Rule Discovery for Exploratory Causal Reasoning

Kailash Budhathoki
Max Planck Institute for Informatics and
Saarland University, Germany
kbudhath@mpi-inf.mpg.de

Mario Boley*
Monash University, Australia
mario.boley@monash.edu

Jilles Vreeken
CISPA Helmholtz Center for Information Security and
Max Planck Institute for Informatics, Germany
vreeken@cispa.saarland

Abstract

We study the problem of discovering reliable causal rules from observational data. Traditional descriptive rule discovery techniques do not suffice to this end, as they struggle with the consistent detection of (potentially rare) conditions that have a strong effect on an output variable of interest. Among the sources of inconsistency are that naive empirical effect estimations have a high variance, and, hence, their maximization is highly optimistically biased unless the search is artificially restricted to high frequency events. Secondly, observational effect measurements are often highly unrepresentative of the underlying causal effect because they are skewed by the presence of confounding factors. This is a concern especially in scientific data analysis.

To address these issues, we present a novel descriptive rule discovery approach based on reliably estimating the conditional effect given the potential confounders. We demonstrate that the corresponding score is a conservative and consistent effect estimator, identify the admissible data generation process under which causal rule discovery is possible, and derive an efficient optimization algorithm that successfully detects valuable rules on a multitude of real datasets. Important for both causal and associational data exploration, the presented approach naturally allows for iterative rule discovery, where new non-redundant rules can be found by treating previously discovered rules as confounders in subsequent iterations.

1 Introduction

Determining cause and effect from observational data—that is, from data that was not generated through carefully randomized trials—is among the most important problems in science. Over the years, we have gained a large amount of understanding of what is theoretically possible (Pearl, 2009; Spirtes et al., 2000) which in turn has led to range of methods that, under strict assumptions, can extract partially directed causal graphs from data (Spirtes et al., 2000; Chickering, 2002), up to identify the most likely causal direction between pairs of variables (Shimizu et al., 2006).

Although both impressive and useful, stating that there exists a causal relationship from a set of variables $\mathbf{X}$ towards a certain variable of interest Y does not fully satisfy ones curiosity; for a domain expert it is of particular interest to know those conditions under which the effect is visible, such as the specific combinations of drugs that lead to severe side-effects. That is, we are interested in discovering

*Part of this work was done when the author was at the Max Planck Institute for Informatics.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

causal *rules* (σ s) from observational data that consistently maximize rule effect formalized as

$$e(\sigma) = \mathbb{E}\left[Y \mid \sigma = \text{true}\right] - \mathbb{E}\left[Y \mid \sigma = \text{false}\right] \ .$$

Though simple to state, this task is not only computationally hard; algorithmic solutions also have to cope with an intricate combination of two semantic problems—one statistical and one structural.

The statistical problem is the well-known phenomenon of overfitting. This phenomenon results from the high variance of the naive empirical (or "plug-in") estimator of e for rules with too small sample sizes for either of the two events, $\sigma = \text{true}$ or $\sigma = \text{false}$. Combined with the maximization task over a usually very large rule language, this variance turns into a strong positive bias that dominates the search and causes essentially random results of either extremely specific or extremely general rules.

The structural problem is often referred to as Simpson's paradox: even strong and confidently measured effects of a rule might not actually reflect true domain mechanisms, but can be mere artifacts of the effect of other variables. Notably, such confounding effects can not only attenuate or amplify the marginal effect of a rule on the target variable, in the most misleading cases they can even result in sign reversal.

In this paper, we present a theoretically sound approach to the discovery of causal large effect rules that remedies each of the aforementioned problems.

1. To address the overfitting problem, we propose to measure and optimize the *reliable* effect of a rule. In contrast to the plug-in estimator, we propose a conservative empirical estimate of the population effect, that is not prone to overfitting. Additionally, and in contrast to other known rule optimization criteria, it is also *consistent*, i.e., with increasing amounts of evidence (data), the measure converges to the *actual* population effect of a rule.
2. To address the structural problem, we propose to control for the effect of a given set of potential confounder variables $\mathbf{Z}$. In particular, we identify the admissible data generation process under which it is possible to discover truly causal rules. While in practice the set of control variables will rarely be complete, i.e., not contain all potential confounders, this approach can rule out specific alternative explanations of findings as well as eliminate misleading observations caused by selected observables that are known to be strong confounders. In fact, this pragmatic approach is warranted not only by the usual lack of knowledge about the causal structure of the domain, but is also a necessity due to the limited availability of data.
3. We develop a practical algorithm for efficiently discovering the top-k reliable effect rules. In particular, we show how the optimization function can be cast into a branch-and-bound approach based on computationally efficient tight, selection unaware, optimistic estimators.

In sum, we propose a practically applicable rule induction technique that is able to discover true causal rules, and therewith insights to the application domain underlying a dataset. Moreover, our approach lends itself naturally to an iterative data mining approach in which we can discover insights beyond the factors included in $\mathbf{Z}$. We support our claims by experiments on real-world datasets as well as by reporting the required computation times on a large set of benchmark datasets.

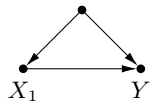
2 Rules for Causal Reasoning

Suppose that we have a population $\mathcal{U}$ of individuals u. Let Y be a binary **target** variable, and $\mathbf{X}$ be a set of **description** variables measured on the individuals. Let $\mathcal{Y} = \{0, 1\}$ be the domain of Y. Let $\mathcal{X}_j$ be the domain of a description variable $X_j \in \mathbf{X}$, which can be real or categorical. The domain of $\mathbf{X}$ is an m-dimensional outer product space $\boldsymbol{\mathcal{X}} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$. Let $\boldsymbol{\pi}$ be a set of predicates, where each predicate $\pi \in \boldsymbol{\pi}$ is either an equality or an inequality constraint on one of the description variables $X_j \in \mathbf{X}$, e.g. $\pi \equiv X_j < v$ for some threshold v. A **rule** descriptor $\sigma : \mathcal{U} \to \{\text{true}, \text{false}\}$ is a logical conjunction of predicates, i.e. $\sigma \equiv \pi_1 \wedge \pi_2 \wedge \cdots \wedge \pi_\ell$ where $\pi_j \in \boldsymbol{\pi}$. We denote by $\mathcal{L}$ the set of all possible rules that can be formulated in this way, also referred to it as **rule language**.

Let $\mathbf{X}_\sigma \subseteq \mathbf{X}$ be the subset of description variables over which predicates of σ are defined. We denote an intervention on $\mathbf{X}_\sigma$, such as setting $\mathbf{X}_\sigma$ to a vector value $\mathbf{x}$, by a do-expression of the form $do(\mathbf{X}_\sigma = \mathbf{x})$, or simply $do(\mathbf{x})$ (Pearl, 2009, Chap. 3.4). For a rule σ, we define an **unbiased intervention** $do(\sigma)$ as the randomized operation of satisfying σ by setting $\mathbf{X}_\sigma$ to some $\mathbf{x}$ such that $\sigma(\mathbf{x}) = \text{true}$ according to the probabilities $p(\mathbf{X}_\sigma = \mathbf{x} \mid \sigma = \text{true})$. Our goal is to find those rules σ

$p(Z_1 = 1)$	$p(Z_1 = 0)$
0.7	0.3

Z_1	$p(X_1 = 1 \mid Z_1)$	$p(X_1 = 0 \mid Z_1)$
1	0.9	0.1
0	0.5	0.5



Z_1	X_1	$p(Y = 1 \mid X_1, Z_1)$	$p(Y = 0 \mid X_1, Z_1)$
1	1	0.8	0.2
1	0	0.7	0.3
0	1	0.2	0.8
0	0	0.2	0.8

Figure 1: A causal Bayesian network representing causal influences among three variables. Each node is accompanied by its corresponding conditional probability distribution table.

that maximize the **causal effect** defined as the difference between the expectations of Y under two interventions $do(\sigma)$ and $do(\neg\sigma)$, i.e.

$$\begin{aligned} e(do(\sigma)) &= \mathbb{E}\left[Y \mid do(\sigma)\right] - \mathbb{E}\left[Y \mid do(\neg\sigma)\right] \\ &= p(Y = 1 \mid do(\sigma)) - p(Y = 1 \mid do(\neg\sigma)) \ . \end{aligned}$$

Unfortunately, as we consider observational data, we do not have direct access to these post-intervention distributions. In the following subsection we will identify means by which the causal effect can still be estimated under certain circumstances.

2.1 From Observational to Causal Effect

What we can do is approximate the causal effect from observational data by computing the observational effect of σ on Y by simply conditioning it on $\sigma =$ true resp. $\sigma =$ false, i.e.

$$\begin{aligned} e(\sigma) &= \mathbb{E}\left[Y \mid \sigma = \text{true}\right] - \mathbb{E}\left[Y \mid \sigma = \text{false}\right] \\ &= p(Y = 1 \mid \sigma = \text{true}) - p(Y = 1 \mid \sigma = \text{false}) \ . \end{aligned}$$

However, unless special circumstances hold, the observed conditional probability $p(Y = 1 \mid \sigma = \text{true})$ will not be the same as the post-intervention probability $p(Y = 1 \mid do(\sigma))$; they are only the same when assignment of values to $\mathbf{X}_\sigma$ is randomized. In observational data, the description variables are seldom randomized, however, and hence we typically have $e(\sigma) \neq e(do(\sigma))$. That is, in general, rule effect measure $e(\sigma)$ measures association rather than causation.

In particular, we need to be aware that the observed effect of a rule σ on the target Y may be an artifact due to variations in some other factors ($\mathbf{Z}$), also called "covariates" or "confounders", that (co-)cause the observed effect. Formally, a variable Z is a confounder of X and Y, if it influences both X and Y. To illustrate this, let us consider the causal Bayesian network in Fig. 1. Without any knowledge of the causal graph, the effect of a rule $\sigma \equiv X_1 = 1$ on variable Y would be measured as

$$\begin{aligned} e(\sigma) &= p(Y = 1 \mid \sigma = \text{true}) - p(Y = 1 \mid \sigma = \text{false}) \\ &= p(Y = 1 \mid X_1 = 1) - p(Y = 1 \mid X_1 = 0) = 0.75 \ . \end{aligned}$$

However, if we adjust the observed effect using Z_1—that is, we partition the data into groups that are homogeneous relative to Z_1, estimate the effect of σ on Y in each homogeneous group, then average the results—that influences both X_1 and Y, we get the adjusted effect of

$$\begin{aligned} e_{\text{adj}}(\sigma) = p(Z_1 = 1)\,(&p(Y = 1 \mid X_1 = 1, Z_1 = 1) - p(Y = 1 \mid X_1 = 0, Z_1 = 1)) + \\ p(Z_1 = 0)\,(&p(Y = 1 \mid X_1 = 1, Z_1 = 0) - p(Y = 1 \mid X_1 = 0, Z_1 = 0)) = 0.15 \ . \end{aligned}$$

In general, confounders can not only amplify or attenuate the marginal effect of σ, in the extreme cases we can even observe sign reversal. This phenomenon in which the association between a pair of variables reverses sign when conditioned on the third variable is also known as Simpson's paradox.

To avoid such paradoxical conclusions, we have to adjust the observational effect for confounders, or more generally all **spurious paths** (or "back-door" paths) from Y to $\mathbf{X}_\sigma$, i.e. any undirected path from Y to any variable $X_j \in \mathbf{X}_\sigma$ that has an incoming edge into X_j (Pearl, 2009, Chap. 6). We say that a set of variables $\mathbf{Z}$ satisfies the **backdoor criterion** for $\mathbf{X}_\sigma$ and Y if it blocks[2] all their

[2]For the formal definition of "blocking", please refer to (Pearl, 2009, Def. 1.2.3). For simplicity, we will assume a simplified graphical structure and can simply think of blocking variables as lying on a path.

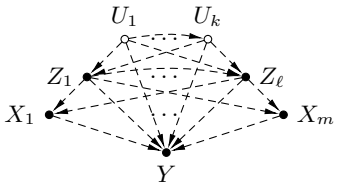


Figure 2: A skeleton causal graph that represents the data-generating process of an admissible input to causal rule discovery (see Def. 1). In the causal graph, there are no outgoing edges from $\mathbf{X}$ to $\mathbf{Z}$; no edges between variables in $\mathbf{X}$; and no outgoing edges from the latent variables $\mathbf{U}$ to $\mathbf{X}$.

spurious paths and there is no direct path from $\mathbf{X}_\sigma$ to $\mathbf{Z}$. Given a set of variables $\mathbf{Z}$, with domain $\boldsymbol{\mathcal{Z}}$, satisfying the back-door criterion, we can estimate the effect of an intervention $do(\mathbf{X}_\sigma = \mathbf{x})$ on Y from observational data using the **back-door adjustment formula** (Pearl, 2009, Thm. 3.3.2), which is given by

$$\begin{aligned} p(Y = 1 \mid do(\mathbf{X}_\sigma = \mathbf{x})) &= \sum_{\mathbf{z} \in \boldsymbol{\mathcal{Z}}} p(Y = 1 \mid \mathbf{X}_\sigma = \mathbf{x}, \mathbf{Z} = \mathbf{z}) p(\mathbf{Z} = \mathbf{z}) \\ &= \mathbb{E}_{\mathbf{Z}} \left[\mathbb{E} \left[Y \mid \mathbf{X}_\sigma = \mathbf{x}, \mathbf{Z} \right] \right] \ . \end{aligned}$$

In order to perform sound causal rule discovery over the rule language $\mathcal{L}$ we have to find a set of variables $\mathbf{Z}$ that satisfies the back-door criterion for all descriptive variables. The situation is further complicated by the potential existence of latent variables ($\mathbf{U}$) in the system. The following definition gives a criterion for an admissible input to causal rule discovery.

Definition 1 (Admissible Input to Causal Rule Discovery)**.** *Let $\mathbf{X}$ be a set of description variables, Y be a target variable, $\mathbf{Z}$ be a set of control variables, and $\mathbf{U}$ be a set of latent variables. Then $\mathbf{X}$, Y, and $\mathbf{Z}$ are admissible input to causal rule discovery if the underlying causal graph of the variables satisfies the following:*

(a) there are no outgoing edges from $\mathbf{X}$ to $\mathbf{Z}$,

(b) there are no edges between description variables $\mathbf{X}$, and

(c) there are no outgoing edges from the latent variables $\mathbf{U}$ to $\mathbf{X}$.

Such a set $\mathbf{Z}$ is then called an admissible set of control variables.

In Fig. 2, we show the skeleton of causal graph of an input system with an admissible set of control variables. A dashed edge from a node u to v indicates that u potentially affects v. Condition (a) ensures that any intervention on $\mathbf{X}$ does not affect $\mathbf{Z}$. As such, it rules out the possibility of opening a spurious path between $\mathbf{X}$ and Y via $\mathbf{Z}$ when conditioning on $\mathbf{Z}$—which can happen in case of a collider structure. By virtue of condition (b), when considering a subset of description variables $\mathbf{X}_\sigma \subset \mathbf{X}$, we do not have to condition on the remaining description variables $\mathbf{X} \setminus \mathbf{X}_\sigma$ to block any spurious path between $\mathbf{X}_\sigma$ and Y. Condition (c) ensures that, by conditioning on $\mathbf{Z}$, we block any spurious path between $\mathbf{X}$ and Y via $\mathbf{U}$.

With the definition of admissible control variables we can now state how the causal effect of a rule $\sigma \in \mathcal{L}$, i.e., the effect of the unbiased intervention $do(\sigma)$, can be computed from observational data. Let $\mathbf{Z}$ be a set of control variables. For a rule $\sigma \in \mathcal{L}$ we define the **conditional effect** as

$$e(\sigma \mid \mathbf{Z}) = \mathbb{E}_{\mathbf{Z}} \left[\mathbb{E} \left[Y \mid \sigma = \text{true}, \mathbf{Z} \right] \right] - \mathbb{E}_{\mathbf{Z}} \left[\mathbb{E} \left[Y \mid \sigma = \text{false}, \mathbf{Z} \right] \right] \ .$$

As the following theorem shows, this rule measure is indeed capable of capturing the causal effect if the set of controls variables is admissible.

Theorem 1. *Let $\mathbf{Z}$ be an admissible set of control variables for a target Y, and a set of description variables $\mathbf{X}$. Then, for any rule $\sigma \in \mathcal{L}$, we have $e(\sigma \mid \mathbf{Z}) = e(do(\sigma))$.*

Exceptional cases aside, in practice, we often do not know the complete causal diagram, and hence do not know if we are considering—or have even measured—all of $\mathbf{Z}$. In an attempt to block any path other than the direct ones between $\mathbf{X}$ and Y, a naive approach would be to include as many variables in $\mathbf{Z}$ as possible. This, however, likely leads to measuring association rather than causation. A second problem, which we will discuss in the next section, is that the more control variables we consider, the more the confidence of our estimate of the conditional effect is reduced. In short, we need to be careful in selecting $\mathbf{Z}$.

One strategy would be to initialize $\mathbf{Z}$ with our best guess, which could be the empty set, and then discover the rules with the strongest conditional effect. From these, we can then carefully select

those that we wish to add to $\mathbf{Z}$, and then iterate to investigate whether there exist strong effect rules between $\mathbf{X}$ and Y conditioned on $\mathbf{Z}$. While this does not guarantee we discover the true $\mathbf{Z}$, it does provide a natural approach to causal exploration—as well as to iterative data mining, where we wish to discover hypotheses that explain the data beyond what we already know (Hanhijärvi et al., 2009).

2.2 Statistical Considerations

In practice, we want to estimate the conditional effect of a rule from a sample drawn from the population. Let $\mathcal{S} = \{u_i = (\mathbf{x}_i, y_i, \mathbf{z}_i)\}_{i=1}^{N}$ be the sample, with $\mathbf{x}_i \in \boldsymbol{\mathcal{X}}$, $y_i \in \mathcal{Y}$ and $\mathbf{z}_i \in \boldsymbol{\mathcal{Z}}$. We assume that the sample is a stratified sample. Let $\hat{\mathbb{E}}[y]$ be the expectation of a random variable y based on its empirical distribution $\hat{p}$. Let $\mathcal{Z}_j$ be the domain of a control variable $Z_j \in \mathbf{Z}$, which can be real or categorical. The domain of $\mathbf{Z}$ is an ℓ-dimensional outer product space $\boldsymbol{\mathcal{Z}} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_\ell$. The naive estimator of the conditional effect is the estimator based on the empirical distribution $\hat{p}$, i.e. the **plug-in** estimator, which is defined as

$$\begin{aligned}\hat{e}(\sigma \mid \mathbf{Z}) &= \hat{\mathbb{E}}\left[\hat{\mathbb{E}}[Y \mid \sigma = \text{true}, \mathbf{Z}] - \hat{\mathbb{E}}[Y \mid \sigma = \text{false}, \mathbf{Z}]\right] \\ &= \sum_{\mathbf{z} \in \boldsymbol{\mathcal{Z}}} \hat{p}(\mathbf{Z} = \mathbf{z})\Big(\hat{p}(Y = 1 \mid \sigma = \text{true}, \mathbf{Z} = \mathbf{z}) - \hat{p}(Y = 1 \mid \sigma = \text{false}, \mathbf{Z} = \mathbf{z})\Big) \\ &= \sum_{\mathbf{z} \in \boldsymbol{\mathcal{Z}}} \hat{p}(\mathbf{z})(\hat{p}_\sigma - \hat{p}_{\neg\sigma}),\end{aligned}$$

where $\hat{p}_\sigma = \hat{p}(Y = 1 \mid \sigma = \text{true}, \mathbf{Z} = \mathbf{z})$, and $\hat{p}_{\neg\sigma} = \hat{p}(Y = 1 \mid \sigma = \text{false}, \mathbf{Z} = \mathbf{z})$, and $\hat{p}(\mathbf{z})$ is a shorthand for $\hat{p}(\mathbf{Z} = \mathbf{z})$. In a stratified sample, the empirical distribution of the control variable $\hat{p}(\mathbf{z})$ is the same as its distribution in the population $p(\mathbf{z})$. As the empirical distribution is an unbiased and a consistent estimator of the population distribution, the plug-in estimator is an **unbiased** and a **consistent** estimator of the conditional effect.

Although unbiased and consistent, the plugin estimator shows high variance for rules with overly small sample sizes for either of the two events, $\sigma =$ true or $\sigma =$ false. To illustrate this, in Fig. 4 (left), we show the score distribution for the plug-in estimator for a very specific rule of five conditions, and see that while it is close to the true conditional effect, it shows very high variance in small samples. This high variance is problematic, as it leads to overfitting: if we use this estimator for the optimisation task over a very large space of rules, the variance will turn into a strong positive bias—we will overestimate the effects of rules from the sample—that dominates the search, and we end up with random results of either extremely specific or extremely general rules.

We address this problem of high variance by introducing bias to the plugin-estimator. In particular, we introduce bias in terms of our confidence in the point estimates using confidence intervals. Note that we need not quantify the confidence of the point estimate $\hat{p}(\mathbf{z})$ as $\hat{p}(\mathbf{z}) = p(\mathbf{z})$; the point estimates of particular concern are simply the empirical conditional probability mass functions $\hat{p}(Y = 1 \mid \sigma = \text{true}, \mathbf{Z} = \mathbf{z})$, and $\hat{p}(Y = 1 \mid \sigma = \text{false}, \mathbf{Z} = \mathbf{z})$.

In repeated random samples of units with $\sigma =$ true and $\mathbf{Z} = \mathbf{z}$, the number of units with outcome $Y = 1$ is a binomial random variable with the success probability $\hat{p}(Y = 1 \mid \sigma = \text{true}, \mathbf{Z} = \mathbf{z})$. Let n_s be the number of successes in n trials. The one-sided confidence interval of the probability of success, $q = n_s/n$, of a binomially distributed observation, using a normal approximation on the distribution of error, is hence $\beta\sqrt{q(1-q)/n}$, where β is the $1 - \frac{\alpha}{2}$ quantile of a standard normal distribution for an error rate α. For the 95% confidence level, the error rate is $\alpha = 0.05$, thereby $\beta = 1.96$. Observe that the maximum value of $q(1-q)$ is $1/4$, and hence the maximum value of the one-sided confidence interval is $\beta/(2\sqrt{n})$.

Taking a conservative approach, we bias the difference $\hat{p}_\sigma - \hat{p}_{\neg\sigma}$ by subtracting the sum of the maximum values of the one-sided confidence intervals of the point estimates. In a control group $\mathbf{z}$, let n_σ be the number of data points that satisfy $\sigma =$ true, and $n_{\neg\sigma}$ be the number of remaining data points that satisfy $\sigma =$ false. Then the confidence interval corrected difference $\hat{p}_\sigma - \hat{p}_{\neg\sigma}$ is given by

$$\tau(\sigma, \mathbf{z}) = (\hat{p}_\sigma - \hat{p}_{\neg\sigma}) - \left(\beta/(2\sqrt{n_\sigma}) + \beta/(2\sqrt{n_{\neg\sigma}})\right).$$

Note that $\tau(\sigma, \mathbf{z})$ lower bounds the true probability mass difference in the population with confidence $1 - \alpha$. That is, there is a $1 - \alpha$ chance that the true difference is larger than $\tau(\sigma, \mathbf{z})$. For a fixed β,

	$Y=1$	$Y=0$	$\sum$
$\sigma=$ true	a	b	n_σ
$\sigma=$ false	c	d	$n_{\neg\sigma}$
$\sum$	n_1	n_0	n

Figure 3: Contingency table for a control group $\mathbf{z} \in \boldsymbol{\mathcal{Z}}$. We have $n_1 = a + c$, $n_0 = b + d$, $n_\sigma = a + b$, $n_{\neg\sigma} = c + d$ and $n = n_1 + n_0 = n_\sigma + n_{\neg\sigma}$.

the lower bound gets tighter with increasing sample size. In fact, it is easy to see that $\tau(\sigma, \mathbf{z})$ is a consistent estimator of the true probability mass difference in the population; the bias term vanishes asymptotically. More formally, for a fixed finite β, we have

$$\lim_{\min(n_\sigma, n_{\neg\sigma})\to\infty} \frac{\beta}{2\sqrt{n_\sigma}} + \frac{\beta}{2\sqrt{n_{\neg\sigma}}} = 0 \ .$$

As we use empirical probability mass functions, we can express $\tau(\sigma, \mathbf{z})$ in terms of the counts in the contingency tables. Suppose that we have a contingency table as shown in Fig. 3 for a control group $\mathbf{z} \in \mathcal{Z}$. We can then express $\tau(\sigma, \mathbf{z})$ in terms of cell counts as

$$\tau(\sigma, \mathbf{z}) = \frac{a}{n_\sigma} - \frac{c}{n_{\neg\sigma}} - \frac{\beta}{2\sqrt{n_\sigma}} - \frac{\beta}{2\sqrt{n_{\neg\sigma}}} \ .$$

In the extreme case, however, a rule may select all or none of the entities in a control group, resulting in $n_\sigma = 0$ or $n_{\neg\sigma} = 0$, and hence the empirical conditional probability mass functions can be undefined. In practice, we encounter this problem often, compounded both due to specificity of a rule—addition of predicates to a rule—as well as small sample sizes to begin with.

As a remedy, we apply a Laplace correction to the score. That is, we increment count of each cell in the contingency table by one. This way we start with a uniform distribution within each control group. Thus a control group of size n increases to $n + 4$, and the total effective sample size increases to $N + 4|\mathcal{Z}|$. After applying Laplace correction, we have $\hat{p}(\mathbf{z}) = (n + 4)/(N + 4|\mathcal{Z}|)$, and $\tau(\sigma, \mathbf{z})$ is given by

$$\tau(\sigma, \mathbf{z}) = \frac{a+1}{n_\sigma + 2} - \frac{c+1}{n_{\neg\sigma} + 2} - \frac{\beta}{2\sqrt{n_\sigma + 2}} - \frac{\beta}{2\sqrt{n_{\neg\sigma} + 2}} \ .$$

After introducing the bias and applying Laplace correction to the plug-in estimator, we obtain **reliable** estimator of the conditional effect as

$$\hat{r}(\sigma \mid \mathbf{Z}) = \sum_{\mathbf{z}\in\mathcal{Z}} \hat{p}(\mathbf{z})\tau(\sigma, \mathbf{z}) \ .$$

Although biased, $\hat{r}(\sigma \mid \mathbf{Z})$ is still a **consistent** estimator of the conditional effect. Importantly, in contrast to the plug-in estimator, the reliable estimator is much better at generalisation as it avoids overfitting.

Consider the following example to see the generalisation behaviour of the estimators. Suppose that we generate the population using the causal graph in Fig. 1. In addition, we generate five uniformly distributed binary description variables, $X_2, X_3, \ldots, X_6$ that are independent of each other as well as the rest of the variables. We can now numerically estimate the variance of the two estimators for a specific rule, such as $\sigma' \equiv X_1 = 1 \wedge X_2 = 0 \wedge X_3 = 1 \wedge X_4 = 0 \wedge X_5 = 1$, which does not only contain truly causal attribute X_1 but also other four attributes that are independent of the target Y.

To do so, we draw stratified samples of increasing sizes from the population, and report $\hat{e}(\sigma \mid \mathbf{Z})$ and $\hat{r}(\sigma \mid \mathbf{Z})$ scores averaged over 25 simulations along with one sample standard deviation in Fig. 4 (left). We observe that variances of both estimators decrease with increasing sample size. Although the reliable estimator is biased, its variance is relatively low compared to the plug-in estimator. As a result of this low variance, unlike the plug-in estimator, the reliable estimator is indeed able to avoid overfitting, and hence, better at generalisation. Let σ^* be the top-1 rule in the population, i.e. $\sigma^* = \operatorname{argmax}_{\sigma\in\mathcal{L}} e(\sigma \mid \mathbf{Z})$. Let $\varphi^* = \operatorname{argmax}_{\sigma\in\mathcal{L}} \hat{e}(\sigma \mid \mathbf{Z})$, and $\rho^* = \operatorname{argmax}_{\sigma\in\mathcal{L}} \hat{r}(\sigma \mid \mathbf{Z})$. In Fig. 4 (right), we plot $e(\varphi^* \mid \mathbf{Z})$ against $e(\rho^* \mid \mathbf{Z})$. We observe that with increasing sample sizes $e(\rho^* \mid \mathbf{Z})$ is both relatively closer, as well as converges much faster to the reference $e(\sigma^* \mid \mathbf{Z})$, which is in agreement with both theory and intuition.

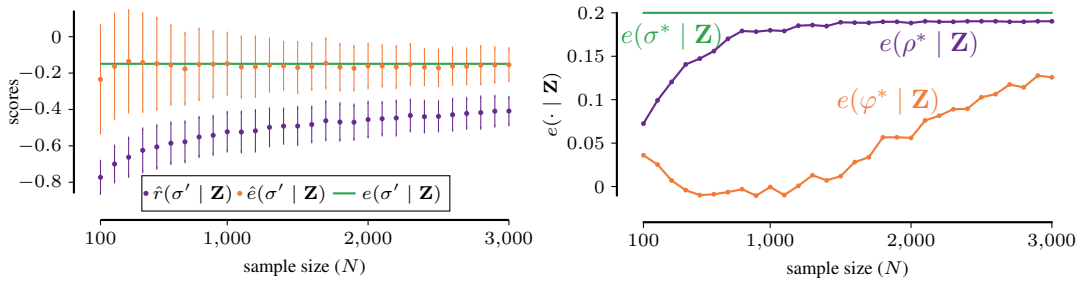


Figure 4: In stratified samples drawn from the population generated using Fig. 1 mixed together with 5 independent random description variables $X_2, \ldots, X_6$, we show (left) variance of the estimators of the conditional effect for a specific rule σ', and (right) generalisation error of the estimates.

3 Discovering Rules

Now that we have a reliable and consistent score for the conditional effect, we turn to the problem of discovering rules that yield maximal reliable conditional effect. Below, we provide the formal problem definition.

Definition 2 (Top-k causal rule discovery)**.** *Given a sample $\mathcal{S}$, and a positive integer k, find the set $\mathcal{F}_k \in \mathcal{L}$, $|\mathcal{F}_k| \leq k$, such that for all $\sigma \in \mathcal{F}_k$ and $\sigma' \in \mathcal{L} \setminus \mathcal{F}_k$, $\hat{r}(\sigma \mid \mathbf{Z}) \geq \hat{r}(\sigma')$.*

Given the hardness of empirical effect maximisation problems (Wang et al., 2005), it is unlikely that the optimisation of the reliable conditional effect allows a worst-case polynomial algorithm. While the exact computational complexity of the causal rule discovery problem is open, here we proceed to develop a practically efficient algorithm using the branch-and-bound paradigm, which is a standard approach in rule discovery.

3.1 Branch-and-Bound Search

The branch-and-bound search scheme (Mehlhorn and Sanders, 2008, Ch. 12) provides a systematic way for efficient exhaustive search by restricting the exploration to promising candidates only. More formally, the goal of branch-and-bound is to find a solution S that optimizes the objective function $f : \Omega \to \mathbb{R}$, among a set of admissible solutions Ω, also called the search space. Let $\mathbf{ext}(\sigma)$, also called the **extension** of a rule σ, be the the subset of data points in a sample described by σ, defined as $\mathbf{ext}(\sigma) = \{u \in \mathcal{S} \mid \sigma(u) = \text{true}\}$. Then the generic search scheme requires the following two ingredients (Boley et al., 2017):

- A **refinement operator** $\mathbf{r} : \mathcal{L} \to \mathcal{P}(\mathcal{L})$ that is monotone, i.e. for $\sigma, \varphi \in \mathcal{L}$ with $\varphi = \mathbf{r}(\sigma)$ it holds that $\mathbf{ext}(\varphi) \subseteq \mathbf{ext}(\sigma)$, and that non-redundantly generates the search space $\mathcal{L}$. That is, for every rule $\sigma \in \mathcal{L}$, there is a unique sequence of rules $\sigma_0, \sigma_1, \ldots, \sigma_\ell = \sigma$ with $\sigma_i = \mathbf{r}(\sigma_{i-1})$.
- An **optimistic estimator** $\bar{f} : \Omega \to \mathbb{R}$ that provides an upper bound on the objective function attainable by extending the current rule to more specific rules. That is, it holds that $\bar{f}(\sigma) \geq f(\varphi)$ for all $\varphi \in \mathcal{L}$ with $\mathbf{ext}(\varphi) \subseteq \mathbf{ext}(\sigma)$.

A branch-and-bound algorithm simply enumerates the search space $\mathcal{L}$ starting from the root ϕ using the refinement operator $\mathbf{r}$ (branch), but based on the optimistic estimator $\bar{f}$ prunes those branches that cannot yield improvement over the best rules found so far (bound). Depending on the branching operator, solutions may be reachable via multiple paths. To avoid evaluating solutions multiple times, more advanced implementations of branch-and-bound enumerate the search space non-redundantly, for example by considering only closures (Uno et al., 2003; Boley and Grosskreutz, 2009).

The optimistic estimator depends on the objective function. Interestingly, there are many optimistic estimators for an objective function f. Not all of these are equally well-suited in practice, as the tightness of the optimistic estimator determines its pruning potential. The ideal choice would therefore be the strictest optimistic estimator, which takes the maximum over all possible refinements of σ, i.e. $\bar{f}(\sigma) = \max\{f(\varphi) \mid \mathbf{ext}(\varphi) \subseteq \mathbf{ext}(\sigma) \text{ for all } \varphi \in \mathcal{L}\}$. Computing this estimator, however, is as hard as the original optimisation problem itself, and therefore not a practical solution. As the

	$Y=1$	$Y=0$	
$\sigma=$ true	a	b	
$\sigma=$ false	c	d	
$\sum$	n_1	n_0	n

	$Y=1$	$Y=0$	
$\sigma'=$ true	a'	b'	
$\sigma'=$ false	c'	a'	
$\sum$	n_1	n_0	n

Figure 5: Contingency tables for (left) σ and (right) its refinement $\sigma' = \mathbf{r}(\sigma)$ for a control group $\mathbf{z}$.

next best choice, we instead look over the valid subsets of the extension of σ without taking $\mathcal{L}$ into account. That is, we consider the **tight optimistic estimator** (Grosskreutz et al., 2008) given by

$$\begin{aligned}\bar{f}(\sigma) &= \max\{f(Q) \mid Q \subseteq \mathbf{ext}(\sigma)\} \\ &\geq \max\{f(\varphi) \mid \mathbf{ext}(\varphi) \subseteq \mathbf{ext}(\sigma) \text{ for all } \varphi \in \mathcal{L}\} \ .\end{aligned}$$

The branch-and-bound search scheme also provides an option to trade-off the optimality of the result for the speed. Instead of asking for the f-optimal result, we can ask for the γ-approximation result for some approximation factor $\gamma \in (0, 1]$. This is done by relaxing the optimistic estimator, i.e. $\bar{f}(\sigma) \geq \gamma f(\varphi)$ for all $\varphi \in \mathcal{L}$ with $\mathbf{ext}(\varphi) \subseteq \mathbf{ext}(\sigma)$. Lower γ generally yields better pruning, at the expense of guarantees on the quality of the solution.

In our problem setting, we define the refinement operator as

$$\mathbf{r}(\sigma) = \{\sigma \wedge \pi_i \mid \pi_i \in \boldsymbol{\pi}, i > \max\{j : \pi_j \in \boldsymbol{\pi}(\sigma)\}\} \ ,$$

where $\boldsymbol{\pi}(\sigma)$ is the set of predicates used in rule σ. That is, we maintain a lexicographical ordering of predicates in the pool, and only extend the current rule with a predicate from the pool that is larger than the largest predicate in current description rule in lexicographical ordering. Next we derive optimistic estimators for the objective function $\hat{r}$.

3.2 Efficient optimistic estimators

To develop optimistic estimators for the reliable estimator $\hat{r}(\sigma \mid \mathbf{Z})$, we first review its definition first,

$$\hat{r}(\sigma \mid \mathbf{Z}) = \sum_{\mathbf{z} \in \mathcal{Z}} \hat{p}(\mathbf{z}) \tau(\sigma, \mathbf{z}) \ .$$

We see that, regardless of σ, $\hat{p}(\mathbf{z})$ remains the same for all control groups $\mathbf{z} \in \mathcal{Z}$. This implies we can obtain an optimistic estimate of $\hat{r}(\sigma \mid \mathbf{Z})$ by simply bounding $\tau(\sigma, \mathbf{z})$ for all $\mathbf{z} \in \mathcal{Z}$. The optimistic estimator of $\hat{r}(\sigma \mid \mathbf{Z})$ is hence given by

$$\bar{r} = \sum_{\mathbf{z} \in \mathcal{Z}} \hat{p}(\mathbf{z}) \bar{\tau}(\sigma, \mathbf{z}) \ ,$$

where $\bar{\tau}(\sigma, \mathbf{z})$ is the optimistic estimator of $\tau(\sigma, \mathbf{z})$ yet to be defined. To derive optimistic estimators $\bar{\tau}(\sigma, \mathbf{z})$, for clarity of exposition we first project $\tau(\sigma, \mathbf{z})$ in terms of free variables a and b, such that we can write

$$\tau(a, b) = \frac{a+1}{a+b+2} - \frac{n_1 - a + 1}{n - a - b + 2} - \frac{\beta_c}{2\sqrt{a+b+2}} - \frac{\beta_c}{2\sqrt{n-a-b+2}} \ .$$

Suppose that we have a contingency table as shown in Fig. 5 (left) for a control group $\mathbf{z}$ with a rule σ. The refinement of σ, $\sigma' = \mathbf{r}(\sigma)$, results in a contingency table as shown in Fig. 5 (right). Note that n_1, n_0, and n do not change within a group regardless of a rule. Now, since $\mathbf{ext}(\sigma') \subseteq \mathbf{ext}(\sigma)$ for any $\sigma' = \mathbf{r}(\sigma)$, we have the following relations: $a' \leq a$ and $b' \leq b$. This implies that the subsets of the extensions of σ will have contingency table counts a' in the range $\{0, 1, \ldots, a\}$, and b' in the range $\{0, 1, \ldots, b\}$. Let $\mathcal{C} = \{0, 1, \ldots, a\} \times \{0, 1, \ldots, b\}$. Then the optimistic estimator of $\tau(\sigma, \mathbf{z})$ can be defined in terms of $\mathcal{C}$ as

$$\bar{\tau}(\sigma, \mathbf{z}) \geq \max_{(a', b') \in \mathcal{C}} \tau(a', b') \ .$$

This suggests that we can get a **tight optimistic estimate** of $\tau(\sigma, \mathbf{z})$ by simply taking the maximum value of τ from all possible configurations $\mathcal{C}$ in linear time. If possible, however, an improvement over this is always desirable. It turns out that we can speed up the computation of the tight optimistic estimator, which we formalize in the following proposition.

Proposition 1. *Let $\mathcal{C} = \{0, 1, \ldots, a\} \times \{0, 1, \ldots, b\}$ be the set of all possible configurations of (a', b') in Fig. 5 (right) that can result from the refinement of a rule σ from the contingency table of Fig. 5 (left). Then the* ***tight optimistic estimator*** *of $\tau(\sigma, \mathbf{z})$ is given by*

$$\bar{\tau}_t(\sigma, \mathbf{z}) = \max_{a' \in \{0,1,\ldots,a\}} \frac{a'+1}{a'+2} - \frac{n_1 - a' + 1}{n - a' + 2} - \frac{\beta_c}{2\sqrt{a'+2}} - \frac{\beta_c}{2\sqrt{n - a' + 2}}.$$

4 Related Work

In rule-based classification the goal is to find a (set of) rules that together optimally predict the target label. Classic approaches include CN2 (Lavrac et al., 2004), and FOIL (Quinlan and Cameron-Jones, 1995). In more recent work, the attention shifted from accuracy to optimizing more reliable scores, such as Area-Under-ROC (Fürnkranz and Flach, 2005). While related, the overall goal in learning classification rules is different than ours; we want to find rules that describe the strong causal effects, rather than separate two classes.

Association rules (Agrawal et al., 1993) are such rules, taking the form $x \rightarrow y$, where x is a set of items, and y is a target item. The interestingness of an association rule is typically measured in terms of its relative occurrence frequency. To get reliable rules, we can impose hard constraints on the relative occurrence frequency of an association rule. Despite that, within this framework we conflate the goal of finding rules with large effect size with the relative occurrence frequency of the rule. Contrast patterns (Dong and Li, 1999; Dong and Bailey, 2012), otherwise known as emerging patterns, are patterns whose supports differ significantly between datasets. As the support of a pattern is an empirical effect measure, without special measures such as taken here, emerging patterns tend to overfit the given sample and hence capture unreliable statements that are not necessarily characteristic of the underlying domain.

Subgroup discovery (Atzmueller, 2015) is a related, but subtly different task. Most subgroup discovery methods optimize a surrogate function based on some statistical null hypothesis test. The resulting objective function are usually some multiplicative combination of coverage and effect and, hence, do not consistently optimize for large effect. Also patterns found through standard subgroup discovery frameworks do not correct for the influence of confounder and are hence purely associational. Closer to our approach is `RAWR` (Kalofolias et al., 2017), which discovers patterns that both have large deviation from the mean of the population, but at the same time are also representative with respect to a univariate binary control variable z in terms of statistical parity. Besides that we introduce a reliable measure of effect, our framework allows for control variables of higher dimensionality that, under the specific circumstances, directly optimizes the causal effect.

Causal falling rule lists (Wang and Rudin, 2017) are sequences of "if-then" rules over the covariates such that the treatment effect decreases monotonically down the list. Our formulation, on the other hand, is aimed at finding top-k interventions, represented by rules, that have maximal effect on the target given $\mathbf{Z}$. Atzmueller and Puppe (2009) propose a semi-automatic approach to discovering causal interactions by mining subgroups, inferring a causal network over these, and visually presenting this to the user. Li et al. (2015) are specifically concerned with discovering causal association rules from observational data. They propose to do so by first mining association rules, and then performing cohort studies per rule. Unlike our setup, the scores they optimize are not statistically consistent.

Overall, despite the importance of the problem, to the best of our knowledge there does not exist a generally applicable, theoretically well-founded, efficient, and reliable approach to discovering rules with strong causal effect from observational data.

5 Experiments

We implemented the branch-and-bound search algorithm in free and open source `realKD`[3] Java library, and provide the source code online.[4] We use a priority-queue based implementation of branch-and-bound search. All experiments were executed single threaded on Intel Xeon E5-2643 v3

[3] `https://bitbucket.org/realKD/`

[4] `https://goo.gl/tcntcT`

machine with 256 GB memory running Linux. We report the results at $\beta = 2.0$, which corresponds to a 95.45% confidence level. Let the **coverage** of a rule σ be the size of the extension of σ relative to the sample size, defined as $\mathbf{cvg}(\sigma) = |\mathbf{ext}(\sigma)|/N$.

5.1 Efficiency

First we assess the efficiency of the branch-and-bound search with the proposed optimistic estimators. To this end, we consider all the standard classification datasets (22 of them) from the `KEEL` repository (Alcalá-Fdez et al., 2011). For each dataset, we select the classification target as the target variable (Y). We binarize a nominal target variable by mapping one of the outcomes to the positive category ($Y = 1$), and the rest to the negative category ($Y = 0$). We select one of the attributes for the set of control variables ($\mathbf{Z}$). We discretize a continuous real-valued description variable into maximum 8 equi-frequent bins. On each dataset, we search for the top-1 result.

In Table 2 (see Appendix), we provide the summary of the datasets along with the efficiency results. For each dataset, we report the target variable (Y), the set of control variables ($\mathbf{Z}$), the number of rows, the number of attributes, the approximation factor (γ) such that the branch-and-bound implementation finishes within an hour, the runtime in seconds, and the number of nodes expanded during the search. We observe that the branch-and-bound search with the tight optimistic estimator retrieves the optimal top-1 result (γ =1.0) within seconds for most datasets, taking up to an hour (or more) for few datasets.

5.2 Qualitative Study on Real-World Data

Next we investigate whether the rules discovered by the proposed method are meaningful. To this end, we consider the `titanic` training set from `Kaggle`.[5] The sinking of RMS Titanic is one of the notorious shipwrecks in history. One of the reasons behind such tragic loss of lives was the lack of lifeboats. During the evacuation, some passengers were treated differently than the others; some groups of people were, hence, more likely to survive than the others. Thus it is of interest to find those groups of people. The dataset contains the demographics and travel attributes of the passengers on board. The target variable of interest is the survival of a passenger.

In Table 1, we present the results of iterative rule mining on this dataset. For every iteration, we report the control variables ($\mathbf{Z}$) used in that run, and the optimal top-1 rule along with their coverage, followed by $\hat{r}(\sigma \mid \mathbf{Z})$ and $\hat{e}(\sigma \mid \mathbf{Z})$ scores. We start without control variables in the first iteration. In the subsequent iterations, we put the top-1 rules discovered from previous iterations in $\mathbf{Z}$.

In the first iteration, without any control variables, we observe that being a female passenger (sex=female) with the first, or the second class ticket (class ≤ 2) has the highest effect on survival with a score of $\hat{r}(\sigma \mid \mathbf{Z}) = 0.576$. It is well-known that passengers from different classes were treated differently during evacuation. What is interesting is that although females were more likely to survive, this only applied to the females from the first and the second class; this is also corroborated by the fact that roughly half of the females from the third class did not survive the mishap compared to the one-tenth from the other classes combined.

In the second iteration, the top-1 rule discovered in the first iteration is used as a control variable. We find that children (age < 12.5) with fewer siblings (sib-sp ≤ 2), and parents on board (par-ch ≥ 1) have highest effect on survival. The fact that this rule came out on top with a coverage of only 4.6% demonstrates that the proposed method can discover rare rules.

In the third iteration, after controlling for the top-1 rules discovered from the previous two iterations, we find that unmarried females (title=Miss) despite paying a low fare (fare < 19.85) have the highest effect on survival with a score of $\hat{r}(\sigma \mid \mathbf{Z}) = 0.003$. In the fourth iteration, we find that the top-3 rules have negative $\hat{r}(\sigma \mid \mathbf{Z})$ scores. Although the $\hat{e}(\sigma \mid \mathbf{Z})$ score is positive for the top-1 rule, the negative $\hat{r}(\sigma \mid \mathbf{Z})$ score indicates that there is no evidence. Therefore we stop after the fourth iteration.

6 Conclusions

We studied the problem of discovering reliable causal rules from observational data. Whereas traditional descriptive rule discovery methods struggle with the consistent detection of conditions

[5] `https://www.kaggle.com/c/titanic/data`

Table 1: Results of iterative rule mining on the `titanic` dataset with "survival" as a target variable. We start without control variables in the first iteration. In the subsequent iterations, we control for the top-1 rules discovered in the previous iterations. "par-ch" stands for the number of parents/children aboard, and "sib-sp" for the number of siblings/spouses aboard.

Itr.	Controls ($\mathbf{Z}$)		Top-3 rules (σ)	$\mathbf{cvg}(\sigma)\%$	$\hat{r}(\sigma \mid \mathbf{Z})$	$\hat{e}(\sigma \mid \mathbf{Z})$
1	$\emptyset$	(σ_1)	class $\leq 2 \wedge$ sex = female	19.07	0.576	0.690
2	$\{\sigma_1\}$	(σ_2)	age $< 12.5 \wedge$ sib-sp $\leq 2 \wedge$ par-ch ≥ 1	4.6	0.239	0.482
3	$\{\sigma_1, \sigma_2\}$	(σ_3)	fare $< 19.85 \wedge$ title = Miss	10.36	0.003	0.222
4	$\{\sigma_1, \sigma_2, \sigma_3\}$	(σ_4)	embarked=C	18.21	-0.036	0.149

under which a strong effect on an output variable of interest happens. Instead, we presented a novel descriptive rule discovery approach based on reliably estimating the conditional effect given the value of potential confounders. We demonstrated that the corresponding score is a conservative and consistent effect estimator, identified the admissible data generation process under which causal rule discovery is possible, and derived an efficient optimization algorithm that successfully detects valuable rules on a multitude of real datasets. Through empirical evaluation we showed our framework is efficient and applicable on datasets of realistic sizes and dimensions. Moreover, and of particular importance for both causal and associational data exploration, we showed that the presented approach naturally allows for iterative rule discovery.

Acknowledgments

Kailash Budhathoki is supported by the International Max Planck Research School for Computer Science (IMPRS-CS) and the CISPA Helmholtz Center for Information Security. All three authors were supported by the Cluster of Excellence MMCI.

References

Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. SIGMOD Rec 22(2):207–216

Alcalá-Fdez J, Fernández A, Luengo J, Derrac J, García S (2011) Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. Multiple-Valued Logic and Soft Computing 17(2–3):255–287

Atzmueller M (2015) Subgroup discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 5(1):35–49

Atzmueller M, Puppe F (2009) A knowledge-intensive approach for semi-automatic causal subgroup discovery. In: Knowledge Discovery Enhanced with Semantic and Social Information, pp 19–36

Boley M, Grosskreutz H (2009) Non-redundant subgroup discovery using a closure system. In: ECMLPKDD, Springer, pp 179–194

Boley M, Goldsmith BR, Ghiringhelli LM, Vreeken J (2017) Identifying consistent statements about numerical data with dispersion-corrected subgroup discovery. Data Mining and Knowledge Discovery 31(5):1391–1418

Chickering DM (2002) Learning equivalence classes of bayesian-network structures. JMLR 2:445–498

Dong G, Bailey J (2012) Contrast Data Mining: Concepts, Algorithms, and Applications, 1st edn. Chapman & Hall/CRC

Dong G, Li J (1999) Efficient mining of emerging patterns: Discovering trends and differences. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, pp 43–52

Fürnkranz J, Flach PA (2005) ROC 'n' rule learning - towards a better understanding of covering algorithms. Machine Learning 58(1):39–77

Grosskreutz H, Rüping S, Wrobel S (2008) Tight optimistic estimates for fast subgroup discovery. In: Machine Learning and Knowledge Discovery in Databases, Springer, pp 440–456

Hanhijärvi S, Ojala M, Vuokko N, Puolamäki K, Tatti N, Mannila H (2009) Tell me something I don't know: randomization strategies for iterative data mining. In: KDD, ACM, pp 379–388

Kalofolias J, Boley M, Vreeken J (2017) Efficiently discovering locally exceptional yet globally representative subgroups. In: 2017 IEEE International Conference on Data Mining, pp 197–206

Lavrac N, Kavsek B, Flach PA, Todorovski L (2004) Subgroup discovery with CN2-SD. JMLR 5:153–188

Li J, Le TD, Liu L, Liu J, Jin Z, Sun B, Ma S (2015) From observational studies to causal rule mining. ACM Trans Intell Syst Technol 7(2):14:1–14:27

Mehlhorn K, Sanders P (2008) Algorithms and Data Structures: The Basic Toolbox, 1st edn. Springer Publishing Company, Incorporated

Pearl J (2009) Causality: Models, Reasoning and Inference, 2nd edn. Cambridge University Press, New York, NY, USA

Quinlan JR, Cameron-Jones RM (1995) Induction of logic programs: FOIL and related systems. New Generation Comput 13(3&4):287–312

Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A (2006) A linear non-gaussian acyclic model for causal discovery. JMLR 7:2003–2030

Spirtes P, Glymour C, Scheines R (2000) Causation, Prediction, and Search. MIT press

Uno T, Asai T, Uchida Y, Arimura H (2003) LCM: an efficient algorithm for enumerating frequent closed item sets. In: FIMI

Wang F, Rudin C (2017) Causal falling rule lists. In: Proceedings of 2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning

Wang L, Zhao H, Dong G, Li J (2005) On the complexity of finding emerging patterns. Theoretical Computer Science 335(1):15–27

Proof of Theorem 1

Proof. By condition (b), $\mathbf{Z}$ also satisfies the back-door criterion for any subset of description variables $\mathbf{X}_\sigma \subset \mathbf{X}$ and Y. Thus the effect of $do(\sigma)$ on the probability of $Y = 1$ can be computed using the back-door adjustment formula as

$$\begin{aligned} p(Y = 1 \mid do(\sigma)) &= \sum_{\mathbf{x}:\sigma(\mathbf{x})=\text{true}} p(Y = 1 \mid do(\mathbf{X}_\sigma = \mathbf{x}))p(\mathbf{X}_\sigma = \mathbf{x} \mid \sigma = \text{true}) \\ &= \sum_{\mathbf{x}:\sigma(\mathbf{x})=\text{true}} \mathbb{E}_{\mathbf{Z}}\left[p(Y = 1 \mid \mathbf{X}_\sigma = \mathbf{x}, \mathbf{Z})\right] p(\mathbf{X}_\sigma = \mathbf{x} \mid \sigma = \text{true}) \end{aligned}$$

using the linearity of expectation, the definition of joint probability, and that $p(\mathbf{X}_\sigma = x, \sigma = \text{true}) = p(\mathbf{X}_\sigma = \mathbf{x})$ for $\mathbf{x}$ with $\sigma(\mathbf{x}) = \text{true}$, this results in

$$\begin{aligned} &= \mathbb{E}_{\mathbf{Z}}\left[\frac{\sum_{\mathbf{x}:\sigma(\mathbf{x})=\text{true}} p(Y = 1 \mid \mathbf{X}_\sigma = \mathbf{x}, \mathbf{Z})p(\mathbf{X}_\sigma = \mathbf{x})}{p(\sigma = \text{true})}\right] \\ &= \mathbb{E}_{\mathbf{Z}}\left[\frac{\sum_{\mathbf{x}:\sigma(\mathbf{x})=\text{true}} p(Y = 1, \mathbf{X}_\sigma = \mathbf{x} \mid \mathbf{Z})}{p(\sigma = \text{true})}\right] \\ &= \mathbb{E}_{\mathbf{Z}}\left[p(Y = 1 \mid \sigma = \text{true}, \mathbf{Z})\right] \\ &= \mathbb{E}_{\mathbf{Z}}\left[\mathbb{E}\left[Y \mid \sigma = \text{true}, \mathbf{Z}\right]\right] . \end{aligned}$$

Similarly the effect of $do(\neg\sigma)$ on the probability of $Y = 1$ can be computed as

$$p(Y = 1 \mid do(\neg\sigma)) = \mathbb{E}_{\mathbf{Z}}\left[\mathbb{E}\left[Y \mid \sigma = \text{false}, \mathbf{Z}\right]\right] .$$

Combining the two, we get

$$\begin{aligned} e(\sigma \mid \mathbf{Z}) &= \mathbb{E}_{\mathbf{Z}}\left[\mathbb{E}\left[Y \mid \sigma = \text{true}, \mathbf{Z}\right]\right] - \mathbb{E}_{\mathbf{Z}}\left[\mathbb{E}\left[Y \mid \sigma = \text{false}, \mathbf{Z}\right]\right] \\ &= p(Y = 1 \mid do(\sigma)) - p(Y = 1 \mid do(\neg\sigma)) \\ &= e(do(\sigma)) . \end{aligned}$$

□

Proof of Proposition 1

Proof. The expression for $\tau(a', b')$ from the contingency table in Fig. 5 (right) is given by

$$\tau(a', b') = \frac{a' + 1}{a' + b' + 2} - \frac{n_1 - a' + 1}{n - a' - b' + 2} - \frac{\beta_c}{2\sqrt{a' + b' + 2}} - \frac{\beta_c}{2\sqrt{n - a' - b' + 2}}.$$

Combining the first and the third term above, we get

$$\lambda_z(a', b') = \frac{2a' + 2 - \beta_c\sqrt{a' + b' + 2}}{2(a' + b' + 2)} - \frac{n_1 - a' + 1}{n - a' - b' + 2} - \frac{\beta_c}{2\sqrt{n - a' - b' + 2}}.$$

Note that if we fix the value of a', then the value of b' that maximises $\tau(a', b')$ has to maximise the first term above, but minimise the other two terms. Observe that $b' = 0$, out of $b' \in \{0, 1, \ldots, b\}$, does both simultaneously. Thus we have the following relation.

$$\tau(a', 0) > \tau(a', b') \text{ for all } b' > 0.$$

Then the tight optimistic estimator of $\tau(\sigma, \mathbf{z})$ is the maximum value over all possible configurations $\mathcal{C}$ given by

$$\begin{aligned} \bar{\tau}_t(\sigma, \mathbf{z}) &= \max_{a' \in \{0,1,\ldots,a\}} \tau(a', 0) \\ &= \max_{a' \in \{0,1,\ldots,a\}} \frac{a' + 1}{a' + 2} - \frac{n_1 - a' + 1}{n - a' + 2} - \frac{\beta_c}{2\sqrt{a' + 2}} - \frac{\beta_c}{2\sqrt{n - a' + 2}}. \end{aligned}$$

□ □

Efficiency Results on Benchmark Datasets

Table 2: Summary of the datasets used for the evaluation along with the results. For each dataset, we report the target variable (Y), the control variables ($\mathbf{Z}$), the number of rows (N), the number of description variables (m), the approximation factor γ for the branch-and-bound implementation to finish within an hour, the runtime in seconds, and the number of nodes expanded during the search.

Dataset	Target (Y)	Controls ($\mathbf{Z}$)	N	m	γ	time (s)	#nodes
adult	class	sex	48,842	13	0.8	1,717	258,575
australian	class	a4	690	13	1.0	146	952,175
automobile	output	engine-type	205	24	1.0	1	15,167
breast	class	age	286	8	1.0	78	420
car	acceptability	safety	1,728	5	1.0	1	33
chess	class	bkblk	3,196	35	1.0	851	1,613,398
connect-4	class	a1	67,557	61	0.3	1,679	140,707
crx	class	a1	690	14	1.0	14	101,621
fars	injury-severity	case-state	100,968	28	0.8	724	22,328
flare	class	prev24hour	1,066	10	1.0	1	32
german	customer	statusAndSex	1,000	19	1.0	8	43,007
housevotes	class	el-salvador-aid	435	15	1.0	1	57
kddcup	class	atr-6	494,020	40	0.99	37	219
kr-vs-k	game	white-king-col	28,056	5	1.0	30	7,304
lymphography	classes	changes-in-lym	148	17	1.0	1	1,666
mushroom	class	gill-size	8,124	21	1.0	1	215
nursery	class	social	12,690	7	1.0	1	279
post-operative	decision	l-core	90	7	1.0	1	258
splice	class	pos1	3,190	59	1.0	1.03	1,855
tic-tac-toe	class	topleft	958	8	1.0	1	488
titanic	survived	sex	891	9	1.0	4.5	26,700
zoo	type	aquatic	101	15	1.0	1	96