Causal heterogeneity discovery by bottom-up pattern search for personalised decision making

Jiuyong Li[1] · Lin Liu[1] · Shisheng Zhang[1] · Saisai Ma[2] · Thuc Duy Le[1] · Jixue Liu[1]

Abstract

In personalised decision making, evidence is required to determine whether an action (treatment) is suitable for an individual. Such evidence can be obtained by modelling treatment effect heterogeneity in subgroups. The existing interpretable modelling methods take a top-down approach to search for subgroups with heterogeneous treatment effects and they may miss the most specific and relevant context for an individual. In this paper, we design a *Treatment effect pattern (TEP)* to represent treatment effect heterogeneity in data. To achieve an interpretable presentation of TEPs, we use a local causal structure around the outcome to explicitly show how those important variables are used in modelling. We also derive a formula for unbiasedly estimating the *Conditional Average Causal Effect (CATE)* using the local structure in our problem setting. In the discovery process, we aim at minimising heterogeneity within each subgroup represented by a pattern. We propose a bottom-up search algorithm to discover the most specific patterns fitting individual circumstances the best for personalised decision making. Experiments show that the proposed method models treatment effect heterogeneity better than three other existing tree based methods in synthetic and real world data sets.

Keywords Personalised decision making · Treatment effect heterogeneity · Treatment effect pattern · Conditional average treatment effect

1 Introduction

We study the problem of identifying the *Treatment effect patterns (TEPs)* which specify subgroups where a treatment has a significant effect on the outcome. For example, chemotherapy is a common cancer treatment, but it is not suitable for all patients. Finding TEPs indicating the types of patients who are benefited (or least benefited) from chemotherapy treatment will be helpful for personalised medicine. For personalised marketing, it will be helpful to identify TEPs indicating the subgroups of customers who will buy a certain product due to a promotion (treatment).

TEPs are different from the discriminative patterns in data mining literature, e.g. emerging patterns [11], contrast sets [5] and subgroups [24, 27]. Discriminative patterns specify subgroups where the distribution of the outcome is significantly different from that outside the subgroups, and they are used for classification. For example, the discriminative pattern: {family background = business} defines a subgroup where the probability of high income (an outcome for illustration only) is larger than that outside the group. The pattern can be used to predict if a person has a high income or not.

TEPs are not aimed at predicting an outcome, but are aimed at determining whether to take a treatment (or an action) in decision making. TEPs take a fixed pair of the treatment and the outcome variables, and represent subgroups where a change in the treatment variable makes a significant change in the outcome. For example, let college education be the treatment and salary be the outcome. The discriminative pattern {family background = business} is not a TEP, as for this subgroup college education would not change their income much (this subgroup of people are likely to have high income anyway based on their family background). A TEP would be {family background = illiterate}. For this subgroup of people, education can make a big impact on their future careers. For example, without a college education, this subgroup of people may nearly all receive a very low salary. After the education,

✉ Jiuyong Li
  jiuyong.li@unisa.edu.au

1 STEM, University of South Australia, Adelaide, Australia

2 Consilium Technology, Adelaide, Australia

30% of individuals in this subgroup receive a salary higher than the median salary in the population. 30% is lower than 50%, the expected percentage of the population having a salary above the median. But for this subgroup, 30% is a big improvement. So this TEP provides strong evidence for personalised decision making on going to college or not.

A summary of the differences between TEPs and discriminative patterns is shown in Table 1.

TEPs are different from high utility patterns [14, 30] studied in recent years. Utility patterns are frequent itemsets (attribute value sets) with the minimum utility based on an internal or external utility measure, whereas TEPs present conditions for a causal relationship between the treatment and the outcome being strong or weak. Utility patterns can be extended to utility rules, but the utility rules capture associations, not causations.

TEPs are designed for personalised decision making. For example, a TEP in an e-commerce application, (new customer = true, multiple channels = true) with the treatment effect of 0.2 (treatment: sending promotional emails; outcome: visiting the online catalogue within one week) provides the company evidence for targeting this group of customers for email promotion since the email promotion causes the online catalogue visit. In a medical application, a TEP (MFAP3L = low, AGR2 = low, ABCC2 = low), where MFAP3L, AGR2, and ABCC2 are genes and the low indicates a gene expression level, with the treatment effect of -0.16 (treatment: chemotherapy; outcome: the survival rate) will discourage a doctor to recommend a patient matching the pattern for a chemotherapy treatment since the treatment does not lead to a positive outcome for this group of patients.

Our work is closely related to treatment effect heterogeneity modelling [3, 21, 22, 28, 35, 43, 44], an active research area in causal inference. We refer readers to the Related Work section for more discussions. Here we focus on tree based modelling methods since we are interested in interpretable modelling considering that interpretation is also crucial in decision making.

Treatment effect heterogeneity modelling is mainly about Conditional Average Treatment Effect (CATE) estimation which needs the causal graph underlying the data. Most existing works do not explicitly use a causal graph. For example, many works assume a given covariate set $\mathbf{X}$, such as in [3, 37]. Firstly, this covariate set is unknown to users. Secondly, even if a covariate set can be found by another algorithm (see discussions in the Related Work section.), the contributions of different variables in a covariate set to CATE estimation are different. For example, confounders which affect both W and Y need to be adjusted in treatment effect estimation, whereas effect modifiers which affect Y but do not affect W [39] do not need to be adjusted in treatment effect estimation but to be conditioned on. Such differentiation is only possible when the causal graph (or local causal structure) is presented.

In our pattern representation, we explicitly represent patterns in a local causal graphic structure and this makes causal semantic clearer. We have also proposed to use a local structure search (instead of a global structure search which can be inefficient) to find the two sets of variables in our problem setting: one set to represent confounders of the treatment and outcome, and the other set to denote effect modifiers of the outcome since two sets play different roles in causal effect estimation. Another advantage of having an explicit presentation of the local causal structure is that users can use their domain knowledge to validate the discovered TEPs since a valid causal graph is supposed to be consistent with the domain knowledge. Such pattern presentation improves the interpretability and usability of a causal effect heterogeneity model greatly.

Tree based methods have been adapted for interpretable causal heterogeneity modelling [3, 37]. These methods employ a top-down approach to recursively split a (sub)population into subgroups with different treatment effects. Their subgroup search is restricted by the choice of root node since all paths include the root, and this limits their capability for capturing significant heterogeneous subgroups.

In this paper, we employ a bottom-up search approach for identifying TEPs (subgroups), starting from the most specific patterns described by the set of all direct causes of the outcome. The patterns with small numbers of records are merged to be statistically significant. The merging process is implemented by generalisation which aims at minimising heterogeneity within a subgroup of a generalised pattern while maximising the specificity of patterns. When using the discovered TEPs, the most specific pattern matching a person's situation is used for personalised decision making.

Table 1 Differences between TEPs & discriminative patterns

	Discriminative patterns $\{W = 1, X_1 = 1\} \rightarrow Y = 1$	TEPs (w.r.t pair (W, Y)) $\{(X_1 = 1, X_2 = *), X_3 = *\}$
Nature	Association between $\{W = 1, X_1 = 1\}$ & $Y = 1$	Causation: changing W leads to a change in Y when $X_1 = 1$.
Semantics	Influence of other variables on the pattern is not indicated.	(X_1, X_2, X_3) are direct causes of Y and (X_1, X_2) confound (W, Y).
Usability	Classification	Personalised decisions

The contributions of our work are summarised in the following:

1. We design a new representation for causal effect heterogeneity modelling, TEPs, which explicitly represent the local causal structure for interpretable modelling and evaluation.
2. We derive solutions to use the local causal structure for unbiased CATE estimation in our problem setting.
3. We develop a bottom up generalisation algorithm to discover TEPs by considering within pattern homogeneity and pattern specificity. The bottom up approach ensures that the most specific pattern is used for predicting CATE for an individual.

2 Problem definition

Let D be a data set containing n records of the triple $(W, Y, \mathbf{X})$, where W is the treatment variable, Y the outcome variable, and $\mathbf{X}$ the set of pretreatment variables representing background conditions and/or characteristics of an individual, denoted by a record in D. Pre-treatment variables are not influenced by W or Y but may influence W or Y. We assume that W has an effect on Y. W takes two values, 1 and 0, standing for treatment and control respectively. and Y is a binary variable.

We are interested in answering the question: *"For a subgroup of individuals, will they benefit from receiving the treatment ($W = 1$)?"*

What we need is to estimate CATE, i.e. the change of Y as a result of changing or intervening on W under condition $\mathbf{X} = \mathbf{x}$. To make the objective formally, we use Pearl's *do* operator [31], a notation commonly used in causal inference literature, to represent an intervention. The *do* operation mimics setting a variable to a certain value (not just passively observing a value) in a real world experiment. The probability given a *do* operation, e.g. prob($y \mid do(W = 1)$)), indicates the probability of $Y = 1$ when W is set to 1, and is different from prob($y \mid W = 1$), the probability of $Y = 1$ when observing $W = 1$.

Let $\mathbf{P} = \mathbf{p}$ (or simply $\mathbf{p}$) where $\mathbf{P} \subseteq \mathbf{X}$ be a pattern which represents a subgroup in the population. For example, (male, professional) is a pattern representing a type of employees. CATE associated with pattern $\mathbf{p}$ is defined as the following.

$$
\begin{aligned}
CATE_{\mathbf{p}}(W, Y) = \ &\text{prob}(y \mid do(W = 1), \mathbf{p}) - \\
&\text{prob}(y \mid do(W = 0), \mathbf{p})
\end{aligned} \tag{1}
$$

When $\mathbf{P}$ is an empty set, $CATE_{\emptyset}(W, Y)$ is the Average Treatment Effect (ATE) in the population, specifically.

$$
\begin{aligned}
ATE(W, Y) = \ &\text{prob}(y \mid do(W = 1)) - \\
&\text{prob}(y \mid do(W = 0))
\end{aligned} \tag{2}
$$

To make CATE estimation close to the individual level, we need $\mathbf{p}$ to be specific. However, the estimated CATE may not be reliable when there are a small number of samples in the subgroup of $\mathbf{p}$. Given a data set, a pattern cannot be too specific since its CATE estimation has to be reliable. In contrast, a general pattern may contain heterogeneous treatment effects within its subgroup. Putting both considerations together, we have the following problems to be tackled in this paper.

Definition 1 (Problem definition) Given a data set D of $(W, Y, \mathbf{X})$ and $\mathbf{X}$ is a pretreatment set of (W, Y), we aim to design and find a set of patterns for personalised decision making. A pattern should be as specific as possible while its subgroup should be large enough for reliable CATE estimation. The CATEs of sub subgroups in the subgroup should have as a small difference as possible.

Equation (1) is conceptual and the CATE of a pattern cannot be estimated directly from data yet. Our next step is to develop an analytic expression to estimate CATE for a pattern from data. Firstly, we will introduce the background of causal graphs and the calculus of intervention.

3 Causal DAG and *do* calculus

A DAG (Directed Acyclic Graph), denoted as $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, is a directed graph where $\mathbf{V}$ contains a set of nodes, $\mathbf{E}$ contains a set of directed edges, and no node has a sequence of directed edges pointing back to itself. If there exists an edge $V_1 \rightarrow V_2$ in $\mathcal{G}$, V_1 is a parent node of V_2 and V_2 is a child node of V_1. We use $PA(V_2)$ to denote the set of all parents of V_2. A path is a sequence of nodes linked by edges regardless of their directions. A directed path is a path on which all the edges follow the same direction. Node V_1 is an ancestor of node V_2 if there is a directed path from V_1 to V_2, and equivalently V_2 is a descendant of V_1. V_2 is a collider if $V_1 \rightarrow V_2 \leftarrow V_3$.

Definition 2 (Markov condition [31]) Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG and $P(\mathbf{V})$ be the probability distribution over $\mathbf{V}$. $P(\mathbf{V})$ and $\mathcal{G}$ satisfy the Markov condition if, $\forall V \in \mathbf{V}$, V is conditionally independent of all of its non-descendants given $PA(V)$.

When the Markov condition holds, the joint distribution of $\mathbf{V}$ is factorised as $\text{prob}(\mathbf{V}) = \prod_{V_i \in \mathbf{V}} \text{prob}(V_i \mid PA(V_i))$.

Definition 3 (Faithfulness [36]) If all the conditional independence relationships in $P(\mathbf{V})$ are entailed by the Markov condition applied to DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, and vice versa, $P(\mathbf{V})$ and $\mathcal{G}$ are faithful to each other.

The faithfulness assumption is to ensure that the DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ represents all the conditional independence relationships in the joint distribution $P(\mathbf{V})$ and vice versa.

The following causal sufficiency assumption is needed when estimating treatment effects in data in addition to the Markov condition and faithfulness assumption.

Definition 4 (Causal sufficiency [36]) For every pair of variables observed in a data set, all their common causes are also observed in the data set.

Given the three assumptions, a DAG learned from data is a causal DAG, and parents are interpreted as the direct causes of their children.

d-Separation as defined below is an important concept to read the conditional independences/dependencies among nodes from a causal DAG.

Definition 5 (*d*-Separation [31]) A path p in a DAG is *d*-separated by a set of nodes $\mathbf{Z}$ if and only if (1) $\mathbf{S}$ contains the middle node, V_k of a chain $V_i \rightarrow V_k \rightarrow V_j$, $V_i \leftarrow V_k \leftarrow V_j$, or a fork $V_i \leftarrow V_k \rightarrow V_j$ in p; and (2) when p contains a collider V_k, i.e. $V_i \rightarrow V_k \leftarrow V_j$, none of V_k and its descendants is in $\mathbf{S}$.

When all paths between V_1 and V_2 are *d*-separated by $\mathbf{S}$ in a DAG, we have $(V_1 \perp\!\!\!\perp V_2 \mid \mathbf{S})$. We call $\mathbf{S}$ *blocks* a set of paths if it *d*-separates all the paths simultaneously.

Now we use DAG for causal effect estimation.

Definition 6 (The backdoor criterion [31]) Given a causal DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, for an ordered pair of variables (W, Y) in $\mathbf{V}$, a set of variables $\mathbf{Z} \subseteq \mathbf{V} \setminus \{W, Y\}$ is said to satisfy the backdoor criterion if (1) $\mathbf{Z}$ does not contain a descendant node of W, and (2) $\mathbf{Z}$ d-separates every path between W and Y, containing an arrow into W.

Once a set $\mathbf{Z}$ satisfying the backdoor criterion with respect to the variable pair (W, Y) is identified. $\text{prob}(y \mid do(W = w), \mathbf{Z})$ is reduced to $\text{prob}(y \mid W = w, \mathbf{Z})$ where $w \in \{0, 1\}$. This means that the causal effect defined by $do()$ operations can be estimated in data. The set $\mathbf{Z}$ is called an *adjustment* (or deconfounding) set relevant to (W, Y).

do-calculus rules [31] are more general criteria for reducing a $do()$ operation to a normal statistical expression, and are used in our derivations of CATEs for patterns. The $do()$ operation on a variable, e.g. $do(X = x)$ in DAG $\mathcal{G}$ can be represented by removing all incoming edges of X from $\mathcal{G}$. Let V_1 and V_2 be two variables in $\mathcal{G}$. $\mathcal{G}_{\overline{V_1}}$ represents the mutilated graph of $\mathcal{G}$ by removing all incoming edges of V_1, $\mathcal{G}_{\underline{V_2}}$ the mutilated graph of $\mathcal{G}$ by removing all outgoing edges of V_2, $\mathcal{G}_{\overline{V_1}, \underline{V_2}}$ the mutilated graph of $\mathcal{G}$ by removing all incoming edges of V_1 and V_2, and $\mathcal{G}_{\overline{V_1}\underline{V_2}}$ the mutilated graph of $\mathcal{G}$ by removing all incoming edges of V_1 and all outgoing edges of V_2. When $\mathbf{V_1}$ or $\mathbf{V_2}$ represents a variable set, the edge removal is then for each variable in the set. The rules of *do*-calculus are given as Theorem 3.4.1 in [31].

4 Bottom up discovery of TEPs

4.1 CATE estimation in the local causal structure

An exemplar sketch of causal DAG in the problem setting is shown in Fig. 1. $\mathbf{A}, \mathbf{A}', \mathbf{F}, \mathbf{F}'$ are parents and ancestors of W and Y respectively. $\mathbf{B}, \mathbf{B}', \mathbf{Z}$ are parents and/or ancestors of both W and Y. $\mathbf{Z}$ is an adjustment set of (W, Y) (to be discussed later in this section). $\mathbf{O}$ contains irrelevant variables which are independent of both W and Y.

In Section 2, pattern $\mathbf{p}$ is defined as a value assignment of set $\mathbf{P} \subseteq \mathbf{X}$. Based on the causal graph and do-calculus, we have the following refinement.

Theorem 1 *Given a variable pair (W, Y) and a set of pretreatment variables $\mathbf{X}$. W has a treatment effect on Y. Patterns defined in $PA'(Y)$ where $PA'(Y) = PA(Y) \backslash W$ capture all treatment effect heterogeneities of W on Y defined by $\mathbf{X}$ and a superset of $PA'(Y)$.*

Proof Let us consider a pattern $\mathbf{X} = \mathbf{x}$. Based on the definition, $\text{CATE}_{\mathbf{x}}(W, Y) = \text{prob}(y \mid do(W = 1), \mathbf{x}) - \text{prob}(y \mid do(W = 0), \mathbf{x})$.

Let w be a value of treatment W. Since the two terms of $\text{CATE}_{\mathbf{x}}(W, Y)$ are the same except for the values of W. We show how the expression with $do(w)$ is simplified.
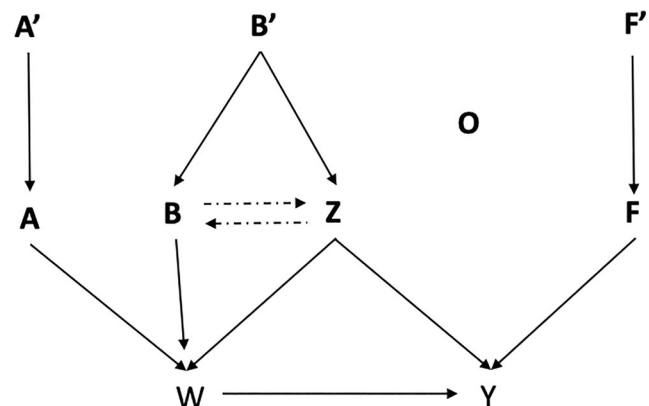

Fig. 1 An illustrative causal DAG in the problem setting. Two dash edges mean alternative paths, either into $\mathbf{Z}$ or $\mathbf{B}$ (not both). $\mathbf{A}, \mathbf{A}', \mathbf{B}, \mathbf{B}', \mathbf{Z}, \mathbf{F}, \mathbf{F}'$ and $\mathbf{O}$ are sets and there are edges between variables in a set. Edges between two sets can be multiple although only one edge is shown

Let $X = \mathrm{PA}'(Y) \cup \mathbf{N}$ where $\mathrm{PA}'(Y) \cap \mathbf{N} = \emptyset$. $\mathbf{N}$ represents the set of all non-parent nodes of Y. Let N_1 be a variable in $\mathbf{N}$, and $\mathbf{N}' = \mathbf{N} \backslash N_1$. We have the following reduction.

$$\mathrm{prob}(y \mid do(w), \mathbf{x})$$
$$= \mathrm{prob}(y \mid do(w), \mathrm{PA}'(Y) = \mathbf{p}, \mathbf{N}' = \mathbf{n}', N_1 = n_1)$$
$$= \mathrm{prob}(y \mid do(w), \mathrm{PA}'(Y) = \mathbf{p}, \mathbf{N}' = \mathbf{n}')$$

(do calculus rule 1 in Theorem 3.4.1 [31])

In the above reduction, the following rationale is used. Firstly, if there are one or more paths linking N_1 to Y in the mutilated graph $\mathcal{G}_{\overline{W}}$ where all the incoming edges of W are removed. N_1 is d-separated from Y by $\mathrm{PA}'(Y)$ in $\mathcal{G}_{\overline{W}}$. Hence, $N_1 \perp\!\!\!\perp Y \mid \mathrm{PA}'(Y)$ (equivalently $N_1 \perp\!\!\!\perp Y \mid \mathrm{PA}'(Y), \mathbf{N}'$ since there are no colliders at $\mathbf{N}'$ in $\mathcal{G}_{\overline{W}}$. Therefore, N_1 is removed from the equation based on do calculus rule 1 in Theorem 3.4.1 [31]. Secondly, if there is not a path linking N_1 to Y in $\mathcal{G}_{\overline{W}}$, $N_1 \perp\!\!\!\perp Y \mid \mathrm{PA}'(Y), \mathbf{N}'$ trivially holds and hence the do calculus rule 1 is applied.

By repeatedly using Rule 1 in Theorem 3.4.1 [31], all variables in $\mathbf{N}'$ are removed from the equation one by one. We obtain the following equation.

$\mathrm{prob}(y \mid do(w), \mathbf{x}) = \mathrm{prob}(y \mid do(w), \mathrm{PA}'(Y) = \mathbf{p})$.

So $\mathrm{CATE}_{\mathbf{x}}$ is determined by a pattern defined by the parents of Y excluding W.

Following the above procedure, any pattern defined by a superset of $\mathrm{PA}'(Y)$ can be reduced to a pattern in $\mathrm{PA}'(Y)$ with the same CATE.

Hence patterns defined in $\mathrm{PA}'(Y)$ capture all treatment effect heterogeneities defined by $\mathbf{X}$ and a superset of $\mathrm{PA}'(Y)$.   □

Theorem 1 reduces the complexity for finding patterns significantly. This is different from feature selection since $\mathbf{A}, \mathbf{A}', \mathbf{B}, \mathbf{B}', \mathbf{Z}', \mathbf{F}'$ are all associated with Y. The strength of association between two adjacent variables may not be stronger than that between two unadjacent variables. For example, the association between $\mathbf{A}$ (or $\mathbf{A}'$) and Y could be stronger than the association between $\mathbf{Z}$ (or $\mathbf{F}$) and Y. So feature selection cannot find the parents of Y.

The parents of Y can be found in a causal graph. In some real world applications, parents of Y are known by domain experts since they are direct causes of Y. The parents of Y can be learned in data in our problem setting and we will discuss learning parents in data in Section 4.4.

$\mathrm{PA}'(Y)$ contains confounders and the parents of Y only. Confounders are variables that affect both (the selection of) treatment W and effect Y, and hence need to be adjusted in treatment effect estimation. In graphical terms, Confounders have paths into both W and Y in our problem setting. Let set $\mathbf{Z}$ be parents of Y and parents (or ancestors) of W. $\mathbf{F} = \mathrm{PA}'(Y) \backslash \mathbf{Z}$ is the set of parents of Y only, and they do not have paths into W. In our problem setting, $\mathbf{F} \perp\!\!\!\perp W$ since variables in $\mathbf{F}$ are not parents or ancestors of W. We separate $\mathbf{Z}$ from other variables because of the following property.

Corollary 1 *Set $\mathbf{Z}$ is a minimal adjustment set for pair (W, Y) and the average treatment effect of W on Y is $ATE(W, Y) = \sum_z (\mathrm{prob}(y \mid W = 1, \mathbf{z}) - \mathrm{prob}(y \mid W = 0, \mathbf{z})) \mathrm{prob}(\mathbf{z})$*

Proof Set $\mathbf{Z}$ blocks all the backdoor paths of (W, Y) since $\mathbf{F} = \mathrm{PA}'(Y) \backslash \mathbf{Z}$ do not have backdoor paths into W. According to Theorem 1, set $\mathbf{Z}$ is an adjustment set and the $ATE(W, Y)$ is calculated by the summation shown. A subset of $\mathbf{Z}$ leaves a backdoor path unblocked, and does not satisfy Theorem 1. Hence, set $\mathbf{Z}$ is minimal.   □

The parents of Y only (which are d-separated from W by an empty set) are effect modifiers, e.g. $\mathbf{F}$. The average treatment effects between (W, Y) conditioned on different values of $\mathbf{F}$ are different [39].

4.2 The minimal TEP set

Now we can define treatment effect patterns to represent the causal heterogeneity in data.

Definition 7 (Treatment effect patterns (TEPs)) Given a variable pair (W, Y) and a set of pretreatment variables $\mathbf{X}$. Let $\mathbf{P} = \mathrm{PA}'(Y) \subseteq \mathbf{X}$. A TEP is a value set $\mathbf{P} = \mathbf{p}$ representing a subgroup of population and its associated treatment effect is $\mathrm{CATE}(W, Y)_{\mathbf{p}}$. To represent the local causal structure around Y, a TEP is represented as $\{(\mathbf{Z} = \mathbf{z}), \mathbf{F} = \mathbf{f}\}$ where $\mathbf{Z} \cup \mathbf{F} = \mathbf{P}, \mathbf{Z} \cap \mathbf{F} = \emptyset, \mathbf{Z}$ denotes a set of confounders and $\mathbf{F}$ stand for a set of effect modifiers.

Let us use $\mathrm{PA}'(Y) = \{X_1, X_2, X_3\}$, $\mathbf{Z} = \{X_1, X_2\}$ and $\mathbf{F} = \{X_3\}$ as an example. $\mathbf{p_1} = \{(X_1 = 1, X_2 = 0), X_3 = 1\}$ is a TEP.

Definition 8 (Specific and general TEPs) A TEP $\mathbf{p}$ is one of the most specific patterns if all its values are specified. A general pattern contains one or more unspecific values '$*$', and represents the union of subgroups of two or more most specific TEPs. When we consider the relationship between two TEPs, we drop unspecified values. If $\mathbf{p_2} \subset \mathbf{p_1}$, TEP $\mathbf{p_2}$ is more general than TEP $\mathbf{p_1}$ or TEP $\mathbf{p_1}$ is more specific than TEP $\mathbf{p_2}$.

For example, consider $\mathbf{p_1} = \{(X_1 = 1, X_2 = 0), X_3 = 1\}$ and pattern $\mathbf{p_2} = \{(X_1 = 1, X_2 = *), X_3 = 1\}$. pattern $\mathbf{p_2}$ is more general than pattern $\mathbf{p_1}$ or pattern $\mathbf{p_1}$ is more specific than pattern $\mathbf{p_2}$.

Note that $X = *$ in a TEP does not mean simply dropping variable X as in the traditional emerging patterns [11], contrast sets [5] and subgroups [24, 27] since an unspecified value of a variable in $\mathbf{Z}$ affects the CATE estimation as discussed below.

Now we derive $\mathrm{CATE}(W, Y)$ when there are unspecified values, i.e. '$*$'s. Let $\mathbf{Z} = \mathbf{Z_1} \cup \mathbf{Z_2}$ and $\mathbf{F} = \mathbf{F_1} \cup \mathbf{F_2}$ where $\mathbf{Z_1}$ and $\mathbf{F_1}$ contain specified values and $\mathbf{Z_2}$ and $\mathbf{F_2}$ contain unspecified values.

Theorem 2 *In the problem setting,* $\mathrm{CATE}_{\mathbf{p}}(W, Y) = \sum_{z_2}(prob(y \mid W = 1, \mathbf{z_1}, \mathbf{z_2}, \mathbf{f_1}) - prob(y \mid W = 0, \mathbf{z_1}, \mathbf{z_2}, \mathbf{f_1}))prob(\mathbf{z_2} \mid \mathbf{z_1}, \mathbf{f_1})$

Proof Based on the definition, $\mathrm{CATE}_{\mathbf{p}}(W, Y) = \mathrm{prob}(y \mid do(W = 1), \mathbf{z_1}, \mathbf{f_1}) - \mathrm{prob}(y \mid do(W = 0), \mathbf{z_1}, \mathbf{f_1})$ since $\mathbf{P} = \{\mathbf{Z_1}, \mathbf{F_1}\}$.

Let w be a value of treatment W. Since the two terms of $\mathrm{CATE}_{\mathbf{p}}(W, Y)$ are the same except for the values of W. We will show how $do(W = w)$ (shorted as $do(w)$) is reduced a do free expression.

$$\mathrm{prob}(y \mid do(w), \mathbf{z_1}, \mathbf{f_1}) = \sum_{\mathbf{z_2}} \mathrm{prob}(y, \mathbf{z_2} \mid do(w), \mathbf{z_1}, \mathbf{f_1})$$

$$= \sum_{\mathbf{z_2}}(\mathrm{prob}(y \mid do(w), \mathbf{z_1}, \mathbf{z_2}, \mathbf{f_1})\mathrm{prob}(\mathbf{z_2} \mid do(w), \mathbf{z_1}, \mathbf{f_1}))$$

$$= (\mathrm{Rule\ 2}) \sum_{\mathbf{z_2}}(\mathrm{prob}(y \mid w, \mathbf{z_1}, \mathbf{z_2}, \mathbf{f_1})\mathrm{prob}(\mathbf{z_2} \mid do(w), \mathbf{z_1}, \mathbf{f_1}))$$

$$= (\mathrm{Rule\ 3}) \sum_{\mathbf{z_2}}(\mathrm{prob}(y \mid w, \mathbf{z_1}, \mathbf{z_2}, \mathbf{f_1})\mathrm{prob}(\mathbf{z_2} \mid \mathbf{z_1}, \mathbf{f_1}))$$

In the second last step of reduction, do calculus rule 2 in Theorem 3.4.1 [31] has been used. In the mutilated graph $\mathcal{G}_{\underline{W}}$ where edge $W \rightarrow Y$ is removed, W is d-separated from Y by $\mathbf{Z_1}, \mathbf{Z_2}$. There are no colliders at $\mathbf{F_1}$. Hence, $W \perp\!\!\!\perp Y \mid \mathbf{Z_1}, \mathbf{Z_2}, \mathbf{F_1}$ in $\mathcal{G}_{\underline{W}}$ and "do" is removed from $do(W)$.

In the last step of reduction, do calculus rule 3 in Theorem 3.4.1 [31] has been used. In the mutilated graph $\mathcal{G}_{\overline{W}}$ where edges into W have been removed, W is d-separated from $\mathbf{Z_2}$ by the empty set. W is d-separated from $\mathbf{Z_2}$ by set $\{\mathbf{Z_1}, \mathbf{F_1}\}$ since there are no colliders at $\mathbf{Z_1}, \mathbf{F_1}$. Hence, $W \perp\!\!\!\perp \mathbf{Z_2} \mid \mathbf{Z_1}, \mathbf{F_1}$ in in $\mathcal{G}_{\overline{W}}$ and $do(\mathbf{Z_2})$ is removed from the equation.

Therefore, the Theorem is proved. $\square$

The CATE of the most general pattern, such as, $\mathbf{p_3} = \{(X_1 = *, X_2 = *), X_3 = *\}$, is the $\mathrm{ATE}(W, Y)$ in the population.

We are interested in significant patterns with reliable statistics.

Definition 9 (Significant patterns) Pattern $\mathbf{p}$ is significant if the difference $\Delta = |\mathrm{prob}(y \mid W = 1, \mathrm{PA}'(Y) = \mathbf{p}) - \mathrm{prob}(y \mid W = 0, \mathrm{PA}'(Y) = \mathbf{p})|$ is greater than 0 statistically.

We use a critical ratio statistic as in [13] to test the significance of difference Δ. Based on the values of W and Y, we obtain the following cross table where $n_{*j} = n_{1j} + n_{0j}$, $n_{i*} = n_{i1} + n_{i0}$, and $n_{\mathbf{p}}$ is the total number of samples in subgroup $\mathbf{p}$.

|	$Y = 1$	$Y = 0$	Total	
$W = 1$	n_{11}	n_{10}	n_{1*}	$r_1 = n_{11}/n_{1*}$
$W = 0$	n_{01}	n_{00}	n_{0*}	$r_0 = n_{01}/n_{0*}$
total	n_{*1}	n_{*0}	$n_{\mathbf{p}}$	$\bar{r} = n_{*1}/n_{\mathbf{p}}$

$\Delta = |r_1 - r_0|$ is significantly grater than 0 if $z > z_c$ where $z = \frac{|r_1 - r_0| - \frac{1}{2}(\frac{1}{n_{1*}} + \frac{1}{n_{0*}})}{\sqrt{\bar{r}(1-\bar{r})(\frac{1}{n_{1*}} + \frac{1}{n_{0*}})}}$ and z_c is the critical value at a confidence level. For example, when the confidence level is 95%, $z_c = 1.96$.

The most specific TEPs and their general TEPs form a lattice in space $\mathbf{X}$. The number of TEPs can be large. We aim at finding the minimal set of TEPs that explain every individual with the most specific TEP. A TEP *covers* a record in a data set if the TEP is a subset of the record when unspecified values in the TEP are dropped.

Definition 10 (The minimal significant TEP set) A TEP set is significant and minimal with respect to a data set when 1) each TEP is significant except that the most general TEP may be insignificant; 2) all TEPs in the set cover all records in the data set; and 3) each TEP is the most specific for some records, i.e. it covers at least one record that is not covered by another more specific TEP in the set.

The minimum in the above definition means non-redundancy. A more general TEP is redundant if it does not cover any new records in addition to its more specific TEPs. A redundant TEP is excluded from the minimal significant TEP set. Figure 2 shows the minimal significant TEP set. Note that a record may be covered by more than one TEP. For example, some records are covered by both TEPs $\{(0, *), 0\}$ and $\{(0, *), *\}$ (for $\mathrm{PA}'(Y) = \{(X_1, X_2), X_3\}$). We consider that $\{(0, *), 0\}$ (the more specific one among two) is the TEP covering the records. TEP $\{(0, *), *\}$ is not redundant since it covers records covered by TEP $\{(0, 0), 0\}$ which is not in the minimal significant TEP set. Note that it is possible that there are not enough significant TEPs to cover all instances in a data set and

Fig. 2 An illustration of the minimal significant TEP set whose members are in the box with the solid line. TEPs with '?' treatment effects are insignificant. This example also shows how TEPs are discovered via pattern generalisation
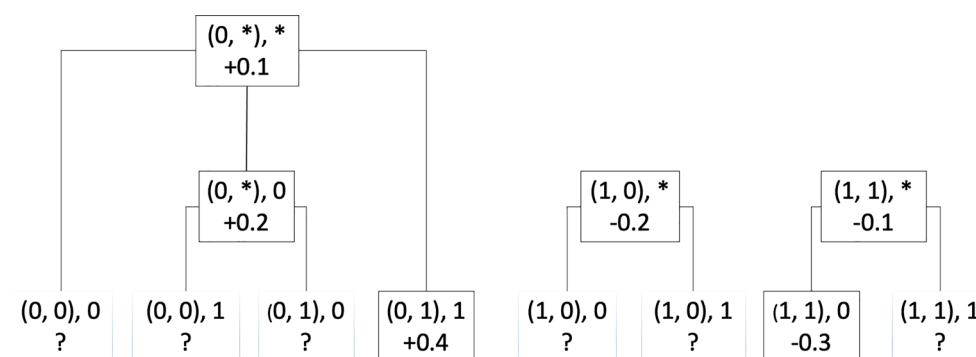
those uncovered instances are explained by the most general TEP corresponding to ATE(W, Y). This is caused by the data limitation, and using ATE(W, Y) to estimate treatment effect is reasonable.

Finding the minimal significant TEP set is to solve a set cover problem, which is NP-hard [7]. We will propose a greedy algorithm to find the minimal TEP set.

4.3 TEP discovery via pattern generalisation

We start with the set of most specific TEPs, some or all of which are insignificant. The main reason for an insignificant TEP is that the subgroup of the TEP is small. We will merge the subgroup with other subgroups to make the TEP of the merged subgroup significant.

Definition 11 (TEP generalisation) Generalisation is a merge process where one or more specified values are replaced by unspecified values '*'s. A generalised TEP represents two or more (if there are more than one unspecified value) most specific TEPs.

An example of TEP generalisation is given in Fig. 2. Patterns $\{(0, 0), 1\}$ and $\{(0, 1), 1\}$ (for PA$'(Y) = \{(X_1, X_2), X_3\}$) are generalised as pattern $\{(0, *), 1\}$. Patterns $\{(0, 0), 0\}$, $\{(0, *), 0\}$ and $\{(0, 1), 1\}$ are generalised as patterns $\{(0, *), *\}$.

There are two constraints in the generalisation.

1. The generalisation should involve as small heterogeneity as possible. A generalised TEP denotes the number of subgroups represented by a set of most specific TEPs with different treatment effects. The differences between the treatment effects should be as small as possible to make the resulted causal effect represent the treatment effects of all subgroups well.
2. The generalisation should keep the specificity as high as possible. An unspecified value means a loss of specificity. The higher the speciality, the treatment effect represented by a TEP closer to the individual treatment effect. The lower the speciality, the treatment effect represented by a TEP closer to the average causal

effect in the population. For the purpose of personalised decision making, we wish a TEP is as specific as possible, and hence the number of '*' values should be minimised. A bottom up approach as proposed in this paper has an advantage over other existing top-down partition approaches to produce specific patterns.

We use the following measure to quantify the heterogeneity.

Definition 12 (Diversity) Let a generalised TEP **p** represent k most specific TEPs, $\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_k}$, and θ be the average treatment effect of k the TEPs. The diversity of treatment effect of pattern **p** is DV$(\mathbf{p}) = \frac{1}{k} \sum_1^k (\text{CATE}_{\mathbf{p_k}} - \theta)$.

In the merge process, we prefer a merger with the smallest diversity.

The specificity loss is measured by the number of '*' in a generalised TEP. To minimise the loss, the TEPs to be merged should have the smallest edit distance (or the number of different values).

The generalisation can be modelled as a multiple objective optimisation problem following the two constraints. We design a level-wise generalisation algorithm by using the ϵ-constraint strategy for a Pareto optimal solution [29]. In each step, we constrain the specificity loss to the smallest possible loss, and search for the generalisation to minimise heterogeneity. More specifically, the search strategy is as the following.

1. for each insignificant pattern, find its closest patterns with the smallest edit distance (to maximise the specificity).
2. In the set of closest patterns, choose a pattern to generalise resulting in the smallest diversity in the generalised pattern (to minimise the heterogeneity).

Let diversity dv_0 be the diversity of the most general TEP in the data set. We do not merge patterns resulting in a diversity larger than dv_0 since in this case, the average treatment effect represents the individuals in the generalised pattern better.

Algorithm 1 Discovering trEatment Effect Patterns (DEEP).

Input: Data set D of treatment variable W, outcome variable Y, and pretreatment variables $\mathbf{X}$; Confidence level for independence tests and significance test.
Output: the minimal TEP set $\mathbf{S}$.
1: call a local structure learning algorithm to find $PA(Y)$
2: let $\mathbf{Z} = PA(Y) \setminus W$ and $\mathbf{F} = \emptyset$
3: **for** each $X \in PA(Y)$ **do**
4:     **if** $(X \perp\!\!\!\perp W)$ **then**
5:         $\mathbf{Z} \leftarrow \mathbf{Z} \setminus \{X\}$ and $\mathbf{F} = \mathbf{F} \cup X$
6:     **end if**
7: **end for**
8: let $\mathbf{S} = \emptyset, \overline{\mathbf{S}} = \emptyset$ and $\mathbf{A} = \emptyset$
9: project D on $(W, PA'(Y), Y)$ and save as D'
10: stratify D' by $PA'(Y)$ and obtain all most specific patterns
11: update cross tables for all most specific patterns
12: calculate CATE of all most specific patterns and add them to $\mathbf{A}$
13: add significant patterns to $\mathbf{S}$
14: let $\overline{\mathbf{S}} = \mathbf{A} - \mathbf{S}$
15: let dv_0 be the diversity of the most general TEP
16: calculate pairwise pattern distances between patterns in $\mathbf{A}$ and store the distances in matrix $\mathbf{M}$
17: let d be the smallest distance in $\mathbf{M}$
18: **while** $(\overline{\mathbf{S}} \neq \emptyset$ and $d < |PA'(Y)|)$ **do**
19:     let $\mathbf{L}$ contain pattern pairs in $\mathbf{A}$ with pattern distance of d and at least one pattern in each pair being insignificant
20:     generalise a pair with the closest CATEs to $\mathbf{p}$
21:     **if** $DV_{\mathbf{p}} > dv_0$ **then**
22:         remove the pair from $\mathbf{L}$ and continue
23:     **end if**
24:     add pattern $\mathbf{p}$ to $\mathbf{A}$ and $\overline{\mathbf{S}}$ and remove the patterns used to generalise $\mathbf{p}$ from $\mathbf{A}$ and $\overline{\mathbf{S}}$
25:     remove all pairs in $\mathbf{L}$ used to generalising $\mathbf{p}$
26:     **if** pattern $\mathbf{p}$ is significant **then**
27:         add $\mathbf{p}$ to $\mathbf{S}$ and remove $\mathbf{p}$ from $\overline{\mathbf{S}}$
28:     **end if**
29:     update pairwise pattern distance matrix $\mathbf{M}$
30:     let d be the smallest value in $\mathbf{M}$
31: **end while**
32: add the most general pattern to $\mathbf{S}$
33: output $\mathbf{S}$

4.4 Algorithm

Based on the above discussions, we propose a DEEP algorithm to find the minimal set of significant TEPs in Algorithm 2. The algorithm consists of three phases: Finding the local causal structure $\{\mathbf{Z}, \mathbf{F}\}$; Initialisation of the most specific TEPs; and Generalising for discovering significant TEPs. After discussing the three phases, we will discuss the complexity of the algorithm and how to use TEPs for personalised decision making.

4.4.1 Finding the local causal structure $\{\mathbf{Z}, \mathbf{F}\}$ (Lines 1 - 7)

Ideally, a causal DAG is given by domain experts, and $\mathbf{Z}$ and $\mathbf{F}$ are read from the DAG. However, in most applications, a causal DAG is unavailable.

For finding $PA(Y)$ from data, one straightforward way is to learn an entire causal DAG from data. However, learning an entire DAG is computationally expensive or intractable with high dimensional data.

Local structure discovery [2], i.e. discovering $PC(Y)$, the set of Parents (direct causes) and Children (direct effects) of the target Y is sufficient for our algorithm. In our problem setting, Y does not have descendants, and hence, $PC(Y) = PA(Y)$. Several algorithms have been developed for discovering $PC(Y)$, such as PC-Select [6], MMPC (Max-Min Parents and Children) [38] and HITON-PC [1]. These algorithms use the framework of constraint-based Bayesian network learning and employ conditional independence tests for finding the PC set of a variable. Their performance is very similar. We chose MMPC because of its newly updated implementation [23]. This is implemented in Line 1.

To distinguish sets $\mathbf{Z}$ and $\mathbf{F}$ in $PA'(Y)$, we use the following property to find $F \in \mathbf{F}$. $F \in PA'(Y)$ is a parent of Y only if $F \perp\!\!\!\perp W$ in data. This is because edges (F, Y) and (W, Y) form a collider at Y. This is implemented in Lines 2-7.

4.4.2 Initialisation of the most specific TEPs (Lines 8 - 15)

Three sets $\mathbf{S}, \overline{\mathbf{S}}$ and $\mathbf{A}$ initialised in Line 8 are used to store significant, insignificant and all TEPs, respectively. The data set is projected to variable set $PA'(Y)$ in Line 9 since TEPs are defined in $PA'(Y)$. Stratification is used to count the cross table for each pattern in Lines 10-11. The CATE of each TEP is calculated by its cross table in Line 12. The significant patterns passing the statistical test are added to the TEP set in Line 13. The diversity of the most general TEP is calculated in line 15 and assigned to dv_0.

4.4.3 Generalising for discovering significant TEPs (Lines 16 - 32)

Immediately after the above initialisation steps, all patterns in set $\mathbf{A}$ are most specific without the unspecified values '∗'. Pairwise edit distances of all patterns are calculated and stored in matrix $\mathbf{M}$ in Line 16 and the shortest distance is found in Line 17. Note, in distance calculation, an unspecified value '∗' and a specified value (1 or 0) are different. Two unspecified values are also different since they may represent different values. Lines 18-31 are for generalisation, and this process stops when $\overline{\mathbf{S}}$ is empty or TEPs in $\overline{\mathbf{S}}$ are nearly generalised to their most general form (only one specific value left). To prepare for generalisation, all pattern pairs with the shortest edit distance are found and added to list $\mathbf{L}$ and the pairs involving both significant TEPs are excluded from the list since we aim at finding the minimal significant TEP set. In list $\mathbf{L}$, the pattern pair with the smallest difference among their treatment effects is generalised. If the diversity of the generalised pattern

is larger than that of the most general TEP, dv_0. The generalised pattern is discarded and the pair is removed from **L**. This is implemented in Lines 21-22. The TEPs used in the generalisation are replaced by the generalised pattern in both sets **A** and $\overline{\mathbf{S}}$. If the generalised pattern is significant, it is removed from the insignificant pattern set and added to the significant TEP set. The generalisation distance matrix is updated using the generalised patterns and the shortest pattern distance is found.

After the loop, the most general pattern with all '∗' values is added for those uncovered records by TEPs in the data or coming test records that do not occur in the training data set.

4.4.4 Using TEPs for personalised decisions

Significant TEPs identified from data are used for personalised decision making. Match an individual's record to the most specific TEP in the minimal significant TEP set. If more than one TEP match the record with the same specificity, the one with the largest n (the cardinality of its covering set) is chosen. The treatment effect of the individual is estimated as the CATE of the TEP. The treatment is recommended to the individual if the CATE is positive, and the treatment is not recommended otherwise.

4.4.5 Time complexity

Finding $PA(Y)$ takes $O(|\mathbf{X}||PA(Y)|^{k+1})$ where k is the size of the maximal conditional set for conditional independence test (usually $k = 3 - 6$) by MMPC [23, 38]. The initialisation of the most specific patterns takes $O(n \log(n))$ of time due to stratification. The pattern generalisation in the worst case takes $O(4^{|PA'(Y)|})$ when all the most specific patterns are generalised, and in most cases, it takes less time. The overall time complexity is $O(|\mathbf{X}||PA(Y)^{k+1}| + n \log(n) + O(4^{|PA'(Y)|}))$. So the complexity is determined by the number of parents of the outcome variable. DEEP works for the data sets where the number of parents of the outcome is not many.

5 Experiments

5.1 Baseline methods and parameter setting

We compare DEEP with two state-of-the-art methods for causal effect heterogeneity modelling, Causal Tree (CT) [3], and Interaction Tree (IT) [37], and one uplift modelling method, Uplift Decision Tree (UpliftDT) [33]. All three methods are tree based, and their interpretability is comparable to DEEP's since a tree path can be interpreted as a pattern. Other causal heterogeneity and uplift modelling models do not provide the same interpretability and hence are not compared.

We use the CT implementation available at https://github.com/susanathey/causalTree by the authors of [3]. For IT, we use the R implementation available at http://biopharmnet.com/subgroup-analysis-software/. The default parameters are used for the two methods. UpliftTree is obtained from https://causalml.readthedocs.io/en/latest/methodology.html#uplift-tree. Euclidean distance is used since it performs best in the authors' work [33]. Other parameters are kept as the default.

The parameters of DEEP are listed as follows. The confidence level for testing significant patterns in DEEP is set as 95%. We have employed the R implementation of MMPC [23] for PC discovery, and set $\max_k$ as 3, p value as 0.05 and gSquare for independence tests.

5.2 Evaluation of synthetic data sets

This part aims at evaluating the quality of TEPs for modelling causal heterogeneity. The ground truth CATEs are necessary and hence the evaluation has been conducted in synthetic data sets.

We have used the code in [3] to generate synthetic data sets. Variables are binarised using their means since DEEP deals with binary variables. The numbers of variables are set as 20, 40, 60, 80 and 100 respectively, and the data set size is fixed at 10,000 for all. The number of parents of Y,

Table 2 PEHE of different methods on 10 synthetics data sets using 10-cross validation

#Var	CT	DEEP	IT	UpliftDT
20	0.201 (0.007)	**0.116** (0.005)	0.154 (0.004)	0.401 (0.070)
40	0.240 (0.005)	**0.109** (0.005)	0.171 (0.007)	0.394 (0.008)
60	0.257 (0.006)	**0.108** (0.003)	0.214 (0.010)	0.410 (0.007)
80	0.278 (0.005)	**0.105** (0.003)	0.267 (0.012)	0.396(0.005)
100	0.285 (0.007)	**0.107** (0.004)	0.312 (0.014)	0.400 (0.006)
Ave	0.252 (0.006)	**0.109** (0.004)	0.223 (0.009)	0.400 (0.019)

The standard deviation is shown in the parentheses. The smallest PEHE for each data set is highlighted

Table 3 MAPE of different methods on 10 synthetics data sets using 10-cross validation

#Var	CT	DEEP	IT	UpliftDT
20	86.0 (4.3)	**51.4** (4.5)	72.3 (3.5)	246.1 (6.2)
40	103.0 (2.8)	**45.0** (3.2)	73.3 (4.3)	212.8 (7.2)
60	114.1 (3.7)	**47.2** (1.3)	87.5 (6.5)	234.9 (4.4)
80	125.1 (4.3)	**46.4** (1.7)	110.7 (6.3)	255.2 (7.4)
100	128.3 (5.1)	**45.4** (1.4)	124.3 (9.9)	224.5 (5.7)
Ave	111.3 (4.0)	**47.1** (2.6)	93.6 (6.1)	234.7 (6.2)

The standard deviation is shown in the parentheses. The smallest MAPE for each data set is highlighted

i.e. $|\mathbf{Z} \cup \mathbf{F}|$, is 8 in all data sets. 10 data sets are generated randomly in each setting.

The ground truth CATEs are known and hence PEHE and MAPE are used for evaluating the quality of models. The Precision in Estimation of Heterogeneous Effects (PEHE) [19] measures the mean squared error of estimated CATEs. i.e. $\text{PEHE} = \frac{1}{n} \sum_{i}^{n} (\hat{\tau}(\mathbf{x}_i) - \tau(\mathbf{x}_i))^2$ where $\hat{\tau}(\mathbf{x}_i)$ and $\tau(\mathbf{x}_i)$ are estimated CATE and ground truth CATE of individual $\mathbf{x}_i$ respectively. The Mean Absolute Percentage Error (MAPE) is $\frac{1}{n} \sum_{i}^{n} |\frac{\hat{\tau}(\mathbf{x}_i) - \tau(\mathbf{x}_i)}{\tau(\mathbf{x}_i)}| \times 100\%$. PEHE and MAPE are obtained by 10-cross validation in each data set and averaging over 10 data sets.

DEEP performs better than three other methods in terms of both PEHE and MAPE as shown in Tables 2 and 3. This is because that DEEP keeps the information as specific as possible and hence predicts CATEs better than others.

5.3 Evaluation on real world data sets

We evaluate the methods on four real world data sets which are briefly described in Table 4. Criteo uplift prediction dataset [10] is an open-access large scale data set. We have randomly sampled 200,000 records from the original data set. The Hillstrom's Email dataset is from https://blog.minethatdata.com/2008/ 03/minethatdata-e-mail-analytics-and-data.html. The Marketing campaign data set is part of the Information R-package (https://cran.r-project.org/web/packages/ Information/index.html). In the data sets, numerical variables have been binarised by their medians. The US Census (KDD) data set is from the UCI Machine Learning Repository [4]. We have selected the following attributes for easy interpretation: 'College degree' (the treatment), 'Income > 50K' (the outcome), 'Age < 30', 'Age > 60', 'Work-in-Private", 'Work-in-Government", 'Self-employed', 'Professional', and 'Full time', and 'Sex'.

Since there are no ground truth CATEs in the real world data sets, we cannot use PEHE and MAPE to assess the quality of the methods. Instead, we use prediction accuracy for the assessment. A predicted CATE indicates the chance of improvement of an individual if s/he takes the treatment. We cannot assess the accuracy of each prediction. However, we can estimate the cumulative improvement of a group of individuals. In a test data set, all individuals are ranked by their predicted CATEs, and are then partitioned into 10 groups: Decile 1 to 10 groups with CATE in the descending order. If a model is good, the observed difference, $\text{prob}(y \mid W = 1) - \text{prob}(y \mid W = 0)$, in the 10 groups will monotonically decrease with the increase of the decile indexes. The higher quality a model is, the steeper the declining rate. The decile plots have been used for assessing the quality of uplift models [17].

Table 4 A brief description of the real world data sets

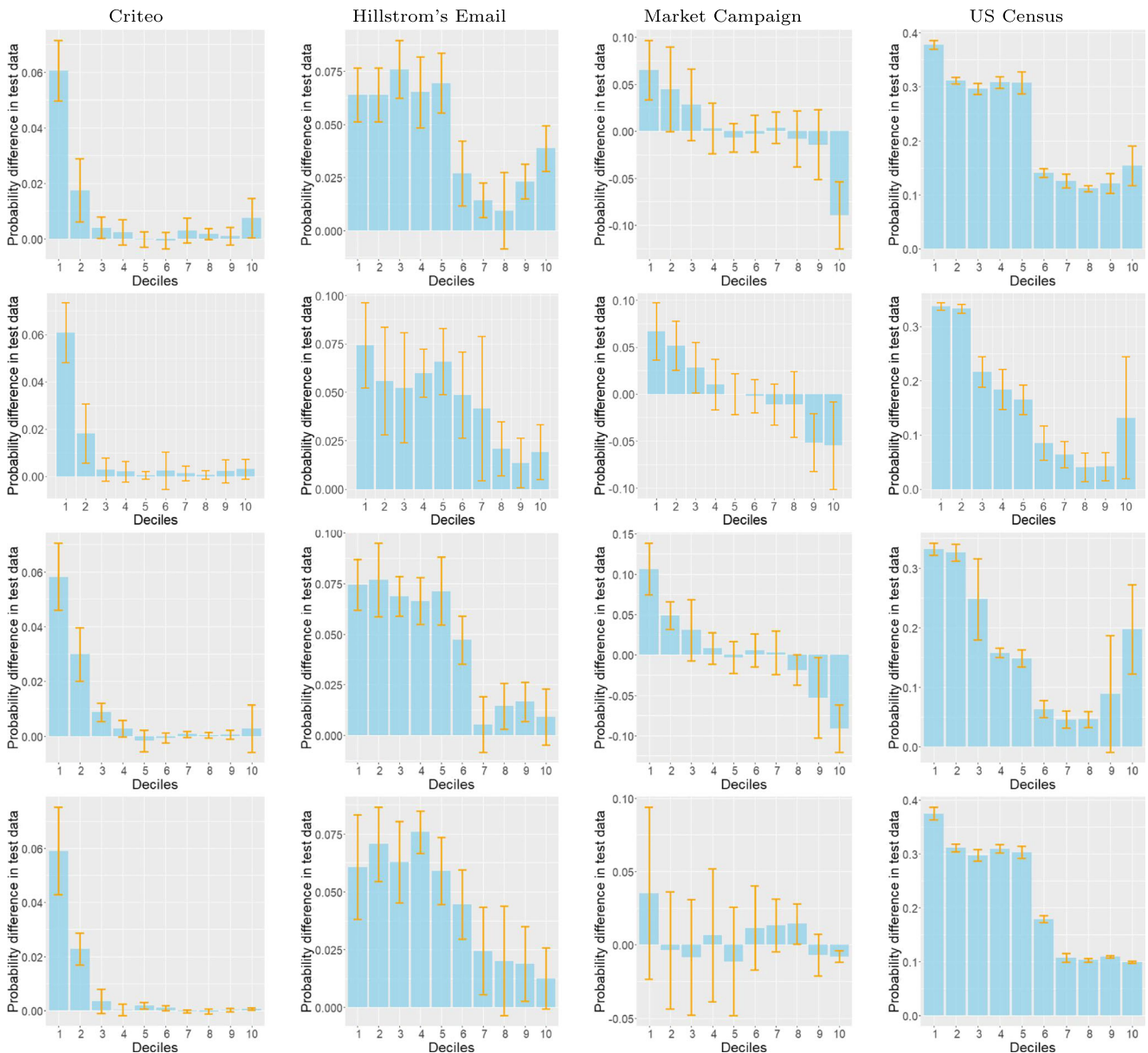| Name	Criteo	Hillstrom's Email	Marketing Campaign	US Census
#Records	200000	42693	20000	348128
#Var	12	29	67	8
Treatment	Promotional email	Women's email	Marketing offer	College degree
Outcome	Visit (4.7%)	Visit (12.9%)	Accept (20.0%)	Income$\geq$50K (8.7%)

Fig. 3 Decile plots of four methods in four data sets. From the 1st row to the last: CT, DEEP, IT and UplifDT. DEEP shows a more consistent decline in four decile plots

The decile plots in Fig. 3 show that DEEP performs overall more consistent than other methods. In data sets Criteo, Hillstrom's email and US Census, DEEP performs better than others since it presents a steep declining curve. In the Market Campaign data set, DEEP's performance is very competitive with CT and IT and better than UpliftDT. No other algorithms perform as consistent as DEEP in all four data sets. The results have been obtained by 10 times 2-fold cross validation.

The number of patterns (or paths from the root to leaves) are shown in Table 5. DEEP does not discover too many patterns and this is due to the significant test for a pattern. A tree based method is able to find many subgroups by

increasing the tree height, but a tree based method does not have flexibility like DEEP since all patterns from a tree are constrained by the variable at the root: all patterns include a value of the root variable. In contrast, DEEP does not have such a constraint and can model any heterogeneous subgroups.

5.4 Time efficiency

We apply DEEP and three other methods to the synthetic data sets of 20, 40, 60, 80, and 100 variables with 10,000 records for testing their scalability with the number of variables, and to the synthetic data sets of 5K, 10K, 20K,

Table 5 The number of patterns (or paths from the root to leaves in a tree) from different methods in four data sets. The standard deviation is shown in the parentheses

	CT	DEEP	IT	UpliftDT
Criteo	14.3 (4.6)	39.4 (6.1)	4.7 (1.3)	31.8 (2.5)
Hillstrom	47.5 (7.9)	24.9 (3.8)	2.4 (1.0)	38.7 (4.6)
Market	75.1 (11.0)	58.6 (13.5)	10.4 (4.1)	30.5 (2.3)
US Census	8.7 (2.4)	75.0 (1.7)	24.3 (2.1)	3.7 (1.9)

40K, 60K, and 80K with 40 variables for testing their scalability with the data set size. Results are shown in Fig. 4.

With the number of variables, the scalability of all four methods is good. Relatively, UpliftDT is the fastest since it does not estimate CATEs or test reliability for tree splits. CT is the slowest since it needs to estimate propensity scores for CATE estimation. Logics regression is used in the propensity score estimation. DEEP and IT perform similarly.

With the number of records, DEEP and UpliftDT perform very well. The increase in data set size improves the time efficiency of DEEP since pattern generalisation is an expensive part of DEEP. With the increase in data set size, the number of significant patterns at the most specific level is increasing, and this reduces the overall number of patterns to be merged. Since CT uses logistic regression for propensity score estimation, its performance deteriorates quickly with the increase in data set size. IT grows a large tree firstly and then prunes it back to a small tree. In the pruning process, cross validation is used to determine whether to retain a leaf node or not, and this leads to its low scalability with the data set size. The scalability of DEEP the data set size is good. DEEP works for large sized data sets.

6 Related work

Great research efforts have been made on treatment effect estimation within two major frameworks: graphic causal modelling [31] and potential outcome modelling [20]. The work in this paper falls into the former.
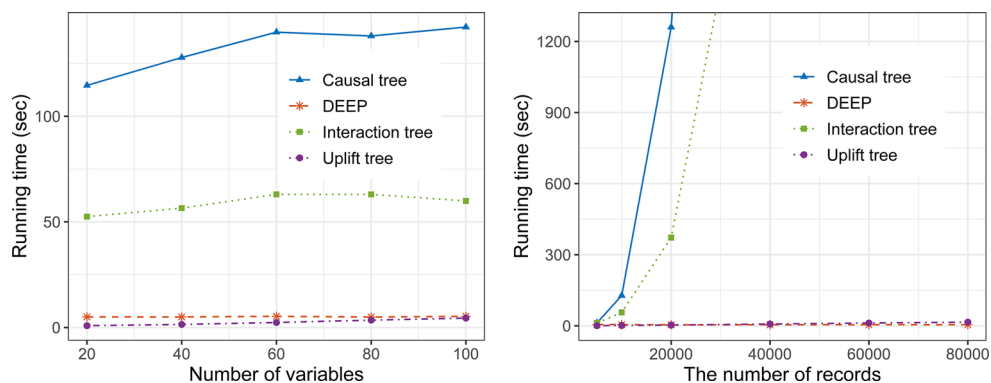
CATEs are commonly analysed to detect treatment effect heterogeneity and we are interested in data driven analysis. Su et al. [37] used recursive partitioning to construct the interaction tree (IT) for treatment effect estimation in different subgroups by adapting the CART [25]. Athey et al. [3] proposed to use honest estimation for tree partition and causal effect estimation, and built the Causal Tree (CT) based on the CART [25] to find the subpopulations with heterogeneous treatment effects. Wager and Athey further proposed a random forest based method for causal effect heterogeneity modelling [41]. A meta-learning method [22] was proposed for causal heterogeneity modelling with unbalanced treated and control samples. In recent years, some algorithms have been presented using deep learning techniques [28, 35, 43, 44]. Interesting readers are referred to a survey [16] and an evaluation paper [21].

Uplift modelling is closely linked to causal heterogeneity modelling as shown in [17, 45]. Due to the page limit, we refer readers to the recent surveys [9, 15]. Uplift modelling is normally assumed in data from a well designed randomised experiment and hence probability difference in the treated and control groups has been used as CART without adjustment. Therefore, it is not clear whether the uplift modelling methods can be used in observational data. Again only tree based methods are of our interest because of the interpretability. Rzepakowski and Jaroszewicz adapted decision trees for uplift modelling [33, 34].

A covariate set in causal inference should satisfy the unconfoundedness assumption (i.e. conditional ignorability [32]). VanderWeele and Shpitser [40] have proposed a covariate set to be the union of causes of the treatment

Fig. 4 Time efficiency with the number of variables (left), and with the number of records (right) of four methods

and causes of the outcome without knowing the underlying causal structure. de Luna et al. [8] have proposed a method to reduce a covariate set to the minimal sets under the unconfoundedness assumption, and an implementation based on the Bayesian network has been reported in [18]. Entner et al. [12] have proposed a method to find covariate sets using conditional independence tests. These works focus on ATE estimation instead of CATE estimation and they have not elaborated on the role of confounders and effect modifiers in CATE estimation. PC (parent and child) discovery algorithms, such as PC-Select [6], MMPC (Max-Min Parents and Children) [38] and HITON-PC [1], can be considered covariate selection algorithms when data sets contain pretreatment variables (both ancestral nodes of treatment and the outcome in a causal graph term).

Causal rules [26, 46] and causal patterns [42] concern multiple treatments, not causal heterogeneity.

7 Conclusions

We have proposed TEPs to represent treatment effect heterogeneity in a population. TEPs encode the local causal structure which gives users an overview of causal relationships around the outcome variable. Users can evaluate TEPs discovered in data based on the consistency between the local causal structure and their domain knowledge, and can also use their believed local causal structure to guide TEP discovery. We have developed the DEEP algorithm to identify TEPs using a bottom up approach which ensures that each TEP is as specific as possible while its subgroup has the smallest possible treatment effect heterogeneity. When using the discovered TEPs, the most specific TEP matching a person's situation is used for personalised decision making. The experiments show that the DEEP models the treatment heterogeneity better than three existing tree based methods in both synthetic and real world data sets and DEEP is efficient among the comparison methods. Our future work will apply the DEEP to assist personalised decision making in various applications and extend the TEP for other types of variables other than binary variables.

Data Availability Statements The four real world data sets are publicly available from the links provided in the paper. The synthetic data sets will be available from the corresponding author on reasonable request.

References

1. Aliferis C, Tsamardinos I, Statnikov A (2003) Hiton: a novel Markov blanket algorithm for optimal variable selection. In: AMIA Annual symposium proceedings, vol 2003. American Medical Informatics Association, pp 21–25
2. Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD (2010) Local causal and Markov blanket induction for causal discovery and feature selection for classification Part I: algorithms and empirical evaluation. J Mach Learn Res 11:171–234
3. Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. Proc Natl Acad Sci 113(27):7353–7360
4. Bache K, Lichman M (2013) UCI machine learning repository. http://archive.ics.uci.edu/ml
5. Bay SD, Pazzani MJ (2001) Detecting group differences: mining contrast sets. J Data Mining Knowl Discov 5(3):213–246
6. Bühlmann P, Kalisch M, Maathuis MH (2010) Variable selection in high-dimensional linear models: partially faithful distributions and the pc-simple algorithm. Biometrika 97(2):261–278
7. Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) Introduction to algorithms. MIT Press and McGraw-Hill, Cambridge
8. De Luna X, Waernbaum I, Richardson TS (2011) Covariate selection for the nonparametric estimation of an average treatment effect. Biometrika 98(4):861–875
9. Devriendt F, Moldovan D, Verbeke W (2018) A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: a stepping stone toward the development of prescriptive analytics. Big Data 6(1):13–41
10. Eustache D, Artem B, Renaudin C, Massih-Reza A (2018) A large scale benchmark for uplift modeling. In: Proceedings of the AdKDD and TargetAd workshop. ACM
11. Dong G, Li J (1999) Efficient mining of emerging patterns: discovering trends and differences. In: Proceedings of ACM international conference on knowledge discovery and data mining (KDD), pp 43–52
12. Entner D, Hoyer P, Spirtes P (2013) Data-driven covariate selection for nonparametric estimation of causal effects. In: Artificial intelligence and statistics, pp 256–264
13. Fleiss JL, Levin B, Paik MC (2003) Statistical methods for rates and proportions, 3rd edn. Wiley
14. Gan W, Lin JC, Fournier-Viger P, Chao H, Tseng VS, Yu PS (2021) A survey of utility-oriented pattern mining. IEEE Trans Knowl Data Eng 33, 4:1306–1327
15. Gubela R, Bequé A, Lessmann S, Gebert F (2019) Conversion uplift in e-commerce: as systematic benchmark of modeling strategies. Int J Inf Technol Decis Mak 18(03):747–791
16. Guo R, Cheng L, Li J, Hahn PR, Liu H (2020) A survey of learning causality with data: problems and methods. ACM Comput Surv 53(4):75:1–75:37
17. Gutierrez P, Gérardy J-Y (2017) Causal inference and uplift modelling: a review of the literature. In: Proceedings of international conference on predictive applications and APIs, pp 1–13

18. Häggström J (2018) Data-driven confounder selection via Markov and Bayesian networks. Biometrics 74(2):389–398
19. Hill JL (2011) Bayesian nonparametric modeling for causal inference. J Comput Graph Stat 20(1):217–240
20. Imbens GW, Rubin DB (2015) Causal inference for statistics, social, and biomedical sciences. Cambridge University Press
21. Knaus MC, Lechner M, Strittmatter A (2021) Machine learning estimation of heterogeneous causal effects: empirical Monte Carlo evidence. Econ J 24:134–161
22. Künzel SR, Sekhon JS, Bickel PJ, Yu B (2019) Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci 116(10):4156–4165
23. Lagani V, Athineou G, Farcomeni A, Tsagris M, Tsamardinos I (2017) Feature selection with the R package MXM: discovering statistically equivalent feature subsets. J Stat Softw 80(7): 1–25
24. Lemmerich F, Puppe F (2011) Local models for expectation-driven subgroup discovery. In: Proceedings of international conference on data mining, pp 360–369
25. Leo B, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth International Group 8:452–456
26. 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):1–27
27. Li J, Liu J, Toivonen H, Satou K, Sun Y, Sun B (2014) Discovering statistically non-redundant subgroups. Knowl-Based Syst 67:315–327
28. Louizos C, Shalit U, Mooij J, Sontag D, Zemel R, Welling M (2017) Causal effect inference with deep latent-variable models. In: Proceedings of international conference on neural information, pp 6449–6459
29. Miettinen K (1998) Nonlinear multiobjective optimization. Springer US
30. Nawaz MS, Fournier-Viger P, Yun U, Wu Y, Song W (2022) Mining high utility itemsets with hill climbing and simulated annealing. ACM Trans Manag Inform Syst 13(1):4:1–4:22
31. Pearl J (2009) Causality: models, reasoning, and inference, 2nd edn. Cambridge University Press
32. Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55
33. Rzepakowski P, Jaroszewicz S (2010) Decision trees for uplift modeling. In: IEEE International conference on data mining, pp 441–450
34. Rzepakowski P, Jaroszewicz S (2012) Decision trees for uplift modeling with single and multiple treatments. Knowl Inf Syst 32(2):303–327
35. Shalit U, Johansson FD, Sontag D (2017) Estimating individual treatment effect: generalization bounds and algorithm. In: Proceedings of international conference on machine learning, pp 3076–3085
36. Spirtes P, Glymour CC, Scheines R (2000) Causation, predication, and search, 2nd edn. The MIT Press
37. Su X, Tsai C-L, Wang H, Nickerson DM, Li B (2009) Subgroup analysis via recursive partitioning. J Mach Learn Res 10:141–158
38. Tsamardinos I, Brown LE, Aliferis CF (2006) The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn 65(1):31–78
39. VanderWeele TJ, Robins JM (2007) Four types of effect modification: a classification based on directed acyclic graphs. Epidemiology 18(5):561–568
40. VanderWeele TJ, Shpitser I (2011) A new criterion for confounder selection. Biometrics 67(4):1406–1413
41. Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc 113(523):1228–1242
42. Yadav P, Steinbach M, Castro MR, Caraballo PJ, Kumar V, Simon G (2019) Frequent causal pattern mining: a computationally efficient framework for estimating bias-corrected effects. In: Proceedings of IEEE international conference on big data, pp 1981–1990
43. Yao L, Li S, Li Y, Huai M, Gao J, Zhang A (2018) Representation learning for treatment effect estimation from observational data. In: Advances in neural information processing systems, pp 2638–2648
44. Yoon J, Jordon J, van der Schaar M (2018) GANITE: estimation of individualized treatment effects using generative adversarial nets. In: Proceedings of international conference on learning representations
45. Zhang W, Li J, Liu L (2022) A unified survey on treatment effect heterogeneity modeling and uplift modeling. ACM Comput Surv 54(8):1–36
46. Jin Z, Li J, Liu L, Le TD, Sun B, Wang R (2012) Discovery of causal rules using partial association. In: Proceedings of IEEE international conference on data mining, pp 309–318

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
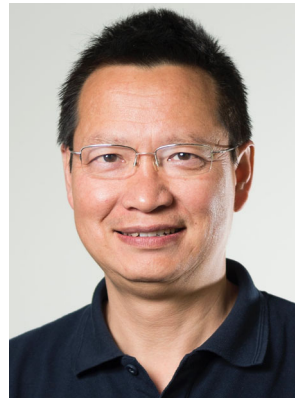
Jiuyong Li received his PhD degree in computer science from Griffith University, Brisbane, Australia, in 2002. He is currently a Professor at the University of South Australia, Australia. His main research interests include data mining, causal discovery, privacy and fairness, and bioinformatics. His research work has been supported by seven Australian Research Council Discovery projects and he has led several industry projects.

Lin Liu received her PhD in computer systems engineering from the University of South Australia (UniSA) in 2006. She is currently an Associate Professor at UniSA. Her research interests include data mining, causal inference, and bioinformatics.
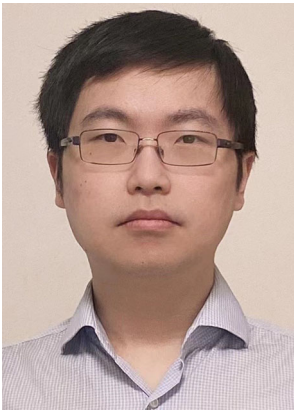
Shisheng Zhang finished his Bachelor of Science at Tongji University in Shanghai, China. He received his Master of Professional Engineering from the School of Biomedical Engineering at the University of Sydney in Australia. After that, Shisheng worked as a research student in STEM at the University of South Australia for a year. He is currently a PhD student at the University of New South Wales. His research interests include machine learning and medical image analysis.

Thuc Duy Le is an associate professor at UniSA STEM, University of South Australia. He is an ARC DECRA fellow (2020–2022) and was an NHMRC ECR fellow (2017–2019). His research interests include causal discovery methods and their applications in Bioinformatics.

Saisai Ma received the bachelor's and master's degrees in electronic engineering from Hohai University, China, in 2010 and 2013 respectively, and the PhD degree in computer science from the University of South Australia in 2017. Currently He is a senior data scientist with Consilium Technology. His research interests include data mining, machine learning, causal inference and their applications.

Jixue Liu got his PhD in computer science from the University of South Australia. He is an Associate Professor in the University. He has published widely in databases and artificial intelligence. His work covers the topics of integrity constraint discovery, data analytics in texts and time series, deep learning in engineering areas, entity linking, fairness computing, privacy in data, XML functional dependencies and data integration and transformation.