (22), Equation (21) can be re-written as

$$X^* = \underset{X \in F \setminus S}{\arg\max} \{I(X;C) - I(X;S) + I(X;S|C)\}. \tag{23}$$

To reduce computational costs in the search for X^* in Equation (23), different non-causal feature selection methods make different assumptions, and thus adopt different level of approximations when calculating $I(X;S)$ and $I(X;S|C)$ by using a linear combination of low-order mutual information terms.

In the following, we will explore these assumptions and approximations in relation to Equation (23). Using a general BN to represent the relation of all features and the class attribute, we have

$$P(C|F) \propto P(C|Pa(C)) \prod_{i=1}^{n} P(F_i|Pa(F_i)). \tag{24}$$

A naive BN is a restricted BN, which considers the class attribute C as a special variable that has no parents and each of the remaining variables in the network only has the class attribute C and a fixed number of other features as its parents [17]. Let k represent the maximum number of parents (excluding the class attribute) a feature can have, we call the naive BN a k-dependency (k-DB) naive BN. A 0-DB network (as illustrated in Figure 2(a)) is the commonly known naive Bayes (NB) network. A NB network assumes that each variable only has one parent, i.e., C, and all features are conditionally independent of each other given C. A 1-DB network (as illustrated in Figure 2(b)) is known as a Tree-Augmented Naive (TAN) Bayes network, which allows each variable to have at most one other feature in addition to C as its parent. A 2-DB network (see an example in Figure 2(c)) relaxes NB's and TAN's independence assumptions by allowing each feature to have a maximum of two other features as parents to generalize to higher degrees of variable interactions.

Let $ncl\_pa(F_i)$ denote the set of parents of F_i excluding the class attribute C, in a k-DB naive BN, Equation (24) becomes

$$P(C|F) \propto P(C) \prod_{i=1}^{n} P(F_i|C, ncl\_pa(F_i)),\ |ncl\_pa(F_i)| = k \ \&\ |pa(C)| = 0. \tag{25}$$

5.2.1 Approximations Under 0-DB (NB) Structural Assumptions.

The following NB network assumption ($k = 0$) is often made by non-causal feature selection methods.

Assumption 1. In a NB network, $\forall F_i, F_j \in F$ and $i \neq j$, F_i and F_j are assumed to be conditionally independent given the class attribute C, that is, $P(F_i, F_j|C) = P(F_i|C)P(F_j|C)$.

By Assumption 1, Equation (25) is transformed into

$$P(C|F) \propto P(C)\prod_{i=1}^{n} P(F_i|C),\ |ncl\_pa(F_i)| = 0\ \&\ |pa(C)| = 0. \tag{26}$$

By Assumption 1 and Equation (26), in Equation (23), the class-conditional relevancy $I(X;S|C)$ is calculated as Equation (27) as follows.

$$\begin{aligned} I(X;S|C) &= E_{x,s,c}\log\tfrac{P(X,S|C)}{P(S|C)P(X|C)} \\ &= E_{x,s,c}\log\tfrac{P(S|C)P(X|C)}{P(S|C)P(X|C)} \\ &= 0. \end{aligned} \tag{27}$$

Then under Assumption 1 and Equation (27), Equation (23) becomes

$$\arg\max_{X\in F\setminus S}\{I(X;C) - I(X;S) + I(X;S|C)\} = \arg\max_{X\in F\setminus S}\{I(X;C) - I(X;S)\}. \tag{28}$$

Since the redundancy term $I(X;S) = H(S) - H(S|X)$, and by the chain rule of entropy, we have $H(S|X) = \sum_{F_i\in S} H(F_i|F_{i-1},\ldots,F_1,X)$. If we further employ Assumption 2 below to restrict the interactions between a feature in S and a feature in $F\setminus S$, in Equation (23), $I(X;S) = 0$ holds.

Assumption 2. For $\forall F_i \in S$ and $\forall F_j \in F\setminus S$, $P(F_i,F_j) = P(F_i)P(F_j)$.

By Assumptions 1 and 2, the objective function in Equation (23) is simplified to the following, which is only based on the mutual information between a feature and the class attribute:

$$X^* = \arg\max_{X\in F\setminus S} I(X;C). \tag{29}$$

The objective in Equation (29) is the MIM criterion initially presented in [28].

Assumption 2 is a strong assumption that the features in S and the features in $F\setminus S$ are pairwise independent. To deal with the redundancy between features, we discuss Assumption 3 below, which is less restrictive than Assumption 2.

Assumption 3. The selected features in S are conditionally independent given an unselected feature $X \in F\setminus S$, that is, $P(S|X) = \prod_{i=1}^{|S|} P(F_i|X)\ (F_i \in S)$.

Since $I(X;S) = H(S) - H(S|X)$, by the chain rule and Assumption 3, we have

$$\begin{aligned} I(X;S) &= H(S) - \textstyle\sum_{i=1}^{|S|} H(F_i|X) \\ &= H(S) - \textstyle\sum_{i=1}^{|S|} H(F_i) + \textstyle\sum_{i=1}^{|S|} I(F_i;X). \end{aligned} \tag{30}$$

Since at each time, $\forall X \in F\setminus S$, the first two terms in Equation (30) are the same, then $I(X;S)$ is decomposed into a sum of pairwise mutual information terms. Further based on Assumption 1, $I(X;S|C) = 0$, then the objective function in Equation (23) becomes:

$$X^* = \arg\max_{X\in F\setminus S}\left\{I(X;C) - \sum_{i=1}^{|S|} I(F_i;X)\right\}. \tag{31}$$

Equation (31) is the criterion of "max-relevance and min-redundancy" [37]. Based on Equation (31), Battiti [6] presents the following MIFS criterion:

$$X^* = \arg\max_{X\in F\setminus S}\left\{I(X;C) - \beta\sum_{i=1}^{|S|} I(F_i;X)\right\}. \tag{32}$$

$\beta \in [0,1]$ in the MIFS criterion is a penalty for balancing the relevance and redundancy terms. When $\beta = 0$, Equation (32) becomes Equation (29), that is, the MIM criterion. As $\beta = 1$, Equation (32) is reduced to Equation (31). If $\beta = 1/|S|$, Equation (32) becomes

$$X^* = \arg\max_{X \in F \setminus S} \left\{ I(X;C) - \frac{1}{|S|} \sum_{i=1}^{|S|} I(F_i;X) \right\}. \tag{33}$$

Equation (33) is the well-known mRMR (max-Relevance and Min-Redundancy) algorithm presented in [37]. Meanwhile, from Equation (33), we can see that as the size of S increases, Equation (33) will tend asymptotically towards Equation (29). There are other feature selection methods based on the idea of max-relevance and min-redundancy shown in Equation (31), such as the representative feature selection algorithm, Fast Correlation Based Filter (FCBF) [62]. FCBF divides the "max-relevance and min-redundancy" criterion into two steps, that is, the forward step (max-relevance) and backward step (min-redundancy).

- Forward step: FCBF selects a subset of features S that $\forall X \in S$, $I(C;X) > 0$, then sorts the features in S by their mutual information with C in descending order.
- Backward step: beginning with the first feature $X \in S$, if $\exists Y \in S \setminus X$ such that $I(X;Y) > I(X;C)$, then it removes Y from S as a redundant feature to X. The FCBF algorithm is terminated until the last feature in S is checked.

At the forward step, FCBF only selects features that are relevant to C, and this implies Assumption 1. The backward step implies Assumption 3. At the backward step, for X, $Y \in S$, if $I(X;C) > I(Y;C)$ and $I(X;Y) > I(X;C)$, then Y can be removed from S. FCBF does not need to specify the number of selected features in advance. Instead, FCBF uses a threshold δ ($\delta > 0$) at the forward step and keeps features satisfying $I(C;X) \geq \delta$.

5.2.2 Approximations with 1-DB (TAN) Structural Assumptions. Under Assumption 1, in Equation (23), $I(X;S|C) = 0$ holds. A TAN BN relaxes Assumption 1 to allow each feature to be dependent on one other feature in addition to C and makes the following assumption, i.e., Assumption 4, which states that the features within S are class-conditionally independent given an unselected feature $X \in F \setminus S$ and C.

Assumption 4. $\forall F_i, F_j \in S$ and $i \neq j$, F_i and F_j are assumed to be conditionally independent given an unselected feature $X \in F \setminus S$ and C, that is, $P(F_i, F_j|X, C) = P(F_i|C,X)P(F_j|C,X)$.

Thus for a TAN BN, Equation (25) becomes

$$P(C|F) \propto P(C) \prod_{F_j \in F,\ F_i \in F \setminus F_j} P(F_i|C,F_j),\ |ncl\_pa(F_i)| = 1\ \&\ |pa(C)| = 0. \tag{34}$$

Then by the chain rule, we get $H(S|X,C) = \sum_{F_i \in S} H(F_i|X,C)$. By Equation (27), $I(X;S|C) = 0$ only and if only Assumption 1 holds, and thus by Assumption 4, $I(X;S|C)$ can be decomposed as follows.

$$\begin{aligned} I(X;S|C) &= H(S|C) - H(S|X,C) \\ &= H(S|C) - \textstyle\sum_{F_i \in S} H(F_i|X,C) \\ &= H(S|C) - \textstyle\sum_{F_i \in S} \{H(F_i|C) - I(F_i;X|C)\}. \end{aligned} \tag{35}$$

Since $H(S|C) - \sum_{F_i \in S} H(F_i|C)$ in Equation (35) is the same for $\forall F_i \in S$ and meanwhile assuming that Assumption 3 holds for feature interactions between the selected features in S and the unselected feature in $F \setminus S$, then by Equation (30) (under Assumption 3) and Equation (35) (under Assumption 4), Equation (23) becomes

$$X^* = \arg\max_{X \in F \setminus S} \{I(X;C) - \Sigma_{F_i \in S} I(X;F_i) + \Sigma_{F_i \in S} I(X;F_i|C)\}. \tag{36}$$

Brown et al. [11] have proposed that many mutual information-based non-causal feature selection methods can fit within the following parameterized criterion. β and γ play the role of balancing factors (in general $\beta \in [0,1]$ and $\gamma \in [0,1]$).

$$X^* = \underset{X \in \{F \setminus S\}}{\arg\max} \left\{ I(X;C) - \beta \sum_{F_i \in S} I(X;F_i) + \gamma \sum_{F_i \in S} I(X;F_i|C) \right\}. \tag{37}$$

If $\beta = 1/|S|$ and $\gamma = 1/|S|$, then we have

$$X^* = \underset{X \in F \setminus S}{\arg\max} \left\{ I(X;C) - \frac{1}{|S|} \Sigma_{F_i \in S} I(X;F_i) + \frac{1}{|S|} \Sigma_{F_i \in S} I(X;F_i|C) \right\}. \tag{38}$$

Using Equation (38) for feature selection, the representative algorithm is the JMI algorithm [57]. If $\beta = 1$ and $\gamma = 1$, Equation (37) is reduced to Equation (36) used by the CIFE algorithm [30]. The CMIM method [16] adopts an objective function as follows:

$$X^* = \underset{X \in F \setminus S}{\arg\max} \left\{ I(X;C) - \max_{F_i \in S} \{ I(X;F_i) - I(X;F_i|C) \} \right\}. \tag{39}$$

5.2.3 *Approximations with 2-DB Structural Assumptions.*

To deal with a higher-order dependency between features, the recent work in [52] calculates $I(X;S)$ in Equation (23) by exploring the 2-DB structure assumptions.

The 2-DB structure relaxes NB's and TAN's independence assumptions by allowing each feature to have at most two features as parents, i.e., $|ncl\_pa(F_i)| = 2$, in addition to C, and makes the following assumptions.

Assumption 5a. $\forall F_i \in S$ and $\forall F_j \in S$ $(i \neq j)$ are assumed to be conditionally independent given an unselected feature $X \in F \setminus S$ and any feature $Y \in F \setminus \{F_i \cup F_j\}$, that is, $P(F_i, F_j|X, Y) = P(F_i|X, Y)P(F_j|X, Y)$.

Assumption 5b. For $\exists F_j \in S$ and $\forall F_i \in F \setminus F_j$ are conditionally independent given an unselected feature $X \in F \setminus S$, that is, $P(F_j, F_i|X) = P(F_i|X)P(F_j|X)$.

Thus, with a 2-DB structure, Equation (25) is transformed into Equation (40).

$$P(C|F) \propto P(C) \prod_{i=1 (F_i \in F \setminus \{F_j \cup F_\omega\})}^{n} P(F_i|C, F_j, F_\omega)), \ |ncl\_pa(F_i)| = 2 \ \& \ |pa(C)| = 0. \tag{40}$$

With the structure assumptions, the redundancy term $I(X;S)$ in Equation (23) is computed as follows. Since $I(X;S) = H(S) - H(S|X)$ under Assumptions 5a and 5b, $H(S|X)$ is calculated as follows.

$$\begin{aligned} H(S|X) &= -\textstyle\sum_{i=1}^{|S|} \sum_{F_1,\ldots,F_i,X} P(F_1,\ldots,F_i,X) \log P(F_i|F_{i-1},\ldots,F_1,X) \\ &= P(F_1,\ldots,F_j,X) \log P(F_j|F_{j-1},\ldots,F_1,X) \\ &\quad + \textstyle\sum_{i=1 (i \neq j)}^{|S|-1} P(F_{i-2},\ldots,F_1,F_j,X) \log P(F_i|F_{i-2},\ldots,F_1,F_j,X) \\ &= H(F_j|X) + \textstyle\sum_{i=1, i \neq j}^{|S|-1} H(F_i|F_j,X). \end{aligned} \tag{41}$$

By Equation (41), $I(X;S)$ is decomposed as Equation (42) as follows.

$$\begin{aligned} I(X;S) &= H(S) - H(S|X) \\ &= H(S) - \{H(F_j|X) + \textstyle\sum_{i=1, i \neq j}^{|S|-1} H(F_i|F_j,X)\} \\ &= H(S) - H(F_j) + I(F_j;X) - \textstyle\sum_{i=1, i \neq j}^{|S|-1} \{H(F_i|F_j) - I((F_i, X|F_j)\}. \end{aligned} \tag{42}$$

In Equation (42), at each iteration, for $\forall X \in F \setminus S$, $H(S) - H(F_j) - \sum_{i=1, i \neq j}^{|S|-1} H(F_i|F_j)$ is the same. Meanwhile, to avoid the need of checking which feature in S satisfying Assumption 5b, by averaging over all features in S, we have

$$\begin{aligned} X^* = &\arg\max_{X\in F\setminus S}\{I(X;C)+H(S|C)-H(S|C,X) \\ &-\tfrac{1}{|S|}\Sigma_{F_i\in S}\{I(X;F_i)+\Sigma_{F_j\in S, i\neq j}I(X;F_j|F_i)\}. \end{aligned} \tag{43}$$

If we employ Assumption 4 for $I(X;S|C)$ in Equation (43), we get the following objective function in Equation (44) used by the RelaxMRMR algorithm proposed by [52].

$$\begin{aligned} X^* = &\arg\max_{X\in F\setminus S}\{I(X;C)+\Sigma_{F_i\in S}I(X;F_i|C) \\ &-\tfrac{1}{|S|}\Sigma_{F_i\in S}\{I(X;F_i)+\Sigma_{F_j\in S, i\neq j}I(X;F_j|F_i)\}\}. \end{aligned} \tag{44}$$

5.2.4 Time Complexity and Sample Requirement of Non-causal Feature Selection. In this section, we will analyze the time complexity and sample requirement of non-causal feature selection methods. Under the k-DB structural assumption, the most common family of non-causal feature selection methods decompose Equation (21) into different objective functions, such as Equations (29), (31), (36), or (44), in a linear combination of low-order mutual information terms. By these objective functions, non-causal feature selection methods greedily select the ψ features with the highest mutual information scores [22]. The time complexity of non-causal feature selection methods depends on ψ. Solving Equation (44) requires $O(\psi^3 n)$ mutual information computations. Equation (31) and Equation (36) need $O(\psi^2 n)$ pairwise comparisons, while Equation (29) (the MIM criterion) only requires $O(n)$ pairwise comparisons. However, how to determine a good value of the user-defined parameter ψ for optimal feature selection is not an easy problem.

The sample requirement of a non-causal feature selection method depends on the number of samples needed to assure reliable computation of mutual information or independence tests. With discrete data, χ^2 (chi-square) test and G^2 test (a variant of chi-square test) are commonly used to determine the independence of two variables. For a reliable independence test between X and C given the current conditioning set S, the minimum number of data samples N is

$$N \geq \xi \times r_X \times r_C \times r_S, \tag{45}$$

where r_X and r_C represent the numbers of possibles values (i.e., levels) of X and C, respectively, and $r_S = \prod_{i=1}^{|S|} r_{F_i}$, $F_i \in S$, i.e., the multiplication of the numbers of possible values of all features in S. ξ is often set to 5 as suggested by Agresti [1]. As ξ is a constant, the lower bound of the required data samples N is only determined by r_X, r_C, and r_S where r_S plays the key role in Equation (45).

In the article, since we formulate feature selection using mutual information, Equation (46) below shows that the mutual information between two variables is proportional to the value of association of the two variables calculated by G^2 test [58], which guarantees the correctness of using Equation (45) above to discuss the sample requirement of non-causal feature selection methods.

$$\frac{1}{2N}G^2(X;C) = I(X;C) \;\&\; \frac{1}{2N}G^2(X;C|S) = I(X;C|S). \tag{46}$$

To obtain the lower bounds of required samples of the non-causal feature selection methods, assuming that X_{max}, Y_{max}, and W_{max} are the three features with the largest discrete values, the minimum numbers of data samples required by Equation (29) (MIM), Equation (31) (MIFS, mRMR, and FCBF), Equation (36) (JMI and CMIM), and Equation (44) (RelaxMRMR) are bounded by $r_{X_{max}} \times r_C$, $r_{X_{max}} \times r_{Y_{max}}$, $r_{X_{max}} \times r_{Y_{max}} \times r_C$, $r_{X_{max}} \times r_{Y_{max}} \times r_{W_{max}}$, respectively. Since the existing major non-causal feature selection methods calculate $I(X;C|S)$ using linear combination of low-order mutual information terms (i.e., the size of S in r_S in Equation (45) is never bigger than 1), the sample requirement of non-causal feature selection is not high.

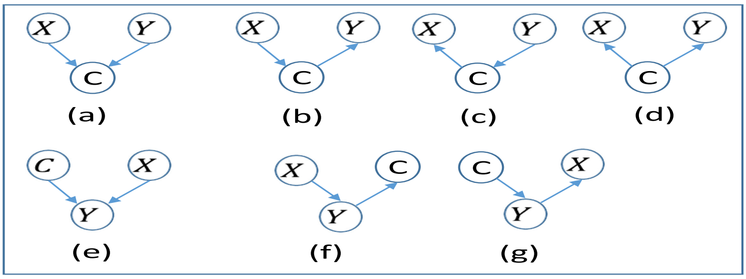


Fig. 5. The three-way causal interactions and non-causal feature selection.

5.2.5 Discussion. Let X be the candidate feature under consideration, and Y a previously selected feature. In Equations (29), (31), (36), and (44), we can see that those methods only consider at most one of the selected features when evaluating X. Therefore, in the following, by representing the interactions among the three variables X, Y, and C (class attribute) using BN structures from Figure 5(a)–(g), firstly, we discuss some properties between X, Y, and C, i.e., Properties 1–4 below. Secondly, with those properties, we will investigate the causal interpretations of Equations (29), (31), (36), and (44). Through the discussion, we will show that the major non-causal feature selection methods driven by the simplified objective functions shown in Equations (29), (31), (36), and (44) prefer direct causes, direct effects, and spouses of C to the features which are not in $MB(C)$. When X and Y are parents or children of C as shown in Figure 5(a)–(d), we have the following properties.

Property 1. *If X and Y are both direct causes (parents) of C, i.e., the class attribute C is a common-effect of the two features, as shown in Figure 5(a), then (1) $I(X;C) > I(X;Y)$, (2) $I(X;Y|C) \geq I(X;Y)$, and (3) $I(X;C) - I(X;Y) + I(X;Y|C) > 0$.*

Proof. According to Proposition 3.10, in the case shown in Figure 5(a), $I(X;Y) = 0$ holds. Clearly, $I(X;C) > I(X;Y)$ if $I(X;Y) = 0$. By $I(X;Y;C) = I(X;Y) - I(X;Y|C) = I(X;C) - I(X;C|Y)$, if $I(X;Y) = 0$, $I(X;Y|C) \geq I(X;Y)$ and $I(X;C) - I(X;Y) + I(X;Y|C) > 0$ hold. □

Property 2. *In the causal chain interaction cases in Figure 5(b) and (c) or the common cause interaction case in Figure 5(d), where X or Y is a direct cause of C, i.e., X, Y and C form a causal chain or X and Y are the common effect of C), (1) $I(X;C) > I(X;Y)$ and (2) $I(X;C) - I(X;Y) + I(X;Y|C) > 0$.*

Proof. From Figure 5(b)–(d), according to the Markov condition in Definition 3.3, $I(X;Y|C) = 0$. By $I(X;Y|C) - I(X;Y) = I(X;C|Y) - I(X;C)$, $I(X;C) > I(X;Y)$ and $I(X;C) - I(X;Y) + I(X;Y|C) > 0$ hold. □

Since $I((Y,X);C) = I(Y;C) + I(X;C|Y)$ and $I(X;Y;C) = I(X;Y) - I(X;Y|C) = I(X;C) - I(X;C|Y)$, i.e., $I(X;C|Y) = I(X;C) - I(X;Y) + I(X;Y|C)$, we have $I((X,Y);C) = I(Y;C) + I(X;C) - I(X;Y) + I(X;Y|C)$, or $I((X,Y);C) - I(Y;C) = I(X;C) - I(X;Y) + I(X;Y|C)$. From Properties 1 and 2 above, we know that if X a direct cause or a direct effect of C, $I(X;C) - I(X;Y) + I(X;Y|C) > 0$, therefore, $I((X,Y);C) - I(Y;C) > 0$, indicating that in the case when X is a direct cause or direct effect of C, X and Y together provide more information about C than Y alone does. When X is a spouse of C as shown in Figure 5(e), we have the following property.

Property 3. *If X is a spouse of C through Y i.e., Y is a child of both X and C, as shown in Figure 5(e), $I(X;Y|C) > I(X;Y)$ and $I(X;C) - I(X;Y) + I(X;Y|C) > 0$.*

Proof. By Proposition 3.10, $I(X;C) = 0$ holds in Figure 1(e). Since $I(X;Y|C) = I(X;Y) + I(X;C|Y) - I(X;C)$, $I(X;Y|C) > I(X;Y)$ and $I(X;C) - I(X;Y) + I(X;Y|C) > 0$ holds. □

Table 2. Non-causal Feature Selection: Objective Functions and Causal Interpretations

Objective function	Representative algorithm	Causal interpretation
Equation (29): $X^* = \arg\max_{X \in F \setminus S} I(X;C)$	MIM	prefer X^* which is a direct cause or direct effect of C
Equation (31): $X^* = \arg\max_{X \in F \setminus S} \{I(X;C)$ $- \sum_{i=1}^{\lvert S \rvert} I(F_i;X)\}$	MIFS, mRMR, FCBF	prefer X^* which is a direct cause or direct effect of C
Equation (36): $X^* = \arg\max_{X \in F \setminus S} \{I(X;C)$ $-\Sigma_{F_i \in S}\{I(X;F_i) - I(X;F_i \vert C)\}\}$	JMI, CIFE, CMIM	prefer X^* which is a direct cause, direct effect, or spouse of C
Equation (44): $X^* = \arg\max_{X \in F \setminus S} \{I(X;C)$ $+\Sigma_{F_i \in S} I(X;F_i \vert C) - \frac{1}{\lvert S \rvert}\Sigma_{F_i \in S}\{I(X;F_i)$ $+\Sigma_{F_j \in S, i \neq j} I(X;F_j \vert F_i)\}\}$	RelaxMRMR	prefer X^* which is a direct cause, direct effect, or spouse of C

Property 3 provides a causal interpretation for the class-conditional relevancy in Equation (36). If X is a spouse of C and Y is the common child of X and C, $I(X;Y|C) - I(Y;X) > 0$. Since $I((X,Y);C) = I(Y;C) + I(X;C) - I(X;Y) + I(X;Y|C)$, then even if $I(X;C) = 0$, $I((X,Y);C)$ provides more information than $I(Y;C)$. This shows that although a spouse of C is not a direct cause or a direct effect of C, from the viewpoint of class-conditional relevancy view, Property 3 confirms that the spouses of C are strongly relevant features.

Property 4. *If Y is a direct cause or a direct effect of C, and X is an indirect cause or an indirect effect of Y, as shown in Figure 5(f)–(g), then (1) $I(X;Y) > I(X;C)$, (2) $I(X;C) + I(X;Y|C) - I(X;Y) = 0$, and (3) $I(Y;C) > I(X;C)$.*

Proof. By the Markov condition, in Figure 5(f)–(g), $I(X;C|Y) = 0$ holds. By $I(X;Y|C) - I(X;Y) = I(X;C|Y) - I(X;C)$, $I(X;Y) \geq I(X;C)$ and $I(X;C) + I(X;Y|C) - I(X;Y) = 0$. Then by $I(Y;C|X) - I(Y;C) = I(X;C|Y) - I(X;C)$, $I(Y;C) - I(X;C) = I(Y;C|X)$. Since $I(Y;C|X) > 0$, $I(Y;C) > I(X;C)$ holds. □

With Properties 1–4, we analyze the causal interpretations of Equations (29), (31), (36), and (44), and our observations are summarized in Table 2. These observations illustrate that the major non-causal feature selection methods prefer direct causes, direct effects, or spouses of C to the features which are not in $MB(C)$ and further confirm that the strongly relevant features belong to $MB(C)$. Specifically, we get the following observations, and these observations will be validated by the experiments in Section 7.1.

- If S is empty, $\forall X \in PC(C)$, i.e., X is a direct cause or effect of C, for any of its ancestors or descendants $F_i \in F \setminus PC(C)$, $I(X;C) > I(F_i;C)$ holds by Property 4. Thus, Equations (29), (31), (36), and (44) will add C's direct causes and effects first to S.
- With Properties 1–4, the term $I(X;C) - I(X;F_i)$ in Equations (29), (31), (36), and (44) prefers direct causes and direct effects of C (i.e., $PC(C)$), while the term $I(X;F_i|C)$ in Equations (36) and (44) prefers spouses of C. Specifically, MIFS, mRMR, and FCBF that are based on or that employ Equation (31) prefer the features $PC(C)$ to be added to S and do not attempt to identify spouses of C, since Properties 1 and 2 state that only when both X and F_i belong to $PC(C)$, $I(X;C) > I(X;F_i)$ holds. Equations (36) and (44) attempt to discover not only $PC(C)$, but also spouses of C, since if X is a spouse of C, there exists a feature F_i, i.e., the common child of C and X, to make $I(X;C) - I(X;F_i) + I(X;F_i|C) > 0$.

- Assuming that currently $S = \{F_i\}$. If $F_i \in ch(C)$, i.e., F_i is a direct effect or a child of C. For two candidate features, $X \in PC(C)$ and W which is a descendant of C and $W \notin ch(C)$, by Property 4, Equations (29), (31), (36), and (44) would prefer X to W. For example, assume that $X \rightarrow C \rightarrow F_i \rightarrow W$, then $I(X;C) - I(X;F_i) + I(X;F_i|C) > 0$ by Property 1 and $I(W;C) - I(W;F_i) + I(W;F_i|C) = 0$. For MIFS and mRMR, $I(X;C) - I(X;F_i) > 0$ while $I(W;C) - I(W;F_i) < 0$, and for FCBF, $I(X;C) > I(X;F_i)$ while $I(F_i;C) > I(W;C)$ and $I(W;F_i) > I(W;C)$. Thus MIFS, mRMR, and FCBF prefer X to W. If $F_i \in pa(C)$, $X \in PC(C)$, W is an ancestor of C and $W \notin pa(C)$ (for example, $W \rightarrow X \rightarrow C \rightarrow F_i$), for X and W, with a similar analysis above, Equations (29), (31), (36), and (44) would add X to S.

5.3 Causal Feature Selection: Assumptions and Approximations

As discussed at the beginning of Section 5 and in the previous sections, non-causal feature selection methods make assumptions on the correlation relationships among predictive features and the class attribute under the naive BN assumptions. Causal feature selection methods do not have such restrictions on the structure of the (causal) BN representing the dependence relationships of all the variables, including the class attribute and all features. However, in order to learn a (causal) BN or the local network structure around the class variable, causal feature selection methods employ the Markov condition (assumption) (Definition 3.3 in Section 3.1), faithfulness assumption (Definition 3.4 in Section 3.1), and causal sufficiency (Definition 3.5 in Section 3.1) for the correctness and causal meaning of the features selected.

Assuming S is the feature set currently selected, $ch(C)$ is the children of C, $Des(C)$ is the descendants of C, and $ND(C)$ is the ancestors of C, by the Markov condition, we can get the following properties.

Property 5. *For an unselected feature $X \in F \setminus S$, if $X \in ND(C) \setminus pa(C)$ and $pa(C) \subseteq S$, X is conditionally independent of C given S, that is, $I(X;C|S) = 0$.*

Property 6. *For an unselected feature $X \in F \setminus S$, if $X \in \{Des(C) \setminus ch(C)\})$ and $pa(X) \subseteq S$, X is conditionally independent of C given S, that is, $I(X;C|S) = 0$.*

With the properties, most existing causal feature selection methods are designed to solve Equation (21) (i.e., maximizing $I(X;C|S)$) with a forward-backward strategy based on the following lemmas.

Lemma 5.1. *$\forall F_i \in PC(C)$ and $\forall S \subseteq F \setminus F_i$, $I(C;F_i|S) > 0$.*

Proof. By Proposition 3.9, $\forall F_i \in PC(C)$ and $\forall S \subseteq F \setminus F_i$, $F_i \not\perp\!\!\!\perp C|S$ holds. By Lemma 4.4, the lemma holds. □

Lemma 5.2. *If F_i is a spouse of C via $F_j \in ch(C)$ (i.e., F_j is a common child of F_i and C), $\exists S \subseteq F \setminus \{F_i, F_j\}$ such that $I(C;F_i|S) = 0$ and $I(C;F_i|F_j \cup S) > 0$.*

Proof. Since C and F_i are not directly connected by an edge, by Proposition 3.9, there must exist a subset S such that C and F_i are independent given S, that is, $I(C;F_i|S) = 0$. By Proposition 3.10, C and F_i are conditionally dependent given any subset containing F_j, i.e., the common child of F_i and C, thus, the lemma is proven. □

In this section, we will analyze the search strategies taken by the existing causal feature selection methods for solving Equation (21). All theorems and lemmas are discussed with the assumption that all independence tests (mutual information calculation) are reliable.

5.3.1 A Simultaneous MB Discovery Strategy by Conditioning on the Entire S for Calculating $I(X;C|S)$ in Equation (21). The simultaneous MB discovery strategy aims to find PC and spouses of C simultaneously without distinguishing PC from spouses during the MB discovery. This approach adopts the forward and backward steps to greedily discover $MB(C)$ for maximizing Equation (21), i.e., sequentially maximizing $I(X;C|S)(X \in F \setminus S)$ at the forward step (max-relevance) and minimizing $I(C;Y|S \setminus Y)(Y \in S)$ at the backward step (min-redundancy) by conditioning on the entire S currently selected. This simultaneous discovery strategy has been employed by two representative algorithms, IAMB and Inter-IAMB [48]. The assumptions and search strategies of IAMB and inter-IAMB are discussed as follows.

IAMB. The forward and backward steps of IAMB for the sequential optimization of Equation (21) are as follows.

- **Forward step.** At each iteration, S is the set of features currently selected, and for each candidate feature within $F \setminus S$, the one satisfying $\arg\max_{X \in F \setminus S} I(X;C|S)$ and $I(X;C|S) > 0$ is added to S. The forward step is terminated until $\forall X \in F \setminus S, I(C;X|S) = 0$.
- **Backward step.** IAMB sequentially removes from S the false positive $Y \in S$ satisfying $I(C;Y|S \setminus Y) = 0$ until $\forall Y \in S, I(C;Y|S \setminus Y) > 0$.

The forward step will add all features in the true $MB(C)$ to S. Due to the greedily strategy, some false positives may enter S at the forward step. For example, assuming $X \notin MB(C)$ and $\exists Y \in MB(C)$ such that $I(X;C|S \cup Y) = 0$. However, when checking $I(X;C|S)$ and at this time $Y \notin S$, $I(C,X|S) > 0$ holds and X will be added to S. Thus, the backward step will remove all the false positives in S by Properties 5 and 6.

Theorem 5.3. *The output of IAMB is the optimal set S^* in Equation (12).*

Proof. Assuming $\overline{S}$ denotes the set $F \setminus S$. At the forward step, at each iteration, $X \in F \setminus S$ is selected that satisfies Equation (47) below.

$$X^* = \arg\max_{X \in F \setminus S} \{I(S;C) + I(C;X|S)\}. \tag{47}$$

At each iteration, for all $X \in F \setminus S$, $I(S;C)$ in Equation (47) is the same. For the IAMB algorithm, by Equation (47), at each iteration, maximizing $I(C;X|S)$ is equivalent to maximizing $I((S,X);C)$. By $I(C;F) = I(C;S) + I(C;\overline{S}|S)$, when $I(C;\overline{S}|S) = 0$, then $I(C;S)$ is maximized. At the forward step, IAMB greedily maximizes $I(C;X|S)$ until $\forall X \in F \setminus S, I(C;X|S) = 0$. Then by Lemma 5.1, all PC of C ($PC(C)$) will be gradually added to S, while by Properties 5 and 6, the ancestors and descendants of C may not be added to S. Let the set $SP(C)$ include all spouses of C, when all PC of C are added to S, by Lemma 5.2, $\forall X \in SP(C), I(X;C|S) > 0$, and thus all spouses of C will be added to S initially during the forward step. In any case, at the end of the forward step, all features in the true $MB(C)$ will have been added to S.

At the backward step, $\exists Y^* \in S$ to be removed from S satisfies

$$Y^* = \arg\min_{Y \in S} I(Y;C|S \setminus Y). \tag{48}$$

By Equation (48), at each iteration, if $I(C;Y|S \setminus Y) = 0$, IAMB will remove Y from S until given any feature $Y \in F \setminus S$, $I(Y;C|S \setminus Y) > 0$. Then all false positives in S are removed, and thus $S = MB(C)$. By Theorem 4.5, the theorem is proved. □

Inter-IAMB. IAMB suffers from the problem of the addition of false positives to S at the forward step, then makes the size of S possibly become high-dimensional. The Inter-IAMB strategy mitigates the problem by interleaving the forward and backward steps of IAMB to keep S as small as possible, then maximizes $I(C;X|S)$ for $X \in F \setminus S$ and minimizes $I(C;Y|S \setminus Y)$ for $Y \in S$ simultaneously.

Theorem 5.4. *The output of Inter-IAMB is the optimal set S^* in Equation (12).*

Proof. At each iteration, by Equation (47), the forward step adds a new feature $X \in F \setminus S$ that maximizes $I(C;X|S)$ to S. Once the new feature X is added to S, the backward step is triggered immediately and removes features in S (false positives) that minimize Equation (48). By maximizing $I(C;X|S)$ and minimizing $I(Y;C|S \setminus Y)$ simultaneously, the strategy will converge until $\forall X \in F \setminus S$, $I(X;C|S) = 0$ and $\forall Y \in S$, $I(Y;C|\{S \setminus Y\}) > 0$. After the backward step, $S = MB(C)$. Then by Theorem 4.5, the theorem is proved. □

The time complexity of IAMB and Inter-IAMB above is measured in the number of conditional independence tests (association computations) executed. For IAMB and Inter-IAMB, the average time complexity is $O(n|S|)$ and the worst time complexity is $O(n^2)$ where n is the total number of features and in the worst case with $|S| = n$.

Comparing to non-causal feature selection, IAMB and Inter-IAMB both use the entire set of S as the conditioning set for the calculation of $I(X;C|S)$ at each iteration. By Equation (45) in Section 5.2.4, assuming S_{max} is the largest conditioning set during MB search, thus the minimum number of data samples N required by IAMB and Inter-IAMB is $r_{X_{max}} \times r_C \times r_{S_{max}}$. Then the number of data instances required by IAMB and Inter-IAMB will increase exponentially in the size of S. To mitigate this drawback, in the next section, we will discuss a divide-and-conquer strategy.

5.3.2 A Divide-and-conquer Strategy by Conditioning on All Subsets of S for Calculating $I(X;C|S)$ in Equation (21). The main idea behind a divide-and-conquer strategy is that: (1) finding $PC(C)$ and $SP(C)$ separately, and (2) using a feature-subset enumeration strategy to explore subsets of S for discovering $PC(C)$ instead of conditioning on the entire set of S. That is, to calculate $I(C;X|S)$, the divide-and-conquer strategy performs a search for a subset, $S' \subseteq S$ such that if X and C are conditional independent given S', i.e., $I(C;X|S') = 0$, X will not be added to S and will never be considered as a candidate feature again. Then, the minimum number of data samples N required by the divide-and-conquer strategy is $r_{X_{max}} \times r_C \times r_{S'}$ where $0 \leq |S'| \leq |S_{max}|$. Accordingly, on average, the divide-and-conquer strategy requires much smaller number of data samples than IAMB and Inter-IAMB. Specifically, the divide-and-conquer strategy mainly consists of the following two steps for solving Equation (21).

- Discovering $PC(C)$. At each iteration, assuming S is the set of features currently selected, for each candidate feature $X \in F \setminus S$, if $\exists S' \subseteq S$ such that X and C are conditional independent given S', i.e., $I(X;C|S') = 0$, X is discarded and will never be considered as a candidate parent or child of C again, otherwise X is added to S. By Lemma 5.1, after this step, all PC will be added to S.
- Discovering $SP(C)$. By Lemmas 5.1 and 5.2, $\forall X \in SP(C)$, there must exist a subset in $F \setminus X$ such that X and C are conditional independent given this subset. Therefore, all spouses of C cannot be added to S at the PC discovery step. To find $SP(C)$, by Lemma 5.2, $\forall X \in S$, the step employs the PC discovery step to find $PC(X)$, then for each feature $Y \in PC(X)$, if $\exists S' \subseteq F \setminus \{Y, X\}$ such that $I(C;Y|S') = 0$ and $I(C;Y|S' \cup X) > 0$, $Y \in SP(C)$.

There are four representative approaches to instantiate the divide-and-conquer strategy, i.e., max-min heuristic, simple max-heuristic, backward heuristic, and k-greedy heuristic. The representative algorithms include MMMB [47], HITON-MB [3], IPC-MB [18], and STMB [20].

1. The max-min heuristic. The representative algorithm using the strategy is the MMMB algorithm, which includes the following two steps.

(1) **Discovering $PC(C)$ step.** This step includes a forward step and a backward step to find $PC(C)$. To select the feature $X^* \in F \setminus S$ to maximize $I(C;X|S)$, the forward and backward steps are implemented as follows.

—Forward step. The max-min heuristic selects the feature that maximizes the minimum correlation with C conditioned on the subsets of S. Specifically, initially S is an empty set, $\forall X \in F \setminus S$, the minimum correlation, denoted as $corr(C;X|S)$, between C and X conditioning on all possible subsets of S, is calculated as Equation (49) below.

$$corr(C;X|S) = \min_{S' \subseteq S} I(C;X|S'). \tag{49}$$

$X^* \in F \setminus S$ will be added to S if $corr(C;X^*|S) > 0$ and Equation (50) below holds.

$$X^* = \arg\max_{X \in \{F \setminus S\}} corr(C;X|S). \tag{50}$$

The forward step stops until $\forall X \in F \setminus S$, $corr(C;X|S) = 0$.

—Backward step. Each feature in S selected at the forward step will be checked. If $\exists Y \in S$ satisfies Equation (51) below, it will be removed from S and never considered again.

$$\exists S' \subseteq S \setminus Y, I(C;Y|S') = 0. \tag{51}$$

(2) **Discovering $SP(C)$ step.** At the step, the max-min heuristic firstly finds the set of PC for each feature in S found at the forward step. Assuming $X \in PC(C)$ and $Y \in PC(X)$, if $Y \notin PC(C)$ and $\exists S' \subset F \setminus \{X, Y\}$ to make Equation (52) below hold, then Y is a spouse of C.

$$I(C;Y|S') = 0 \; and \; I(C;Y|S' \cup X) > 0. \tag{52}$$

2. Interleaving max-heuristic. The main difference between the max-heuristic and the interleaving max-heuristic is that in the discovering $PC(C)$ step, the interleaving max-heuristic interleaves the forward and backward steps to keep the size of S as small as possible. The representative algorithm using the strategy is the HITON-MB algorithm.

(1) Discovering $PC(C)$ step. In the step, the interleaving max-heuristic uses a simpler forward strategy than the max-min heuristic. Before interleaving forward and backward steps, $\forall X \in F$, the interleaving max-heuristic computes $I(C;X)$ and adds the features that satisfy $I(C;X) > 0$ to the candidate $PC(C)$ set, called $SPC(C)$, in a descending order according to the value of $I(C;X)$. If $I(C;X) = 0$, X will be discarded and never considered as a candidate parent or child again. Then, initially S is an empty set, and for each feature in $SPC(C)$, this strategy interleaves Equations (53) and (54) as follows, until $SPC(C)$ is empty.

—Forward step. $\forall X \in SPC(C)$, if X satisfies Equation (53) below, it will be added to S.

$$X^* = \arg \max_{X \in SPC(C)} I(C;X). \tag{53}$$

—Backward step. Once X is added to S at the forward step, the backward step is triggered. Specifically, $SPC(C) = SPC(C) \setminus X$, and $\forall Y \in S$, if $\exists S' \subseteq S \setminus Y$ satisfies Equation (54) below, Y will be removed from S and never considered again.

$$I(C;Y|S') = 0. \tag{54}$$

(2) Finding spouses. The step is the same as the max-min heuristic in Equation (52).

Comparing to the simultaneous discovery strategy to discover MBs in Section 5.2.1, the strategies in this section perform a subset search within S instead of conditioning on the entire S. Thus,

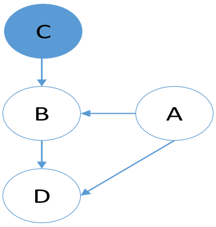


Fig. 6. An example of the false positive D being added to S in the discovering $PC(C)$ step using the max-min heuristic or its interleaving version.

for the max-min heuristic and its interleaving version, the time complexity is $O(n|S|^2 2^{|S|})$ where $|S|$ denotes the largest size of S during forward and backward steps.

Theorem 5.5. *Using the max-min heuristic or its interleaving version, in the discovering $PC(C)$ step, $PC(C) \subseteq S$ [2, 49].*

Theorem 5 states that in addition to $PC(C)$, the output of the discovering $PC(C)$ step, i.e., S, may include some false positives. For example, in Figure 6, assuming C is the target feature, B, A, and D is a child, spouse, and descendant of C, respectively, D will enter and remain in S in the discovering $PC(C)$ step [2]. The explanation is as follows. C and D are dependent conditioning on the empty set, since the path $C \rightarrow B \rightarrow D$ d-connects C and D. By conditioning on B, the path $C \rightarrow B \leftarrow A \rightarrow D$ d-connects C and D by Definition 3.6.

To remove false positives from S, such as D, the two max-min heuristics employ a symmetry correction. The idea behind the symmetry correction is that in a BN, if $X \in PC(C)$, then $C \in PC(X)$. With the symmetry correction, in the discovering $PC(C)$ step, the work in [36, 49] proved that $S = PC(C)$. And with symmetry corrections, the work in [2] proved that the output of the two max-min heuristics is $MB(C)$, that is, $S = MB(C)$, and thus Theorem 5.6 below holds.

Theorem 5.6. *The output of the max-min heuristic (and its interleaving version) employed by MMMB (and HITON-MB) is the optimal set S^* in Equation (12) with symmetry correction.*

3. The backward strategy. The IPCMB [18] and STMB [20] algorithms only employ a backward step to discover $PC(C)$ instead of using a forward-backward strategy. Initially, by setting $S = F$, the backward step removes features from S one by one, instead of greedily adding features to S one by one for maximizing $I(C;S)$. Specifically, in the discovering $PC(C)$ step, for $\forall Y \in S$, if $\exists S' \subseteq S \setminus Y$ and $|S'| = 0$ (i.e., the size of S' equals to 0) such that $I(Y;C|S') = 0$, Y is removed from S. Otherwise, if $\exists S' \subseteq S \setminus Y$ and $|S'| = 1$ such that $I(Y;C|S') = 0$, Y is removed from S. The backward step continues in this way by performing level by level of the size of S', until the size of the current S' is larger than the size of the current S.

This backward strategy employed by IPCMB also finds a superset of $PC(C)$, that is, $PC(C) \subseteq S$. Thus, IPCMB embeds a symmetry correction in the spouse discovery stage to remove false positives in S. To find spouses, IPCMB adopts the same idea with MMMB and HITON-MB.

STMB also employs the backward step to discover $PC(C)$. But STMB has two main differences against IPCMB. Firstly, STMB finds $SP(C)$ in $F \setminus S$, instead of PC of each feature in S. Secondly, STMB uses the found spouses to remove false positives in S found in the discovering $PC(C)$ step instead of using a symmetry correction during the $SP(C)$ discovery step. Specially, assuming S found in the discovering $PC(C)$ step and $SP(C) = \emptyset$, the idea of discovering spouses is summarized below.

- Finding spouses and removing false PC from S: for each feature $X \in F \setminus S$, if $\exists Y \in S$ and $\exists S' \subset F \setminus \{X \cup Y\}$ s.t. $I(C;X|S') = 0$ and $I(C;X|S' \cup Y) > 0$, then X is added to $SP(C)$. Once X is added to $SP(C)$, for each feature $Y \in S$, if $\exists S' \subseteq \{S \cup X\} \setminus Y$ s.t. $I(C;Y|S') = 0$, then Y and X are removed from S and $SP(C)$, respectively. The process terminates until all features in $F \setminus S$ are checked.
- Removing false positives from $SP(C)$ and S: (1) $\forall X \in SP(C)$, if $I(X;C|S \cup SP(C) \setminus X) = 0$, X is removed from $SP(C)$; then (2) $\forall Y \in S$, if $I(Y;C|S \cup SP(C) \setminus Y) = 0$, Y is removed from S.

For the output of IPCMB and STMB, it has been proved that $\{S \cup SP(C)\} = MB(C)$ [18, 20]. Thus, IPCMB and STMB greedily find the optimal set S^* in Equation (12). The time complexity of IPCMB includes finding both $PC(C)$ and $SP(C)$, then the complexity is $O(n2^{|S|} + |S|n2^{|S|}) = O(|S|n2^{|S|})$ where $|S|$ is the largest size of conditional set during search. The worst time complexity of IPC-MB is $O(n2^n + n^2 2^{|S|}) = O(n^2 2^{|S|})$ when all features are PC of C. For STMB, and the average time complexity is $O(n2^{|S|} + |S||F \setminus S|2^{|S|}) = O(|S||F \setminus S|2^{|S|})$, and the worst time complexity is $O(n2^n + n^2 2^{|S|}) = O(n^2 2^{|S|})$.

4. γ-greedy heuristic. In the discovering $PC(C)$ step, as the size of S becomes large, it will be computationally expensive or prohibitive when we perform an exhaustive enumeration over all subsets of S. For example, to check whether X is able to be added to S, in the worst case, the total number of subsets checked is up to $2^{|S|}$. Accordingly, in the discovering $PC(C)$ step, MMMB, HITON-MB, IPCMB, and STMB employ a γ-greedy search method to mitigate this problem. The γ-greedy search checks all subsets of size less than or equal to a user-defined parameter γ ($0 \leq \gamma < |S|$), that is, the maximum size of subsets needed to be checked. In the case of using the γ-greedy heuristic, MMMB, HITON-MB, IPCMB, and STMB return an approximate $MB(C)$ [2].

5.4 Practical Implication

In Sections 5.2 and 5.3, we discussed the BN structural assumptions and analyzed in detail how the assumptions led to the different levels of approximations employed by causal and non-causal feature selection methods for the calculation of $I(X;C|S)$. With the structural assumptions, we are able to fill in the gap in our understanding of the relation between the two types of feature selection methods.

Firstly, the feature sets obtained by causal feature selection methods are closer to $MB(C)$ than those obtained by non-causal feature selection methods. However, our analysis in Sections 5.2 and 5.3 shows that non-causal feature selection methods are much more computationally efficient and have lower sample requirement than causal feature selection methods. The choice of causal or non-causal feature selection methods depends on the size of the dataset under study.

Secondly, the strongly relevant features are the same as the MB of C. This may motivate us to leverage the advantages of both causal and non-causal feature selection methods to develop more efficient and robust new feature selection methods.

Thirdly, causal and non-causal feature selection methods implicitly reduce a full BN classifier to a selective BN classifier by selecting a subset of features S to make the conditional likelihood $P(C|S)$ as close to $P(C|F)$ as possible, as shown in Figure 7.

6 ERROR BOUNDS

In the section, we will discuss the error bounds of non-causal and causal feature selection for understanding the impact of assumptions and approximations made by the two types of methods on classification performance. Since both types of methods are independent of any classifiers, we will analyze the bounds of difference in the information gains between an approximate MB and an exact MB. In Section 3, Equation (17) has presented that if a subset S in D maximizing $I(C;S)$,

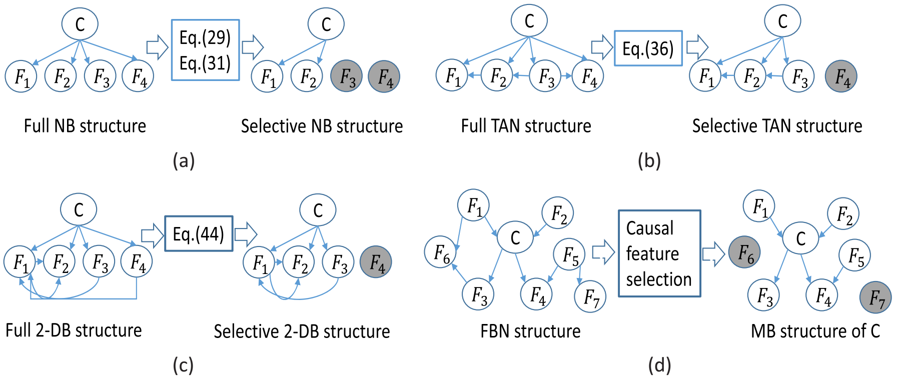


Fig. 7. (a) A NB to a selective NB, (b) a TAN to a selective TAN, (c) a 2-DB to selective 2-DB, and (d) a general BN to a MB-based BN classifier.

then S also maximizes $L(C|S,D)$ and minimizes P_{err}. Based on Equation (17), using information gain, in the following, we will discuss the bounds of the difference between an approximate MB and an exact MB.

6.0.1 Conditioning on the Full S and Its All Subsets (Exact MB Discovery). According to our analysis in Section 5.3, under certain assumptions, causal feature selection algorithms designed with conditioning on both the full S and all of its subsets can find the exact $MB(C)$ from data. Moreover, the algorithms by conditioning on all subsets of S are also able to find the exact $PC(C)$. As $I(C;F \setminus MB(C)|MB(C)) = 0$, $I(C;F) = I(C;MB(C))$. Since $H(P_{err})^{-1} \leq P_{err} \leq 1/2H(C|F)$ (see Equation (14)), Theorem 6.1 gives the minimum upper bound of P_{err}.

Theorem 6.1. $P_{err} \leq 1/2H(C|MB(C))$.

Proof. By Theorem 4.5, $\forall S \subseteq F$, $I(C;MB(C)) \geq I(C;S)$ holds. Since $H(C|MB(C)) = H(C) - I(C;MB(C))$ and $H(C|S) = H(C) - I(C;S)$, we get that $\forall S \subseteq F$, $H(C|MB(C)) \leq H(C|S)$. By Equation (15), the theorem is proven. □

By Equation (16), $\lim_{m\to\infty} -\ell(C|S,D) = KL(p(C|S)||q(C|S)) + H(C|S)$. As $m \to \infty$, $KL(p(C|S)||q(C|S))$ will approach zero, and thus Equation (55) presents that $H(C|MB(C))$ minimizes $-\ell(C|MB(C),D)$.

$$\lim_{m\to\infty} -\ell(C|MB(C),D) \approx H(C|MB(C)). \tag{55}$$

If $S = PC(C)$ holds, by Theorem 4.5, $I(MB(C);C) \geq I(PC(C);C)$ holds. Thus, $H(C|MB(C)) \leq H(C|PC(C))$ holds, and Equation (56) below gives the Bayes error rates of $PC(C)$, that is, $P_{err}(PC(C))$. Since $H(C|MB(C)) \leq H(C|PC(C))$, the upper bound in Equation (56) is looser than that in Equation (55).

$$P_{err}(PC(C)) \leq 1/2H(C|PC(C)). \tag{56}$$

Since $\lim_{m\to\infty} -\ell(C|PC(C),D) \approx H(C|PC(C))$, Equation (57) below gives the upper bound of the conditional log-likelihood of $PC(C)$ in D, that is, $-H(C|MB(C))$.

$$\lim_{m\to\infty} \ell(C|PC(C),D) \leq -H(C|MB(C)). \tag{57}$$

6.0.2 Conditioning on the Subsets of S Up to Size γ (Causal Feature Selection). As we discussed in Section 5.3, the γ-greedy search employed by causal feature selection methods may return an approximate $MB(C)$. Let $AMB(C) \subseteq F$ be an approximate MB of C, by Theorem 4.5, $H(C|MB(C)) \leq$

$H(C|AMB(C))$ holds. Since Theorem 6.1 illustrates that $1/2H(C|MB(C))$ is the minimum upper bound of P_{err}, thus, we get

$$P_{err}(AMB(C)) \leq 1/2H(C|MB(C)). \tag{58}$$

Since $\lim_{m\to\infty} -\ell(C|AMB(C), D) \approx H(C|AMB(C))$ holds, the upper bound of the conditional log-likelihood of any approximate MB of C is

$$\lim_{m\to\infty} \ell(C|AMB(C), D) \leq -H(C|MB(C)). \tag{59}$$

6.0.3 Conditioning on the Subset of Size 0 or 1 (Non-causal Feature Selection). As discussed in Section 5.2, non-causal feature selection algorithms attempt to find $PC(C)$ and some spouses of C. With different values of ψ (i.e., the number of selected features), those strategies may return an approximate $MB(C)$, that is, a superset or a subset of $PC(C)$. In the following, we will focus on discussing the bounds of the superset or subset of $PC(C)$ found by non-causal feature selection methods.

Corollary 6.2. *If* $S1 \subseteq F \setminus PC(C)$ *and* $S = PC(C) \cup S1$,

(1) $-H(C|PC(C)) \leq \lim_{m\to\infty} \ell(C|S, D) \leq -H(C|MB(C))$;
(2) $1/2H(C|MB(C)) \leq P_{err}(S) \leq 1/2H(C|PC(C))$.

Proof. Assuming $\overline{PC(C)} = F \setminus PC(C)$ and $\overline{S} = F \setminus S$. Firstly, we prove that $I(C; \overline{PC(C)}|PC(C)) \geq I(C; \overline{S}|S)$ holds. By $I(C; F) = I(PC(C); C) + I(C; \overline{PC(C)}|PC(C))$, we get

$$\begin{aligned} I(C; F) &= I((\overline{S}, S); C) \\ &= I(S; C) + I(C; \overline{S}|S) \\ &= I((PC(C), S1); C) + I(C; \overline{S}|S) \\ &= I(PC(C); C) + I(S1; C|PC(C)) + I(C; \overline{S}|S). \end{aligned} \tag{60}$$

By the chain rule of mutual information, we can get

$$I(S1; C|PC(C)) = \sum_{j=1}^{|S1|} I(F_j; C|F_{j-1}, \ldots, F_1, PC(C)). \tag{61}$$

Since $S1$ only includes spouses, non-descendants and descendants of C. By Equations (60) and (61), we get the following.

Case 1: if $\exists F_j \in S1$ is a descendant of C and $I(F_j; C|F_{j-1}, \ldots, F_1, PC(C)) > 0$, then $I(C; \overline{PC(C)}|PC(C)) > I(C; \overline{S}|S)$ holds.

Case 2: if $\exists F_j \in S1$ and F_j is a spouse of C, then $I(F_j; C|F_{j-1}, \ldots, F_1, PC(C)) > 0$. Thus, $I(C; \overline{PC(C)}|PC(C)) > I(C; \overline{S}|S)$ holds.

Case 3: if $F_j \in S1$ is a non-descendant of C, by the Markov condition,
$I(F_j; C|F_{j-1}, \ldots, F_1, PC(C)) = 0$, then $I(C; \overline{PC(C)}|PC(C)) = I(C; \overline{S}|S)$.

By $I(C; S) \leq I(C; MB(C))$, $I(C; \overline{S}|S) \geq I(C; \overline{MB(C)}|MB(C))$ holds. Then we get

$$I(C; \overline{PC}|PC(C)) \geq I(C; \overline{S}|S) \geq I(C; \overline{MB(C)}|MB(C)).$$

Then $I(C; PC(C)) \leq I(C; S) \leq I(C; MB(C))$ holds. Thus, we get $H(C|PC(C)) \geq H(C|S) \geq H(C|MB(C))$. Thus, (1) and (2) hold. □

For a subset of $PC(C)$, assuming $S \subset PC(C)$, $\overline{S} = F \setminus S$. If $PC(C) = \{S \cup S'\}$, $I(C; PC(C)) = I(C; S) + I(C, S'|S)$. Since $I(C; S'|S) = \sum_{i=1}^{|S'|} I(F_i; C|F_{i-1}, \ldots, F_1, S)$ holds and $S' \subset PC(C))$, then $I(C; S'|S) > 0$. Thus, $I(C; PC(C)) > I(C; S)$ holds. By $I(C; F) = I(C; PC(C)) + I(C; \overline{PC}|PC(C))$ and

$I(C;F) = I(C;S) + I(C;\overline{S}|S)$, then $I(C;\overline{PC}|PC(C)) < I(C;\overline{S}|S)$. Accordingly, we can get the bounds between $S \subset PC(C)$ and $PC(C)$ in the following:

$$\lim_{m\to\infty} \ell(C|S,D) < -H(C|PC(C)) \text{ } and \text{ } P_{err}(S) < 1/2H(C|PC(C)). \tag{62}$$

By the analysis above, we can see that the errors of causal and non-causal feature selection methods are bounded by $1/2H(C|MB(C))$ and $1/2H(C|PC(C))$, respectively. This indicates that the error bound of non-causal feature selection is looser than that of causal feature selection. Therefore, referring back to Figure 4, our analysis in this section validates that as causal feature selection methods make no assumption on the structure of the BN representing the dependency of variables, their search strategies are able to find the exact $MB(C)$, while the strong assumptions made by non-causal feature selection methods lead to an approximate $MB(C)$ (referring back to Figure 3).

7 EXPERIMENTS

In this section, we will conduct extensive experiments to validate our findings of causal and non-causal feature selection, with the following focuses:

- In Section 7.1, we validate Theorem 4.5 in Section 4.2 (i.e., the MB of C is the optimal set for feature selection), the discussion in Section 5.2.5 (causal interpretations of non-causal feature selection), and the proposed error bounds in Section 6 using a set of synthetic data sampled from a benchmark BN.
- In Section 7.2, as seen in the experiment results, we investigate the impact of different levels of approximations made by causal and non-causal feature selection methods on classification performance, the computational and accuracy performance of causal and non-causal feature selection methods, and the impact of data sample sizes on both methods using 25 various types of real-world datasets, including 6 datasets with large data samples, 6 datasets with extreme small samples, 7 datasets with multiple classes, and 6 class-imbalanced datasets.

To carry out these validations, the following eight representative feature selection methods are selected.

- Five representative causal feature selection methods, including three MB algorithms, IAMB, HITON-MB, and MMMB, and two PC algorithms, HITON-PC and MMPC, we use the implementations of these algorithms obtained from https://github.com/mensxmachina/CausalExplorer_1.5.[2]
- Three representative non-causal feature selection algorithms: mRMR, JMI, and CMIM since these three algorithms provides better tradeoff in terms of accuracy and scalability than the other non-causal feature selection algorithms (especially with small-sized data samples) [11]. We use the implementations of mRMR, JMI, and CMIM obtained from https://github.com/Craigacp/FEAST.

To evaluate the features selected by each algorithm for classification, in all experiments, we use Naive Bayes classifier (NBC), k-Nearest Neighbor (KNN) classifier, and Support Vector Machine (SVM). All experiments were performed on a Window 7 Dell workstation with an Intel(R) Core(TM) i5-4570, 3.20 GHz processor and 8.0 GB RAM, and all eight feature selection methods under comparison are implemented in MATLAB, and NBC, KNN, and SVM are implemented in MATLAB2014 Statistics Toolbox. In the tables in Section 7, the notation “$A \pm B$” denotes that “A”

[2] The source codes of the state-of-the-art causal feature selection methods are also available at https://github.com/kuiy including C++, Matlab, and Python versions.

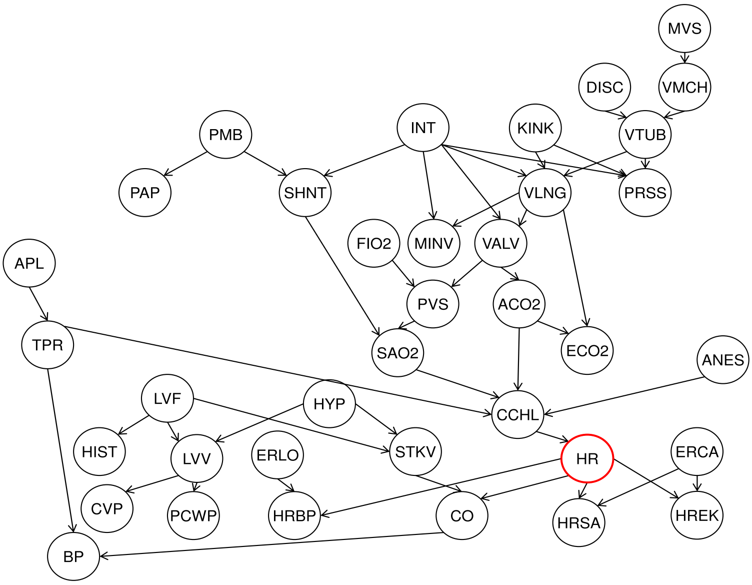


Fig. 8. The ALARM BN.

is the average performance of an algorithm on a dataset, while "B" represents the corresponding standard deviations of the average performance. Due to space limit, for more experiments on real-world datasets, please see the supplement.

7.1 Experiments Using Synthetic Datasets

In this section, we will validate $MB(C)$ is the optimal set for feature selection (Theorem 4.5 in Section 4.2) and the proposed error bounds in Section 6 using a set of synthetic data sampled from A Logical Alarm Reduction Mechanism (ALARM) network, a benchmark and well-known BN modelling an alarm message system for patient monitoring [7]. This network includes 37 variables and the complete structure of the network is shown in Figure 8. Since the MB of each variable can be read from the network, we are able to evaluate the performance of the feature selection methods against the true MBs.

In the ALARM network, we choose the "HR" (Heart Rate) variable as the class attribute for classification. The variable takes three class labels, "low," "normal," and "high," and has the largest MB among all variables, including one parent, four children, and three spouses. We randomly sampled 10 training datasets with 5,000 training cases (large-sized data samples) and 50 training cases (small-sized data samples), respectively. For each training dataset, we randomly sampled a testing dataset with 1,000 testing cases. The reported prediction accuracy is the average accuracy of a classifier using the feature sets selected over the 10 runs of a feature selection method on these 10 training datasets.

In all tables in Section 7.1, "TruePC" and "TrueMB" denote the ground-truths of PC and MB of "HR" in the network, respectively. For validating the discussion of causal interpretations of non-causal feature selection methods in Section 5.2.5, we use the following settings and evaluation metrics.

- We set the parameter ψ, i.e., the numbers of features selected by mRMR, JMI, and CMIM to the size of the true MB (or the true PC) of "HR" in the network, which is denoted as "ψ=nMB" (or "ψ=nPC").

Table 3. Prediction Accuracy of True MB Against Causal and Non-causal Algorithms

Algorithm	5,000 cases		50 cases	
	KNN	NBC	KNN	NBC
TrueMB	**98.99±0.3542**	**98.52±0.3765**	**98.99±0.3542**	**97.27±0.3889**
IAMB	98.65±0.3719	98.34±0.5420	94.67±3.8500	95.06±3.8200
MMMB	**98.99±0.3542**	**98.52±0.3765**	95.02±1.3456	95.13±1.3516
HITON-MB	**98.99±0.3542**	**98.52±0.3765**	95.02±1.3456	95.13±1.3516
mRMR (ψ = nMB)	98.38±0.2898	98.42±0.3360	95.40±1.7321	95.94±1.4152
CMIM (ψ = nMB)	98.17±0.3653	98.38±0.4367	93.25±2.5238	93.49±1.8064
JMI (ψ = nMB)	98.67±0.4523	98.33±0.3889	94.83±1.3259	95.50±1.3968

Table 4. Precision and Recall of Each Algorithm for MB Discovery

Algorithm	5,000 cases		50 cases	
	precision	recall	precision	recall
IAMB	1±0	0.6875±0.0659	1±0	0.1250±0
MMMB	1±0	1±0	0.6885±0.1282	0.9000±0.0791
HITON-MB	1±0	1±0	0.6500±0.0791	0.6500±0.0791
MRMR (ψ = nMB)	0.6250±0	0.6250±0	0.6500±0.0527	0.6500±0.0527
CMIM (ψ = nMB)	0.6250±0	0.6250±0	0.4875±0.1905	0.4875±0.1905
JMI (ψ = nMB)	0.7500±0	0.7500±0	0.6500±0.0791	0.6500±0.0791

- We evaluate the feature sets selected by each algorithm using the precision and recall metrics on the 10 training datasets to observe the percentage of the MB (or the direct causes and direct effects) of "HR" included in the selected features (the output) of each algorithm. The precision metric is the number of true positives in the output (i.e., the variables in the output belonging to the true MB (or PC) of "HR" in the ALARM network) divided by the number of variables in the output of an algorithm. The recall metric is the number of true positives in the output divided by the number of true positives (the number of the true MB (or PC) of "HR" in the alarm network).
- We use the prediction accuracy, the ratio between the number of correct predictions and the total number of testing data samples to validate Theorem 4.5 presented in Section 4.2 and the error bounds proposed in Section 6.

7.1.1 Validation of Theorem 4.5 and the Discussion in Section 5.2.5. In this section, we will validate Theorem 4.5 (i.e., the MB of C is the optimal set for feature selection), and the discussion of causal interpretations of non-causal feature selection in Section 5.2.5.

Validating Theorem 4.5. Table 3 reports the average prediction accuracies and standard deviations using the datasets containing 5,000 and 50 training cases, respectively. Table 3 states that using both KNN and NBC, the true MB of "HR" achieves the highest prediction accuracy. Table 5 shows the number of PC, spouses (SP), and false positives (FP) in the found feature set of each algorithm. These results indicate that classifiers using $MB(C)$ as the feature set achieve the best classifications results, which validates Theorem 4.5 in Section 4.2.

From Tables 4 and 5, we can see that using 5,000 cases, both MMMB and HITON-MB find the exact MB of "HR," and thus get the same prediction accuracy as the true MB of "HR," while the other four algorithms do not find the exact MB of "HR." In addition, from Table 5, we can see that except for IAMB, all feature sets found by the other five algorithms include all variables within the

Table 5. Number of PC, Spouses (SP), and False Positives (FP)

Algorithm	5000 cases			50 cases		
	PC	SP	FP	PC	SP	FP
IAMB	4.3±0.4830	1.2±0.4216	0	1±0	0	0
MMMB	5±0	3±0	0	4.9±0.3162	2.3±0.6749	3.5±1.7159
HITON-MB	5±0	3±0	0	4.9±0.3162	0.3±0.4830	2.8±0.6325
mRMR (ψ = nMB)	5±0	0	3±0	4.9±0.3162	0.3±0.4830	2.8±0.4216
CMIM (ψ = nMB)	5±0	0	3±0	3.9±1.5239	0	3.1±1.1972
JMI (ψ = nMB)	5±0	1±0	2±0	4.9±0.3162	0.3±0.4830	2.8±0.6325

Table 6. Prediction Accuracy of True PC Against Causal and Non-causal Algorithms

Algorithm	5,000 cases		50 cases	
	KNN	NBC	KNN	NBC
TruePC	**98.44±0.3596**	**98.57±0.3199**	**98.44±0.3596**	**97.36±0.3718**
MMPC	**98.44±0.3596**	**98.57±0.3199**	96.66±0.7382	97.01±1.2360
HITON-PC	**98.44±0.3596**	**98.57±0.3199**	96.66±0.7382	97.01±1.2360
mRMR (ψ = nPC)	**98.44±0.3596**	**98.57±0.3199**	96.28±1.2017	96.44±1.0710
CMIM (ψ = nPC)	**98.44±0.3596**	**98.57±0.3199**	95.77±1.7601	96.31±0.8850
JMI (ψ = nPC)	**98.44±0.3596**	**98.57±0.3199**	94.08±3.1435	94.59±1.5975

Table 7. Precision and Recall of Each Algorithm for PC Discovery

Algorithm	5,000 cases		50 cases	
	precision	recall	precision	recall
MMPC	1±0	1±0	0.9714±0.0904	0.9200±0.1033
HITON-PC	1±0	1±0	0.9714±0.0904	0.9200± 0.1033
MRMR (ψ = nPC)	1±0	1±0	0.8600±0.0966	0.8600±0.0966
CMIM (ψ = nPC)	1±0	1±0	0.5800±0.1989	0.5800±0.1989
JMI (ψ = nPC)	1±0	1±0	0.8000±0.0943	0.8000±0.0943

PC set of "HR." This explains why the IAMB, mRMR, JMI, and CMIM are very competitive on the prediction accuracy. CMIM and mRMR cannot find any spouses using 5,000 cases.

Using 50 cases, in Table 3, mRMR gets the highest prediction accuracy among IAMB, MMMB, HITON-MB, CMIM, and JMI using both KNN and NBC, since it finds almost the same PC set as the other rivals, but achieves fewest false positives among all algorithms, as shown in Table 5. This shows that non-causal feature selection algorithms deal with small-sized data samples better than causal feature selection algorithms, which is consistent with our discussions of sample requirement in Section 5.

Validating the discussion presented in Section 5.2.5. Table 6 reports the prediction accuracies using MMPC, HITON-PC, mRMR, JMI, and CMIM, while Table 7 illustrates the precision and recall of each algorithm for PC discovery. From Tables 5–7, we can see that the PC set of a target feature plays a key role in predicting the target. From Tables 6 and 7, we can see that using 5,000 cases (large-sized data samples), the 3 non-causal feature selection methods find the exact PC set of "HR," and thus they get the same prediction accuracy as the true PC set. Even using 50 cases (a small-sized data samples), Table 7 states that the three non-causal feature selection methods still

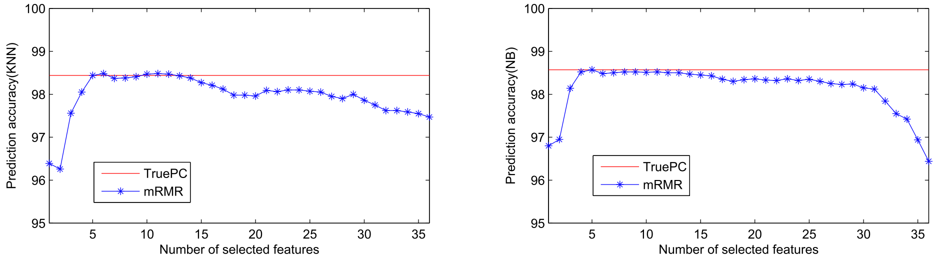


Fig. 9. mRMR and TruePC.

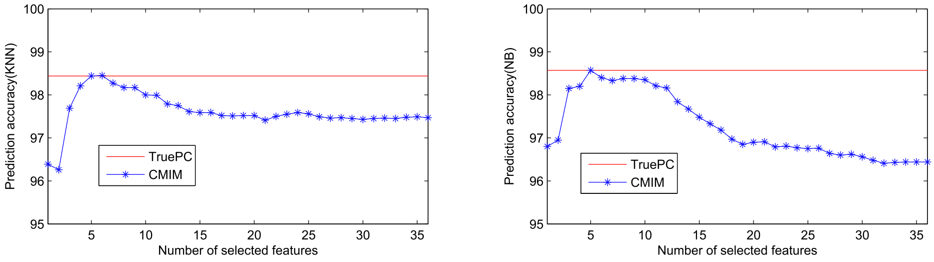


Fig. 10. CMIM and TruePC.

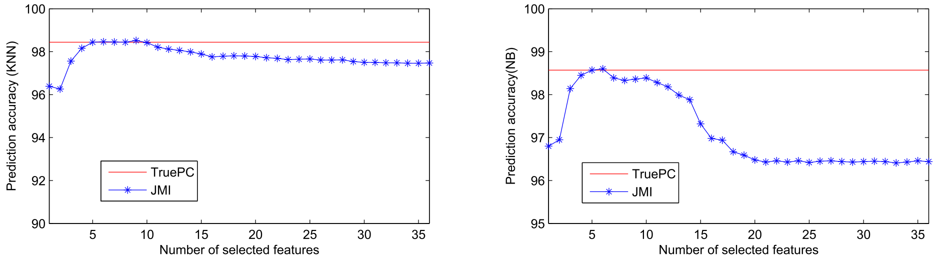


Fig. 11. JMI and TruePC.

prefer the features in the PC set of "HR." Therefore, Tables 5–7 provide strong evidence to support the discussion of causal interpretations of non-causal feature selection in Section 5.2.5.

7.1.2 Validation of Error Bounds Identified in Section 6. In the section, we will examine the proposed error bounds in Section 6. To achieve the goal, we consider the prediction accuracy of the true PC set of "HR" as a baseline using 5,000 data samples, since using the PC set of "HR," the prediction accuracy is almost the same as that using the MB set. Then we check the different prediction accuracies of different feature sets by varying the sizes of the selected feature sets by mMRM, CMIM, and JMI.

From Figures 9 to 11, we can see that the prediction accuracies of mRMR, CMIM, and JMI are bounded by the prediction accuracy of the true PC set. The highest accuracy of the three algorithms was achieved with 5 to 8 selected features, where the PC set of "HR" includes 5 features and the true MB set has 8 features. Thus those results further confirm the bounds proposed in Section 6. In Figure 11, we can see that the feature set selected by JMI containing 4 or 5 features gets the same accuracy as the true PC set of "HR." But in the left figure of Figure 11, as the size of the feature set selected by JMI is up to 8 or 9 features, JMI gets a little higher accuracy, since the feature subset may contain some spouses of "HR" in addition to the true PC set of "HR" as shown in Table 5.

Table 8. Datasets with Large Sample Sizes and a Small Feature-to-sample Ratio

Dataset	Number of features	Number of instances	Number of classes
mushroom	22	5,644	2
kr-vs-kp	36	3,196	2
madelon	500	2,000	2
gisstee	5,000	7,000	2
spambase	57	4,601	2
bankrupty	148	7,063	2

7.2 Evaluation on Real-world Data

In this section, we will conduct extensive experiments with 25 real-world datasets to examine the impact of different levels of approximations made by causal and non-causal methods on their performance, the time complexity, and the impacts of data sample sizes and different types of datasets on causal and non-causal feature selection algorithms, respectively. The 25 datasets are divided into four groups: (1) 6 datasets with large sample sizes and a small feature-to-sample ratio, i.e., "m≫n"; (2) 6 datasets with a small sample-to-feature ratio, i.e., "m≪n"; (3) 7 datasets with multiple classes; and (4) 6 datasets with extremely imbalanced class distributions. By employing NBC, KNN, and SVM, we use the prediction accuracy to evaluate the features selected by each algorithm for classification. Due to page limit, we put the prediction accuracy calculated by NBC and KNN in the supplement. In addition to prediction accuracy used in the previous section, we employ the following metrics to validate all the eight methods:

- number of selected features;
- computational efficiency (running time in seconds);
- AUC: Area Under the ROC (used for imbalanced datasets in Section 7.2.4); and
- kappa statistics (due to page limit, experimental results please see the supplement).

7.2.1 Datasets with Large Sample Sizes and a Small Feature-to-sample Ratio. We select six datasets with large numbers of samples and relatively small numbers of features from the UCI Machine Learning Repository [4], as shown in Table 8. In the experiment, since it is hard to decide in advance a suitable parameter ψ, i.e., the number of selected features, for each of mRMR, CMIM, and JMI, we set the user-defined values for ψ to 5, 10, 15, 20, and 25, respectively, for the algorithm and choose the feature subset with the highest prediction accuracy as the final feature set.

Table 9 reports the prediction accuracy of each algorithm using SVM. From these tables, we can see that the non-causal feature selection methods almost have the same performance as the causal feature selection algorithms. IAMB is a bit better than mRMR, CMIM, and JMI on some datasets. Meanwhile, we can find that using real-world datasets, MMPC and HITON-PC achieve good performance or even better than MMMB and HITON-MB.

Table 11 shows the running time of each algorithm. Clearly, among all causal feature selection algorithms, IAMB is the fastest. MMPC, HITON-PC, HITON-MB, and MMMB need to check the subsets of the feature subset currently selected, therefore they show higher time complexity than IAMB. Moreover, by combining the running time in Table 11 and the number of features selected in Table 10, we can see that more features are selected, more expensive the computations of MMPC, HITON-PC, HITON-MB, and MMMB are. Regarding the time complexity of the non-causal feature selection methods, as we discussed at the beginning of Section 5 and in Figure 4, mRMR, CMIM, and JMI use pairwise comparisons, and thus are faster than all causal feature selection algorithms, and this is validated by the result in Table 11.

Table 9. Prediction Accuracy Using SVM

Dataset	MMPC	HITON-PC	MMMB	HITON-MB	IAMB	mRMR	CMIM	JMI
mushroom	0.9996 ±0.00	0.9996 ±0.00	0.9996 ±0.00	0.9996 ±0.00	**1.00** **±0.00**	0.9986 ±0.00	**1.00** **±0.00**	**1.00** **±0.00**
kr-vs-kp	0.9408 ±0.02	0.9408 ±0.02	0.9383 ±0.02	0.9383 ±0.02	**0.9581** **±0.01**	0.9387 ±0.02	0.9387 ±0.02	0.9387 ±0.02
madelon	0.6030 ±0.02	0.6095 ±0.04	0.6050 ±0.03	0.6050 ±0.03	**0.6200** **±0.03**	0.5970 ±0.03	0.6075 ±0.04	0.6025 ±0.04
gisstee	0.8054 ±0.02	0.8033 ±0.04	0.7650 ±0.33	0.7986 ±0.03	0.8785 ±0.02	0.8857 ±0.01	0.8952 ±0.01	**0.9167** **±0.07**
spambase	0.9270 ±0.01	0.9274 ±0.01	0.9294 ±0.01	**0.9298** **±0.01**	0.9063 ±0.01	0.9234 ±0.01	0.9198 ±0.01	0.9107 ±0.01
bankrupty	0.8856 ±0.01	0.8856 ±0.01	0.8968 ±0.00	0.8976 ±0.00	**0.9036** **±0.00**	0.8856 ±0.00	0.8856 ±0.00	0.8856 ±0.00

Table 10. Number of Selected Features

Dataset	MMPC	HITON-PC	MMMB	HITON-MB	IAMB	mRMR	CMIM	JMI
mushroom	10	10	20	20	3	5/15/10	5/5/5	10/5/5
kr-vs-kp	8	8	19	19	7	10/25/15	5/15/10	5/15/10
madelon	5	5	6	5	6	25/20 /20	5/15/15	20/20/10
gisstee	295	294	1,384	1,402	2	15/20/15	25/25/20	25/20/15
spambase	24	24	45	45	8	20/25/20	15/20/15	10/15/10
bankrupty	29	28	60	56	9	15/20/10	5/15/10	5/25/10

"A/B/C" denotes that "A" represents the number of features with the highest accuracy corresponding to an algorithm using NBC and "B" and "C" are the number of features with the highest accuracy corresponding to an algorithm using KNN and SVM, respectively.

Table 11. Running Time (in seconds)

Dataset	MMPC	HITON-PC	MMMB	HITON-MB	IAMB	mRMR	CMIM	JMI
mushroom	0.89	1.19	42.98	44.12	0.16	0.03	0.01	0.03
kr-vs-kp	0.31	0.33	6.87	6.21	0.43	0.04	0.03	0.07
madelon	0.18	0.21	0.87	0.9448	3.07	0.4	0.03	1.53
gisstee	32,684	65,308	50,929	107,870	12.90	8.38	1.32	52
spambase	35	37	200	203	0.7648	0.09	0.06	0.24
bankrupty	112	95	296	239	2.06	0.31	0.11	1.27

7.2.2 Dataset with High Dimensionality and Small Number of Data Samples. In this section, we will evaluate the eight feature selection methods using the six datasets with high dimensionality and relatively small numbers of samples. Table 12 provides a summary of the datasets. In the following tables reporting the results, "-" denotes that an algorithm fails to obtain any result with a dataset because of excessive running time. We will do the same for the experiments in Sections 7.2.3 and 7.2.4. For mRMR, CMIM, and JMI, we set ψ to the top 5, 10, 15 , . . . , 35, and 40, respectively, then report the results about the feature subset with the highest prediction accuracy.

Table 12. Dataset with High Dimensionality and Small Data Sample Sizes

Dataset	Number of features	Number of instances	Number of classes
prostate	6,033	102	2
dexter	20,000	300	2
arcene	10,000	100	2
dorothea	100,000	800	2
leukemia	7,070	72	2
breast-cancer	17,817	286	2

Table 13. Prediction Accuracy Using SVM

Dataset	MMPC	HITON-PC	MMMB	HITON-MB	IAMB	mRMR	CMIM	JMI
prostate	0.8609 ±0.10	0.9200 ±0.10	0.9364 ±0.05	0.9364 ±0.05	0.9100 ±0.11	**0.9500** **±0.08**	0.9200 ±0.10	**0.9500** **±0.08**
dexter	0.8533 ±0.05	0.8433 ±0.04	0.8533 ±0.05	0.8600 ±0.05	0.7900 ±0.04	**0.9033** **±0.06**	0.9000 ±0.06	0.9000 ±0.05
arcene	0.7236 ±0.19	0.7136 ±0.18	0.7236 ±0.19	0.6914 ±0.16	0.7316 ±0.17	**0.7716** **±0.08**	0.7636 ±0.18	0.7396 ±0.11
dorothea	0.9363 ±0.03	0.9325 ±0.02	-	-	0.9300 ±0.03	0.9363 ±0.02	0.9363 ±0.02	**0.9388** **±0.02**
leukemia	0.9321 ±0.12	-	-	-	0.9446 ±0.10	0.9714 ±0.06	0.9714 ±0.06	**0.9857** **±0.05**
Breast-cancer	0.8252 ±0.06	0.7898 ±0.10	0.7967 ±0.10	0.7967 ±0.10	0.8171 ±0.08	0.8812 ±0.04	0.8567 ±0.06	**0.8707** **±0.03**

Table 14. Running Time (in seconds)

Dataset	MMPC	HITON-PC	MMMB	HITON-MB	IAMB	mRMR	CMIM	JMI
prostate	2	2	41,568	54,315	6.37	2.4	0.09	4
dexter	4	3	31	19	54	16	1	29
arcene	3	3	16	15	20	1	0.3	19
dorothea	59	705	-	-	594	4	4	59
leukemia	10,033	-	-	-	5	2	0.3	3
breast-cancer	9	11	45	43	43.23	17	0.7	31

From Table 13, we can see that the non-causal feature selection methods, mRMR, CMIM, and JMI, significantly outperform the causal feature selection methods, MMPC, HITON-PC, MMMB, HITON-MB, and IAMB. This illustrates that with datasets of high dimensionality and small sample size, as the number of data instances is not enough to support causal feature selection algorithms for reliable conditional independence tests, mRMR, CMIM, and JMI can cope with such datasets. This validates our analysis of sample requirement in Section 5.

Table 14 shows that the computational costs of HITON-PC, MMMB, and HITON-MB are very high and even prohibitive on some datasets, such as prostate, dorothea, and leukemia. The explanation is that the class attribute in each of the datasets may have a large PC set or MB set, as shown in Table 15, then this leads to that MMPC, HITON-PC, MMMB, and HITON-MB need to

Table 15. Number of Selected Features

Dataset	MMPC	HITON-PC	MMMB	HITON-MB	IAMB	mRMR	CMIM	JMI
prostate	9	8	175	98	2	5/20/10	15/5/10	20/20 /10
dexter	8	8	11	10	4	25/20/15	25/20/15	25/25/10
arcene	4	4	5	6	3	5/10/10	15/20/10	35/35/20
dorothea	24	28	-	-	6	15/15/15	40/15/20	10/10/10
leukemia	1,014	-	-	-	1	10/20/15	20/10/10	10/10/10
breast-cancer	8	6	10	7	4	40/35/25	40/25/20	35/40/25

Table 16. Dataset with Multiple Classes

Dataset	Number of features	Number of instances	Number of classes
connect- 4	42	67,557	3
splice	60	3,175	3
waveform	40	5,000	3
landsat	36	6,435	6
lung	325	73	7
lymph	4,026	96	9
NCI9	9,712	60	9

check an exponential number of subsets. On the prostate dataset, we can see that the class attribute has a large size of spouses. However, with a large size of spouses, MMMB and HITON-MB do not achieve significantly better prediction accuracy than MMPC and HITON-PC.

7.2.3 Dataset with Multiple Classes. In this section, we will evaluate the eight feature selection methods using the seven datasets with multiple classes. Table 16 provides a summary of the datasets. As for mRMR, CMIM, and JMI, we set ψ to the top 5, 10, 15, 20, and 25, respectively, then report the results about the feature subset with the highest prediction accuracy.

From Table 17, we can see that given a dataset with a small number of features and a large number of data instances, even if the dataset with multiple classes, MMPC, HITON-PC, MMMB, and HITOM-MB have almost the same prediction accuracy as three non-causal feature selection, and even better than them on some datasets, such as the landsat dataset with six classes.

However, given a dataset with a very small number of data instances and a larger number of classes, MMPC, HITON-PC, MMMB, HITOM-MB, and IAMB fail to select any features due to data inefficiency, while mRMR, CMIM, and JMI seem to work well, especially CMIM. Again this validates that non-causal feature selection algorithms deal with small-sized data samples better than causal feature selection algorithms, which is consistent with our discussions of sample requirement in Section 5. Tables 18 and 19 report the number of selected features and running time of each algorithm. We can see that as expected, mRMR, CMIM, and JMI are faster than MMPC, HITON-PC, MMMB, HITOM-MB, and IAMB.

7.2.4 Dataset with Imbalanced Classes. In this section, we use six class-imbalanced datasets in Table 20 to examine the performance of causal and non-causal feature selection methods. For mRMR, CMIM, and JMI, we set ψ to the top 5, 10, 15, . . . , 25, and 30, then select the feature subset with the highest prediction accuracy as the reporting result.

From Table 21, we can see that all the eight algorithms get good prediction accuracy, but each of them achieves a relatively low AUC, as seen from Table 22. In addition, on both prediction

Table 17. Prediction Accuracy Using SVM

Dataset	MMPC	HITON-PC	MMMB	HITON-MB	IAMB	mRMR	CMIM	JMI
connect-4	**0.7416 ±0.00**	**0.7416 ±0.00**	0.6657 ±0.00	0.6657 ±0.00	0.7096 ±0.00	0.6283 ±0.00	0.7309 ±0.00	0.7309 ±0.00
splice	0.9269 ±0.01	0.9269 ±0.01	0.9052 ±0.02	0.9039 ±0.02	0.8013 ±0.02	**0.9392 ±0.01**	0.9335 ±0.01	**0.9392 ±0.01**
waveform	**0.8488 ±0.01**	**0.8488 ±0.01**	**0.8488 ±0.01**	**0.8488 ±0.01**	0.7210 ±0.02	0.8444 ±0.01	0.8452 ±0.01	0.8452 ±0.01
landsat	**0.8816 ±0.01**	**0.8816 ±0.01**	**0.8816 ±0.01**	**0.8816 ±0.01**	0.7904 ±0.01	0.8558 ±0.01	0.8620 ±0.01	0.8592 ±0.01
lung	-	-	-	-	-	**0.8131 ±0.16**	0.7059 ±0.19	0.7845 ±0.19
lymph	-	-	-	-	-	**0.7729 ±0.10**	0.6993 ±0.10	0.7545 ±0.12
NCI9	-	-	-	-	-	**0.1017 ±0.10**	**0.1017 ±0.10**	**0.1017 ±0.10**

Table 18. Number of Selected Features

Dataset	MMPC	HITON-PC	MMMB	HITON-MB	IAMB	mRMR	CMI	JMI
connect-4	37	37	42	42	7	25/25	25/20	20/15
splice	28	28	50	49	3	25/5	25/5	20/5
waveform	17	17	17	17	3	10/15	10/15	15/15
landsat	36	36	36	36	3	25/25	25/25	25/25
lung	-	-	-	-	-	40/35	40/30	40/35
lymph	-	-	-	-	-	35/40	25/35	20/40
NCI9	-	-	-	-	-	20/30	10/25	40/40

Table 19. Running Time (in seconds)

Dataset	MMPC	HITON-PC	MMMB	HITON-MB	IAMB	mRMR	CMI	JMI
connect-4	2,679	2,949	23,564	23,795	7.85	0.76	1.4	2.33
splice	12	14	34	34	0.30	0.08	0.08	0.25
waveform	4	4	16	16	0.2917	0.05	0.03	0.17
landsat	32	43	798	1019	0.2727	0.08	0.14	0.26
lung	-	-	-	-	-	0.5	0.06	0.7
lymph	-	-	-	-	-	6	00.3	10
NCI9	-	-	-	-	-	10	0.4	22

Table 20. Class-imbalanced Datasets

Dataset	Number of features	Number of instances	Number of classes	ratio
hiva	1,617	4,229	2	3.52%
ohsumed	14,373	5,000	2	5.56%
acpj	28,228	15,779	2	1.3%
sido0	4,932	12,678	2	3.54%
thrombin	13,9351	2,543	2	7.55%
infant	86	5,339	2	6.31%

Table 21. Prediction Accuracy Using SVM

Dataset	MMPC	HITON-PC	MMMB	HITON-MB	IAMB	mRMR	CMIM	JMI
hiva	0.9655 ±0.00	0.9655 ±0.00	0.9660 ±0.00	**0.9674 ±0.01**	0.9662 ±0.00	0.9660 ±0.00	0.9655 ±0.00	0.9662 ±0.00
ohsumed	0.9468 ±0.00	0.9468 ±0.00	0.9450 ±0.00	0.9452 ±0.00	**0.9486 ±0.00**	0.9478 ±0.00	0.9478 ±0.00	0.9480 ±0.00
acpj	**0.9870 ±0.00**	**0.9870 ±0.00**	-	-	**0.9870 ±0.00**	**0.9870 ±0.00**	**0.9870 ±0.00**	**0.9870 ±0.00**
sido0	**0.9643 ±0.00**	**0.9643 ±0.00**	**0.9643 ±0.00**	**0.9643 ±0.00**	**0.9643 ±0.00**	**0.9643 ±0.00**	**0.9643 ±0.00**	**0.9643 ±0.00**
thrombin	0.9504 ±0.01	0.9245 ±0.00	0.9280 ±0.01	0.9504 ±0.01	0.9463 ±0.01	**0.9508 ±0.01**	0.9489 ±0.01	0.9473 ±0.01
infant	0.9560 ±0.01	0.9567 ±0.01	**0.9571 ±0.01**	0.9565 ±0.01	0.9556 ±0.01	0.9563 ±0.01	0.9556 ±0.01	0.9563 ±0.01

Table 22. AUC Using SVM

Dataset	MMPC	HITON-PC	MMMB	HITON-MB	IAMB	mRMR	CMIM	JMI
hiva	0.5325 ±0.03	0.5325 ±0.03	0.5263 ±0.03	**0.5754 ±0.05**	0.5328 ±0.03	0.5236 ±0.03	0.5196 ±0.02	0.5393 ±0.03
ohsumed	0.5285 ±0.02	0.5284 ±0.02	0.5106 ±0.01	0.5124 ±0.01	**0.5615 ±0.02**	0.5494 ±0.03	0.5529 ±0.03	0.5512 ±0.03
acpj	**0.5000 ±0**	**0.5000 ±0**	-	-	**0.5000 ±0**	**0.5000 ±0**	**0.5000 ±0**	**0.5000 ±0**
sido0	**0.5000 ±0.00**	**0.5000 ±0.00**	**0.5000 ±0.00**	**0.5000 ±0.00**	**0.5000 ±0.00**	**0.5000 ±0.00**	**0.5000 ±0.00**	**0.5000 ±0.00**
thrombin	0.7619 ±0.08	0.50 ±0	0.5261 ±0.08	0.7482 ±0.07	**0.7757 ±0.09**	0.7672 ±0.08	0.7592 ±0.08	0.7609 ±0.09
infant	0.6859 ±0.06	0.6931 ±0.06	**0.6962 ±0.05**	0.6944 ±0.06	0.6898 ±0.06	0.6902 ±0.05	0.6871 ±0.04	0.6889 ±0.04

Table 23. Number of Selected Features

Dataset	MMPC	HITON-PC	MMMB	HITON-MB	IAMB	mRMR	CMI	JMI
hiva	7	6	9	7	8	30/30	30/25	20/15
ohsumed	26	26	38	37	8	30/10	20/5	30/5
acpj	16	16	-	-	10	30/15	30/5	20/5
sido0	16	17	62	68	10	30/30	30/30	5/25
thrombin	15	11	38	42	7	30/30	25/30	30/25
infant	5	5	7	6	3	20/25	10/30	25/10

accuracy and AUC, the five causal feature selection methods and the three non-causal feature selection algorithms achieve almost the same performance.

Tables 23 and 24 report the number of selected features and running time of each algorithm. We can see that MMMB and HITON-MB are the slowest algorithms among the nine algorithms under comparison. Thus, we can conclude that both the causal feature selection methods and non-causal feature selection algorithms cannot deal with class-imbalanced datasets well.

Table 24. Running Time (in seconds)

Dataset	MMPC	HITON-PC	MMMB	HITON-MB	IAMB	mRMR	CMI	JMI
hiva	2	2	40	15	13	1	1	4
ohsumed	61	54	563	577	89	32	2	50
acpj	29	3	-	-	905	136	9	175
sido0	33	77	8,928	8,669	206	5	2	13
thrombin	544	133	23,456	17,615	1,429	241	11	291
infant	1	1	1	1	0.5	0.1	0.01	0.1

7.2.5 Summary of Experimental Results. In Section 5, we have analyzed that causal feature selection algorithms calculate higher order mutual information between X and C conditioning on all or a subset of the already selected features S, while the non-causal feature selection methods eventually only look at the pairwise mutual information between X and C without conditioning on other features. Then the number of data instances required by IAMB, MMPC, HITON-PC, MMMB, and HITON-MB will increase exponentially in the size of S, while mRMR, CMIM, and JMI significantly reduce the required number of data instances. Based on the analysis, we summarized the experimental results as follows.

Effectiveness. In Section 7.2.1 and Section 7.2.3, as the number of data instances is enough to support causal feature selection algorithms for reliable conditional independence tests, the five causal feature selection algorithms likely have the chance to achieve the correct MB of the class attribute and thus can achieve better prediction accuracy than the three non-causal feature selection algorithms. This is because classifiers using $MB(C)$ as the feature set achieve the best classification results.

Meanwhile, we note that for mRMR, CMIM, and JMI, their prediction accuracies listed in the tables in Section 7.2 are the highest prediction accuracies among the top 5, 10, 15, 20, 25, and more features selected by mRMR, CMIM, and JMI. In practice, we cannot get such high prediction accuracy using mRMR, CMIM, and JMI since it is hard to select the best value of the parameter ψ (the number of selected features) in advance. However, the causal feature selection algorithms, MMPC, HITON-PC, MMMB, HITON-MB, and IAMB do not need to specify the user-defined parameter ψ.

However, from Sections 7.2.2–7.2.3, we can see that as the number of data instances is not enough to support causal feature selection algorithms, this will lead to a lager number of unreliable conditional independence tests and thus the five causal feature selection algorithms cannot find the correct MB, while the three non-causal feature selection algorithms can tackle very small-sized data samples. Therefore, when a dataset has high dimensionality and small-sized data samples, the three non-causal feature selection algorithms are more practical than the five causal feature selection algorithms.

Efficiency. Since causal feature selection algorithms use all or subsets of the already selected features S, while the non-causal feature selection methods employ pairwise comparisons without conditioning on other features, in the experiments of Sections 7.2.1–7.2.4, as we expected, the three non-causal feature selection methods are faster than the five causal feature selection algorithms.

PC and MB. From Sections 7.2.1–7.2.4, the prediction accuracy (or AUC) of MMPC and HITON-PC is not inferior to that of MMMB and HITON-MB. At the same time, MMPC and HITON-PC are more computationally efficient than MMMB and HITON-MB, since MMPC and HITON-PC do not need to learn spouses. Although the MB of the class attribute is the optimal feature set for feature selection in theory, the PC set of the class attribute plays the crucial role for optimal prediction. Thus MMPC and HITON-PC are more practical than MMMB and HITON-MB in real-world applications.

8 CONCLUSION

In this article, we have proposed a unified view to fill in the gap in the research of the relation between causal and non-causal feature selection methods. With this view, we have analyzed the mechanisms of both types of feature selection methods and have shown that both major approaches to feature selection use different strategies to discover the MB of a class attribute under different BN structural assumptions. In theory, the feature sets obtained by causal feature selection methods are closer to the MB of the class attribute than non-causal feature selection methods, while non-causal methods are more computationally efficient and need fewer data samples than causal methods. With this view, we have provided causal interpretations to the output of non-causal feature selection methods and analyzed the error bounds of causal and non-causal methods. In addition, we have conducted extensive experiments to validate our findings in the article.

From the theoretical and experimental analysis in the article, we can find that both types of feature selection still face many challenges as listed below and we hope this article can stimulate the interest of researchers in machine learning to develop new methods to address these challenges.

- Small sample size. Causal feature selection cannot deal with a dataset with high dimensionality and small sample size. Then leveraging non-causal feature selection to help causal feature selection is a promising way to improve the computational performance and accuracy of causal feature selection methods for high-dimensional and small-sized data sample problems.
- Imbalanced classes. The majority of existing causal and non-causal feature selection methods cannot deal with datasets with imbalanced classes, which exist in many real-world applications. There is a need to develop new feature selection methods to battle this challenging problem.
- Large-sized MBs. A large MB containing hundreds of features makes causal feature selection methods suffer from the data-inefficiency or time-inefficiency problem due to the combinatorial optimization strategy. It is a new and exciting direction to formulating the MB learning problem as a continuous optimization problem for learning large MBs [65].
- Selection of proper parameter values. It is a hard problem for non-causal feature selection to determine a suitable value of ψ. It would be an interesting direction to follow to use the MB property in a BN to help alleviate the parameter selection problem of non-causal feature selection methods.
- Efficiency. Most local-to-global BN learning methods employ causal feature selection methods to learn MBs for constructing structure skeletons. It is interesting to extend non-causal feature selection methods for learning skeletons to improve the computational performance of BN structure learning methods [31].

REFERENCES

[1] Alan Agresti and Maria Kateri. 2011. Categorical data analysis. In *International Encyclopedia of Statistical Science*. Springer, 206–208.

[2] Constantin F. Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D. Koutsoukos. 2010. 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* 11, 7 (2010), 171–234.

[3] Constantin F. Aliferis, Ioannis Tsamardinos, and Alexander Statnikov. 2003. HITON: A novel Markov blanket algorithm for optimal variable selection. In *Proceedings of the AMIA Annual Symposium*. Vol. 2003. American Medical Informatics Association, 21.

[4] Kevin Bache and Moshe Lichman. 2013. UCI machine learning repository. Retrieved from http://archive.ics.uci.edu/ml.

[5] Kiran S. Balagani and Vir V. Phoha. 2010. On the feature selection criterion based on an approximation of multidimensional mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 7 (2010), 1342–1343.

[6] Roberto Battiti. 1994. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks* 5, 4 (1994), 537–550.

[7] Ingo A. Beinlich, Henri J. Suermondt, R. Martin Chavez, and Gregory F. Cooper. 1989. *The ALARM Monitoring System: A Case Study with Two Probabilistic Inference Techniques for Belief Networks*. Springer.

[8] David A. Bell and Hui Wang. 2000. A formalism for relevance and its application in feature subset selection. *Machine Learning* 41, 2 (2000), 175–195.

[9] Gianluca Bontempi and Patrick E. Meyer. 2010. Causal filter selection in microarray data. In *Proceedings of the 27th International Conference on Machine Learning*. 95–102.

[10] Giorgos Borboudakis and Ioannis Tsamardinos. 2019. Forward-backward selection with early dropping. *The Journal of Machine Learning Research* 20, 1 (2019), 276–314.

[11] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. 2012. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research* 13, 1 (2012), 27–66.

[12] Peter Bühlmann, Markus Kalisch, and Marloes H. Maathuis. 2010. Variable selection in high-dimensional linear models: Partially faithful distributions and the PC-simple algorithm. *Biometrika* 97, 2 (2010), 261–278.

[13] Thomas M. Cover and Joy A. Thomas. 2012. *Elements of Information Theory*. John Wiley & Sons.

[14] Manoranjan Dash and Huan Liu. 2003. Consistency-based search in feature selection. *Artificial Intelligence* 151, 1–2 (2003), 155–176.

[15] R. M. Fano. 1961. *Transmission of Information: A Statistical Theory of Communications*. MIT Press.

[16] François Fleuret. 2004. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research* 5, 9 (2004), 1531–1555.

[17] Nir Friedman, Dan Geiger, and Moises Goldszmidt. 1997. Bayesian network classifiers. *Machine Learning* 29, 2–3 (1997), 131–163.

[18] Shunkai Fu and Michel C. Desmarais. 2008. Fast Markov blanket discovery algorithm via local learning within single pass. In *Proceedings of the Conference of the Canadian Society for Computational Studies of Intelligence*. Springer, 96–107.

[19] Keinosuke Fukunaga. 2013. *Introduction to Statistical Pattern Recognition*. Academic Press.

[20] Tian Gao and Qiang Ji. 2017. Efficient Markov blanket discovery and its application. *IEEE Transactions on Cybernetics* 47, 5 (2017), 1169–1179.

[21] Isabelle Guyon, Constantin Aliferis, and André Elisseeff. 2007. Causal feature selection. *Computational Methods of Feature Selection*, H. Liu and H. Motoda (Eds.). CRC Press.

[22] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research* 3, Mar (2003), 1157–1182.

[23] Isabelle Guyon and André Elisseeff. 2006. An introduction to feature extraction. *Feature Extraction*. Springer, 1–25.

[24] Martin E. Hellman and Josef Raviv. 1970. Probability of error, equivocation and the chernoff bound. *IEEE Transactions on Information Theory* 16, 4 (1970), 368–372.

[25] Ron Kohavi and George H. John. 1997. Wrappers for feature subset selection. *Artificial Intelligence* 97, 1 (1997), 273–324.

[26] Daphne Koller and Mehran Sahami. 1995. Toward optimal feature selection. In *Proceedings of the 13th International Conference on International Conference on Machine Learning*. 284–292.

[27] Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22, 1 (1951), 79–86.

[28] David D. Lewis. 1992. Feature selection and feature extraction for text categorization. In *Proceedings of the Workshop on Speech and Natural Language*. Association for Computational Linguistics, 212–217.

[29] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. 2017. Feature selection: A data perspective. *ACM Computing Surveys* 50, 6 (2017), 1–45.

[30] Dahua Lin and Xiaoou Tang. 2006. Conditional infomax learning: An integrated framework for feature extraction and fusion. In *Proceedings of the European Conference on Computer Vision*. Springer, 68–82.

[31] Zhaolong Ling, Kui Yu, Hao Wang, Lei Li, and Xindong Wu. 2020. Using feature selection for local causal structure learning. *IEEE Transactions on Emerging Topics in Computational Intelligence*. (2020).

[32] Zhaolong Ling, Kui Yu, Hao Wang, Lin Liu, Wei Ding, and Xindong Wu. 2019. Bamb: A balanced markov blanket discovery approach to feature selection. *ACM Transactions on Intelligent Systems and Technology* 10, 5 (2019), 1–25.

[33] Dimitris Margaritis and Sebastian Thrun. 2000. Bayesian network induction via local neighborhoods. In *Proceedings of the Advances in Neural Information Processing Systems*. 505–511.

[34] Patrick Emmanuel Meyer, Colas Schretter, and Gianluca Bontempi. 2008. Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing* 2, 3 (2008), 261–274.

[35] Judea Pearl. 2014. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

[36] Jose M. Peña, Roland Nilsson, Johan Björkegren, and Jesper Tegnér. 2007. Towards scalable and data efficient learning of Markov boundaries. *International Journal of Approximate Reasoning* 45, 2 (2007), 211–232.

[37] Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 8 (2005), 1226–1238.

[38] Marko Robnik-Šikonja and Igor Kononenko. 2003. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning* 53, 1–2 (2003), 23–69.

[39] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proceedings of the IEEE*. DOI: 10.1109/JPROC.2021.3058954

[40] Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5, 1 (2001), 3–55.

[41] Alexander Shishkin, Anastasia Bezzubtseva, Alexey Drutsa, Ilia Shishkov, Ekaterina Gladkikh, Gleb Gusev, and Pavel Serdyukov. 2016. Efficient high-order interaction-aware feature selection based on conditional mutual information. In *Proceedings of the Advances in Neural Information Processing Systems*. 4637–4645.

[42] Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. 2012. Feature selection via dependence maximization. *Journal of Machine Learning Research* 13, 47 (2012), 1393–1434.

[43] Xian-fang Song, Yong Zhang, Yi-nan Guo, Xiao-yan Sun, and Yong-li Wang. 2020. Variable-size cooperative coevolutionary particle swarm optimization for feature selection on high-dimensional data. *IEEE Transactions on Evolutionary Computation* 24, 5 (2020), 882–895.

[44] Peter Spirtes, Clark N. Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. Vol. 81. MIT Press.

[45] D. Tebbe and S. Dwyer. 1968. Uncertainty and the probability of error (Corresp.). *IEEE Transactions on Information Theory* 14, 3 (1968), 516–518.

[46] Ioannis Tsamardinos and Constantin F. Aliferis. 2003. Towards principled feature selection: Relevancy, filters and wrappers. In *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*. Morgan Kaufmann Publishers.

[47] Ioannis Tsamardinos, Constantin F. Aliferis, and Alexander Statnikov. 2003. Time and sample efficient discovery of Markov blankets and direct causal relations. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 673–678.

[48] Ioannis Tsamardinos, Constantin F. Aliferis, Alexander R. Statnikov, and Er Statnikov. 2003. Algorithms for large scale Markov blanket discovery. In *Proceedings of the FLAIRS Conference*. Vol. 2. 376–380.

[49] Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning* 65, 1 (2006), 31–78.

[50] Jorge R. Vergara and Pablo A. Estévez. 2014. A review of feature selection methods based on mutual information. *Neural Computing and Applications* 24, 1 (2014), 175–186.

[51] Michel Vidal-Naquet and Shimon Ullman. 2003. Object recognition with informative features and linear classification. In *Proceedings of the 9th IEEE International Conference on Computer Vision*. Vol. 1. 281–281.

[52] Nguyen Xuan Vinh, Shuo Zhou, Jeffrey Chan, and James Bailey. 2016. Can high-order dependencies improve mutual information based feature selection? *Pattern Recognition* 53, May (2016), 46–58.

[53] De Wang, Danesh Irani, and Calton Pu. 2012. Evolutionary study of web spam: Webb spam corpus 2011 versus webb spam corpus 2006. In *Proceedings of the 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom'12)*. IEEE, 40–49.

[54] Hao Wang, Zhaolong Ling, Kui Yu, and Xindong Wu. 2020. Towards efficient and effective discovery of Markov blankets for feature selection. *Information Sciences* 509, January (2020), 227–242.

[55] Jun Wang, Jin-Mao Wei, Zhenglu Yang, and Shu-Qin Wang. 2017. Feature selection by maximizing independent classification information. *IEEE Transactions on Knowledge and Data Engineering* 29, 4 (2017), 828–841.

[56] Bing Xue, Mengjie Zhang, Will N. Browne, and Xin Yao. 2015. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation* 20, 4 (2015), 606–626.

[57] Howard Hua Yang and John Moody. 2000. Data visualization and feature selection: New algorithms for nongaussian data. In *Proceedings of the Advances in Neural Information Processing Systems*. 687–693.

[58] Sandeep Yaramakala. 2004. *Fast Markov Blanket Discovery*. Ph.D. Dissertation. Iowa State University.

[59] Sandeep Yaramakala and Dimitris Margaritis. 2005. Speculative Markov blanket discovery for optimal feature selection. In *Proceedings of the 5th IEEE International Conference on Data Mining*. IEEE, 4.

[60] Kui Yu, Xianjie Guo, Lin Liu, Jiuyong Li, Hao Wang, Zhaolong Ling, and Xindong Wu. 2020. Causality-based feature selection: Methods and evaluations. *ACM Computing Surveys* 53, 5 (2020), 1–36.

[61] Kui Yu, Lin Liu, Jiuyong Li, Wei Ding, and Thuc Duy Le. 2020. Multi-source causal feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 9 (2020), 2240–2256.

[62] Lei Yu and Huan Liu. 2004. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research* 5, December (2004), 1205–1224.

[63] Yiteng Zhai, Yew-Soon Ong, and Ivor W. Tsang. 2014. The emerging "big dimensionality". *Computational Intelligence Magazine, IEEE* 9, 3 (2014), 14–26.

[64] Yong Zhang, Dun-wei Gong, and Jian Cheng. 2015. Multi-objective particle swarm optimization approach for cost-based feature selection in classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 14, 1 (2015), 64–75.

[65] Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. 2018. DAGs with NO TEARS: Continuous optimization for structure learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 9472–9483.

Received August 2019; revised September 2020; accepted November 2020