Causal Decision Trees

Jiuyong Li, Saisai Ma, Thuc Le, Lin Liu, and Jixue Liu

Abstract—Uncovering causal relationships in data is a major objective of data analytics. Currently there is a need for scalable and automated methods for causal relationship exploration in data. Classification methods are fast and they could be practical substitutes for finding causal signals in data. However, classification methods are not designed for causal discovery and a classification method may find false causal signals and miss the true ones. In this paper, we develop a causal decision tree (CDT) where nodes have causal interpretations. Our method follows a well established causal inference framework and makes use of a classic statistical test to establish the causal relationship between a predictor variable and the outcome variable. At the same time, by taking the advantages of normal decision trees, a CDT provides a compact graphical representation of the causal relationships, and the construction of a CDT is fast as a result of the divide and conquer strategy employed, making CDTs practical for representing and finding causal signals in large data sets. Experiment results demonstrate that CDTs can identify meaningful causal relationships and the CDT algorithm is scalable.

Index Terms—Decision tree, Causal relationship, Potential outcome model, Partial association

1 Introduction

Causal relationships can provide better insights into data, as well as actionable knowledge for correct decision making and timely intervening in processes at risk, therefore detecting causal relationships in data is an important data analytics task.

Randomised controlled trials (RCTs) are considered as the gold standard for causal inference in many areas such as medicine and social science [1]. However, it is often impossible to conduct RCTs due to cost or ethical concerns. Causal relationships can also be found by observational studies, such as cohort studies and case control studies [2]. An observational study takes a causal hypothesis and tests it using samples selected from historical data or collected passively over the period of time when observing the subjects of interest. Therefore observational studies need domain experts' knowledge and interactions in data selection or collection, and the process is normally time consuming.

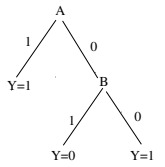
Currently there is a lack of scalable and automated methods for causal relationship exploration in data. These methods should be able to find causal signals in data without requiring domain knowledge or any hypothesis established beforehand. The methods must also be efficient to deal with the increasing amount of data.

Classification methods are fast and have the potential to become practical substitutes for detecting causal signals in data since finding causal relationships is a type of supervised learning when the outcome variable is fixed. Decision trees [3] are a good example of classification methods, and they have been widely used in many areas, including social and medical data analyses.

However, classification methods are not designed with causal discovery in mind. We use the following example to illustrate how a perfect decision tree may not code causal relationships. Figure 1 shows a hypothesised data set and a decision tree built from the data set. The decision tree perfectly classifies the outcome (i.e. Y) using attributes A and B. For example, the path $(A=1) \rightarrow (Y=1)$ with $\text{prob}(Y=1|A=1))=1$ correctly classifies more than half of the records in the data set, but the path does not code a causal relationship between A and Y since, for example, given $C=1$, $\text{prob}(Y=1|A=1)-\text{prob}(Y=1|A=0) = 0$. In other words, when fixing the value of C (aiming to exclude the possible effect of C on Y), a change of A does not result in a change in Y. Such a principle for causal inference is frequently used in our everyday reasoning. For example, we do not conclude that female employees are discriminated just based the observation that on average the salaries of females are lower than those of males. Instead, we need the evidence showing that salary differences can be observed between female and male employees with the same occupations, education levels and work experience (such that the effects of other factors possibly affecting salaries are eliminated). In this data set, there is not sufficient evidence to enable us to draw a causal conclusion (we will give more detailed analysis on this data set in Section 4).

A	B	C	Y	count
0	1	0	0	20
0	0	1	1	20
1	1	1	1	10
1	0	1	1	10
1	0	0	1	20

(a)



(b)

Fig. 1. An example showing that a decision tree does not encode causal relationships (a) An exemplar data set (b) A decision tree of the data set

Causal discovery methods [4], [5] have shown promise for finding real and non-spurious relationships in data although it is arguable whether real causes can be found in data. The major difference between the causal relationship discovery methods and other relationship discovery methods is that the former assesses the relationship between a (potential) cause variable and an outcome variable by removing the effects of ther covariates on the outcome but the latter only considers the relationship between the potential cause variable and the outcome in isolation.

In this paper, we design a causal decision tree (CDT) where nodes have causal interpretations. The paths in a CDT are not interpreted as 'if-then' first order logic rules as in a normal decision tree. Each non-leaf node of a CDT has a causal relationship with

• *All authors are with the School of IT and Mathematical Sciences, University of South Australia, Mawson Lakes, SA 5095, Australia.*
E-mails: {jiuyong.li, thuc.le, lin.liu, jixue.liu}@unisa.edu.au, saisai.ma@mymail.unisa.edu.au

Manuscript received xxx, 2016; revised xxx, 2016.

the outcome variable.

A CDT is not about classification but interpretation. A normal decision tree offering high classification accuracy may provide wrong reasons, and the knowledge represented by it may not be actionable since no real causes are given and thus we may not prevent an outcome from happening. Note that a discriminative feature contributing greatly to classification may not be a causal factor of the outcome since the high correlation between the feature and the outcome may be due to, e.g. a common cause of them or an intermediate variable. In contrast, a CDT tries to provide real causes of an outcome although it is not optimised for classification accuracy. The aim of CDTs is to provide interpretable and actionable knowledge to reduce the effects of undesirable outcomes or to promote the effects of desirable outcomes.

CDTs are developed to take the advantage of efficient tree based search. Graphical causal models, particularly Causal Bayesian networks [5] have achieved great success in causal discovery in the past decades, but efficiency is still a major concern. Our work is closely related to causal Bayesian networks. We will detail the difference of our CDT from Bayesian networks and other Bayesian network related tree representations in Section 7.1.

The potential outcome model [5], [6], [7] forms a base of our causal inference for determining the causal relationship between a predictor variable and the outcome variable when building a CDT. Several causal models have been well established, such as causal Bayesian networks, the potential outcome model and the structural equation model [5]. These models take different approaches for representing and inferring causal relationships, but they have been shown to be complementary and closely related with commonalities at the same time [8]. The potential outcome model is widely accepted for causal inference in epidemiological and social applications, but it has not been commonly applied to large scale data mining for causal signals in an automated manner. With the model, the causal effect of a treatment on the outcome is estimated as the difference between the potential outcome with the treatment and the potential outcome without the treatment. The CDT method evaluates the causal effect of a predictor variable on the outcome using a classic statistical test for partial associations, the Mantel-Haenszel test [9], and considers that there exists a causal relationship between the predictor variable and the outcome if the causal effect is significant.

A number of assumptions are essential to support the claim of a causal relationship. Two fundamental ones that are closely related to our work are the causal sufficiency and faithfulness assumptions [4], [10]. Causal sufficiency requires that all variables are observed and measured and there are no hidden variables. This is to ensure that no hidden cause results in a causal relationship. The faithfulness assumption in our work is that a dependency in data will be discovered by a CDT, and a causal relationship identified by a CDT reflects a true dependency in data. In addition, we also assume that the outcome variable is not a cause of any of the predictor variables.

The main contributions of this paper are as follows:

- We systematically analyse the limitations of decision trees for causal discovery and identify the underlying reasons.
- We propose the CDT method for effectively representing and identifying interpretable causal relationships in data, including context specific causal relationships.

2 Related work

Discovering causal relationships in passively observed data has attracted enormous research efforts in the past decades, due to the high cost, low efficiency and unknown feasibility of experiment based approaches, as well as the increasing availability of observational data. To the credit of the theoretical development by a group of statisticians, philosophers and computer scientists, including Pearl [5], Spirtes, Glymour [11] and others, we have seen graphical causal models playing dominant role in causality discovery. Among these graphical models, causal Bayesian networks (CBNs) [4] have been the most developed and used one.

Many algorithms have been developed for learning CBNs [4], [12]. However in general learning a complete CBN is NP-hard [13] and the methods are able to handle a CBN with only tens of variables, or hundreds if the causal relationships are sparse [4].

Consequently, local causal relationship discovery around a given target (outcome) variable has been actively explored recently as in practice we are often more interested in knowing the direct causes or effects of a variable, especially in the early stage of investigations. The work presented in this paper is along the line of local causal discovery.

Existing methods for local causal discovery around a given target fall into two broad categories: (1) Methods that adapt the algorithms or ideas for learning a complete CBN into local causal discovery, such as PC-Simple [14], [15], a simplified version of the well-known PC algorithm [11] for CBN learning; and HITON-PC [16], which applies the basic idea of PC to find variables strongly (and causally) related to a given target; (2) Methods that are designed to exploit the high efficiency of popular data mining approaches and the causal discovery ability of traditional statistical methods, including the work in [17] and [18], both using association rule mining for identifying causal rules; and the decision tree based approach [19] for finding the Markov blanket of a given variable.

The CDT proposed in this paper belongs to the second category, as it takes advantage of decision tree induction and partial association tests. Comparing to other methods in the category, however, the proposed CDT approach is distinct because it is aimed at finding a sequence of causal factors (variables along the path from the root to a leaf of a CDT) where a preceding factor is a context under which the following factors can have impact on the target, while the other methods identify a set of causal factors each being a cause or an effect of the given target, and they only discover global causal relationships. However, in practice, a variable may not be a cause of another variable globally, but under certain context, it may affect other variables. A CDT provides a way to identify such context specific causal relationships. Additionally because a context specific causal relationship contains information about the conditions in which a causal relationship holds, such relationships are more prescriptive and actionable and thus are more suitable for decision support and action planning.

In terms of using decision trees as a means for causality investigation, except from the above mentioned method for identifying Markov blankets [19], most existing work takes decision trees as a tool for causal relationship representation and/or inference, assuming that the causal relationships are known in advance. Examples include the CPT-trees [20] and causal explanation tree [21] to be discussed in detail in Section 7.1, which are both derived from a known causal Bayesian network. Recently there has been an increasing interest in applying machine learning methods,

including tree-based methods for estimating the heterogeneity of causal effects in different sub-populations [22], [23], [24], [25]. For example, in [25], regression trees are used to partition the population into subsets, and then in each sub-population the causal effect of a known cause is estimated, so that the differences of the effects across the sub-populations can be investigated. However, these methods also assume a known cause, and their focus is on finding the proper sub-populations or contexts so that valid inference of the heterogeneous causal effects can be carried out with respect to the contexts. Unlike all these trees, our CDT is mainly used as a tool for detecting causal relationships in data, without any assumption of known causal relationships.

3 Cause and Effect in the Potential Outcome Framework

Let X be a predictor variable or attribute and Y the outcome attribute where $x \in \{0, 1\}$ and $y \in \mathbb{R}_{\geq 0}$. We aim to identify if there is a causal relationship between X and Y. For easy discussion, we consider that $X = 1$ is a treatment and $Y = 1$ the recovery. We will establish if the treatment is effective for the recovery.

The potential outcome or counterfactual model [5], [6], [7] is a well established framework for causal inference. Here we introduce the basic concepts of the model and a principle for estimating the average causal effect, mainly following the introduction in [26].

With the potential outcome model, an individual i in a population has two potential outcomes for a treatment X: Y_i^1 when taking the treatment and Y_i^0 when not taking the treatment. We say that Y_i^1 is the potential outcome in the treatment state and Y_i^0 is the potential outcome in the control state. Then we have the following definition.

Definition 1 (Individual level causal effect (ICE)) *The individual level causal effect is defined as the difference of two potential outcomes of an individual, i.e.* $\delta_i = Y_i^1 - Y_i^0$.

In practice we can only find out one outcome Y_i^1 or Y_i^0 since one person can be placed in either the treatment group ($X = 1$) or the control group ($X = 0$). One of the two potential outcomes has to be estimated. So the potential outcome model is also called counterfactual model. For example, we know that Mary has a headache (the outcome) and she did not take aspirin (the treatment), i.e. we know Y_i^0. The question is what the outcome would be if Mary took aspirin one hour ago, i.e. we want to know Y_i^1 and to estimate the ICE of aspirin on Mary's condition (having headache or not).

If we had both Y_i^1 and Y_i^0 of an individual we would aggregate the causal effects of individuals in a population to get the average causal effect as defined below, where $E[.]$ stands for the expectation operator in probability theory.

Definition 2 (Average causal effect (ACE)) *The average causal effect of a population is the average of the individual level causal effects in the population, i.e.* $E[\delta_i] = E[Y_i^1] - E[Y_i^0]$.

Note that i is kept in the above formula as other work in the counterfactual framework to indicate individual level heterogeneity of potential outcomes and causal effects.

Assuming that π proportion of samples take the treatment and $(1-\pi)$ proportion do not, and the sample size is large so the error caused by sampling is negligible, given a data set $\mathbf{D}$, the ACE, $E[\delta_i]$ can be estimated as:

$$\begin{aligned} E_{\mathbf{D}}[\delta_i] \;=\; & \pi(E_{\mathbf{D}}[Y_i^1|X_i=1] - E_{\mathbf{D}}[Y_i^0|X_i=1]) + \\ & (1-\pi)(E_{\mathbf{D}}[Y_i^1|X_i=0] - E_{\mathbf{D}}[Y_i^0|X_i=0]) \end{aligned} \quad (1)$$

That is, the ACE of the population is the ACE in the treatment group plus the ACE in the control group, where $X_i = 1$ indicates that an individual takes the treatment, and the causal effect is $(Y_i^1|X_i = 1) - (Y_i^0|X_i = 1)$. Similarly, $X_i = 0$ indicates that an individual does not take the treatment, and the causal effect is $(Y_i^1|X_i = 0) - (Y_i^0|X_i = 0)$.

In a data set, we can observe the potential outcomes in the treatment state for those treated, $(Y_i^1|X_i = 1)$, and the potential outcomes in the control state for those not treated, $(Y_i^0|X_i = 0)$. However, we cannot observe the potential outcomes in the control state for those treated, $(Y_i^0|X_i = 1)$, or the potential outcomes in the treatment state for those not treated, $(Y_i^1|X_i = 0)$. We have to estimate what the potential outcome, $(Y_i^0|X_i = 1)$, would be if an individual did not take the treatment (in fact she has); and what potential outcome, $(Y_i^1|X_i = 0)$, would be if an individual took the treatment (in fact she has not).

With a data set $\mathbf{D}$ we can obtain the following 'naïve' estimation of the ACE:

$$E_{\mathbf{D}}^{naive}[\delta_i] = E_{\mathbf{D}}[Y_i^1|X_i=1] - E_{\mathbf{D}}[Y_i^0|X_i=0] \quad (2)$$

The question is when the naïve estimation (Equation (2)) will approach the true estimation (Equation (1)).

If the assignment of individuals to the treatment and control groups is purely random, the estimation in Equation (2) approaches the estimation in Equation (1). In an observational data set, however, the random assignment is not possible. How can we estimate the average causal effect? A solution is by perfect stratification. Let the differences of individuals in a data set be characterised by a set of attributes $\mathbf{S}$ (excluding X and Y) and let the data set be perfectly stratified by $\mathbf{S}$. In each stratum, apart from the fact of taking treatment or not, all individuals are indistinguishable from each other. Under the perfect stratification assumption, we have:

$$E[Y_i^1|X_i=0, \mathbf{S}=\mathbf{s}_i] = E[Y_i^1|X_i=1, \mathbf{S}=\mathbf{s}_i] \quad (3)$$

$$E[Y_i^0|X_i=1, \mathbf{S}=\mathbf{s}_i] = E[Y_i^0|X_i=0, \mathbf{S}=\mathbf{s}_i] \quad (4)$$

where $\mathbf{S} = \mathbf{s}_i$ indicates a stratum of perfect stratification. Since individuals are indistinguishable in the stratum, unobserved potential outcomes can be estimated by observed ones. Specifically, the mean potential outcome in the treatment state for those untreated is the same as that in the treatment state for those treated (Equation (3)), and the mean potential outcome in the control state for those treated is the same as that in the control state for those untreated (Equation (4)). By replacing Equation (1) with Equations (3) and (4), we have:

$$\begin{aligned} & E_{\mathbf{D}}[\delta_i|\mathbf{S}=\mathbf{s}_i] \\ =\; & \pi(E_{\mathbf{D}}[Y_i^1|X_i=1, \mathbf{S}=\mathbf{s}_i] - E_{\mathbf{D}}[Y_i^0|X_i=1, \mathbf{S}=\mathbf{s}_i]) + \\ & (1-\pi)(E_{\mathbf{D}}[Y_i^1|X_i=0, \mathbf{S}=\mathbf{s}_i] - E_{\mathbf{D}}[Y_i^0|X_i=0, \mathbf{S}=\mathbf{s}_i]) \\ =\; & \pi(E_{\mathbf{D}}[Y_i^1|X_i=1, \mathbf{S}=\mathbf{s}_i] - E_{\mathbf{D}}[Y_i^0|X_i=0, \mathbf{S}=\mathbf{s}_i]) + \\ & (1-\pi)(E_{\mathbf{D}}[Y_i^1|X_i=1, \mathbf{S}=\mathbf{s}_i] - E_{\mathbf{D}}[Y_i^0|X_i=0, \mathbf{S}=\mathbf{s}_i]) \\ =\; & E_{\mathbf{D}}[Y_i^1|X_i=1, \mathbf{S}=\mathbf{s}_i] - E_{\mathbf{D}}[Y_i^0|X_i=0, \mathbf{S}=\mathbf{s}_i] \\ =\; & E_{\mathbf{D}}^{naive}[\delta_i|\mathbf{S}=\mathbf{s}_i] \end{aligned} \quad (5)$$

As a result, the naïve estimation approximates the true average causal effect, and we have the following observation.

Observation 1 *[Principle for estimating average causal effect] The average causal effect can be estimated by taking weighted sum of naïve estimators in stratified sub data sets.*

This principle ensures that each comparison is between individuals with no observable differences, and hence the estimated causal effect is not resulted from other factors than the studied one. In the following, we will use this principle to estimate causal effect in observational data sets.

4 From Normal Decision Trees to Causal Decision Trees

Let $\mathbf{X} = \{X_1, X_2, \ldots, X_m\}$ be a set of predictor attributes where $x_i \in \{0, 1\}$ for $1 \leq i \leq m$, and Y be an outcome attribute where $y \in \{0, 1\}$. Data set $\mathbf{D}$ contains n records taking various assignments of values for $\mathbf{X}$ and Y, each of which represents the record of an observation. Let us assume that $\mathbf{X}$ includes all the attributes for characterising an individual, and the data set is large and thus there is no bias in the sampling process.

4.1 Why a decision tree may not encode causal relationships?

Decision trees [3] are a popular classification model, with two types of nodes: branching and leaf nodes. A branching node represents a predictor attribute and each of its values denotes a choice and leads to another branching node or a leaf node representing a class.

The construction of a decision tree follows a divide and conquer strategy. The most important decision to be made in the construction is to select the branching nodes. Aiming at minimising classification error, the basic idea of the selection is as follows. After samples are split at X, it is desired that the class distribution at X's child nodes (in the obtained subsets of samples) is more skewed than that at X before the splitting. That is, the child nodes should be less impure than X regarding class distribution [3]. When comparing X_i and X_j to decide on a better branching attribute, the impurity at X_i and X_j (before splitting) is the same, so the attribute whose child nodes have smaller impurity is preferred.

A direct measure of a node's impurity is the misclassification error, based on which the impurity of a child node of X is defined as $1\text{–}\max(\text{prob}(Y{=}1|X), \text{prob}(Y{=}0|X))$ [3], where $\text{prob}(Y{=}1|X)$ and $\text{prob}(Y{=}0|X)$ are the fractions of positive and negative samples respectively given X. In practice, other measures such as entropy and Gini index are often used [3] as they provide smoother curves than the misclassification error. In the following, for conceptual discussions and easy comparison with the criterion based on causal effects, we still use the misclassification error. Since minimising the misclassification error is equivalent to maximising the absolute value of $(\text{prob}(Y{=}1|X) - \text{prob}(Y{=}0|X))$, we have the following conceptual definition of a discriminative or branching attribute.

Definition 3 (Discriminative attribute) *Given a data set $\mathbf{D}'$, a discriminative attribute is the attribute X_i such that $|\text{prob}(Y = 1|X_i = 1) - \text{prob}(Y = 0|X_i = 1)|$ is maximised.*

Note that $\mathbf{D}'$ is a sub data set (of $\mathbf{D}$) defined by the attribute values in the prefix path of the current branching node under consideration. It is a context specific data set (see Section 4.3 for details).

In the following, we will discuss why a decision tree may not represent causal relationships.

Firstly, The objective of a discriminative attribute (maximising $\text{prob}(Y = 1|X_i = 1)\text{–}\text{prob}(Y = 0|X_i = 1)$) is different from that of a causal factor (having significant causal effect $\text{prob}(Y = 1|X_i{=}1)\text{–}\text{prob}(Y{=}1|X_i{=}0)$).

Secondly, the estimation of $\text{prob}(Y = 1|X_i = 1)\text{–}\text{prob}(Y = 0|X_i = 1)$ for choosing a discriminative attribute is based on the data set $\mathbf{D}'$, while the estimation of the causal effect of X_i on Y is based on the stratified data set $\mathbf{D}_{\mathbf{S}=\mathbf{s}_i}$ to avoid unfair comparison. For example, let X_i be a treatment and Y the recovery, the comparison between individuals with and without treatments have to be in the same age and gender group and have similar medical conditions. Otherwise, the comparison is meaningless. In other words, when a comparison is within a stratum of a stratified data set, the effect of other attributes on Y is eliminated and hence the difference $(\text{prob}(Y = 1|X_i = 1)\text{–}\text{prob}(Y = 1|X_i = 0))$ reflects the causal effect of X_i on Y.

Essentially the main limitation of a decision tree is that it does not consider other attributes in determining a branching attribute. The choice of a branching attribute does not rely on the causal effect of the attribute on the outcome attribute.

Following Observation 1, the estimation of causal effect should be based on stratified data where the difference of individuals in a stratum is eliminated.

Now we analyse that the perfect classification decision tree in Figure 1 does not code causal relationships.

Example 1 In Figure 1, the path $(A = 1) \rightarrow (Y = 1)$, with the difference in probabilities, $(\text{prob}(Y{=}1|A = 1)\text{–}\text{prob}(Y{=}0|A{=}1)){=}1$, represents a top quality discriminative rule. Let us assume that $A = 1$ is a treatment and Y represents the outcome. To derive the average causal effect of A on Y, we need to stratify the data set such that in each stratum the records are indistinguishable with respect to the stratifying attributes. Here $\{B, C\}$ are the stratifying attributes. The data set in Figure 1 is stratified into four strata and their summaries are as follows:

$\{B,C\}$	Y	
$\{0,0\}$	1	0
$A=1$	20	0
$A=0$	0	0

$\{B,C\}$	Y	
$\{0,1\}$	1	0
$A=1$	10	0
$A=0$	20	0

$\{B,C\}$	Y	
$\{1,0\}$	1	0
$A=1$	0	0
$A=0$	0	20

$\{B,C\}$	Y	
$\{1,1\}$	1	0
$A=1$	10	0
$A=0$	0	0

In the first stratum, all records have $B = 0$ and $C = 0$. There are no records from the control group ($A = 0$), hence we cannot estimate the average causal effect in this stratum. Similarly, we cannot estimate causal effects from the third ($B = 1$, $C = 0$) and fourth ($B = 1$, $C = 1$) strata. In the second stratum ($B = 0$, $C = 1$), $E_{\mathbf{D}}[Y^1|A = 1] - E_{\mathbf{D}}[Y^1|A = 0] = 1 - 1 = 0$. All cases regardless they are treated or not treated have the same outcome. So the causal relationship between A and Y cannot be established.

For paths $(A = 0, B = 1) \rightarrow (Y = 0)$ and $(A = 0, B = 0) \rightarrow (Y = 1)$, let us try to establish a causal relationship between B and Y in the sub data set where $A = 0$.

The two strata by attribute C are summarised as:

C	Y	
0	1	0
$B=1$	0	20
$B=0$	0	0

C	Y	
1	1	0
$B=1$	0	0
$B=0$	20	0

In the two strata ($C = 0$ and $C = 1$) there are only cases in either the treatment group ($B{=}1$) or the control group ($B = 0$). There is no way to estimate the average causal effect, so we cannot establish a causal relationship between B and Y.

In this example, we see that a perfect decision tree does not indicate any causal relationship. In other words, in this data set, there is not enough evidence to support causal relationships.

4.2 A measure for causal effect

Based on the previous discussion, to estimate the causal effect of a predictor attribute Q on the outcome Y, we stratify a data set using $\mathbf{X}\backslash\{Q\}$ so that within each stratum there is no observable difference among the records.

A measure of average causal effect should be able to quantify the difference of outcomes in two groups (treatment and control). For binary outcomes, odds ratio [27] is suitable for measuring the difference of two outcomes. Let the following table summarise the statistics of the k-th stratum, $\mathbf{s}_k$.

$\mathbf{s}_k$	$Y=1$	$Y=0$	total
$Q=1$	n_{11k}	n_{12k}	n_{1k}
$Q=0$	n_{21k}	n_{22k}	n_{2k}
total	$n_{.1k}$	$n_{.2k}$	$n_{..k}$

The odds ratio (measuring the difference of Y between groups $Q = 1$ and $Q = 0$) in the k-th the stratum is $(n_{11k}n_{22k})/(n_{12k}n_{21k})$ or equivalently, $\ln(n_{11k}) + \ln(n_{22k}) - \ln(n_{12k}) - \ln(n_{21k})$.

A question is how to get the aggregated difference over all the strata of a data set. Partial association test [28] is a means to achieve this. Over all the r strata of a data set, the difference can be summarised as:

$$\mathrm{PAMH}(Q,Y) = \frac{(|\sum_{k=1}^{r} \frac{n_{11k}n_{22k} - n_{21k}n_{12k}}{n_{..k}}| - \frac{1}{2})^2}{\sum_{k=1}^{r} \frac{n_{1.k}n_{2.k}n_{.1k}n_{.2k}}{n_{..k}^2(n_{..k}-1)}} \quad (6)$$

This is the test statistic of the Mantel-Haenszel test [9], [28]. The test is used to find direct causal relationship between two variables [28]. The test statistic has a Chi-square distribution (degree of freedom=1). Given a significance level α, if $\mathrm{PAMH}(Q,Y) \geq \chi^2_\alpha$, the null hypothesis that Q and Y are independent in all strata is rejected and the partial association between Q and Y is significant. An example of Mantel-Haenszel test is given in Example 2 in the next section.

4.3 Causal decision trees

Our aim is to build a causal decision tree (CDT) where a non-leaf node represents a causal attribute, an edge denotes an assignment of a value of a causal attribute, and a leaf represents an assignment of a value of the outcome. A path from the root to a leaf represents a series of assignments of values of the attributes and a highly probable outcome value as the leaf.

A CDT differs from a normal decision tree in that each of its non-leaf nodes has a causal interpretation with respect to the outcome, i.e. a non-leaf node and the outcome attribute have a context specific causal relationship as defined below.

Definition 4 (Context) *Let $\boldsymbol{P} \subset \boldsymbol{X}$, then a value assignment of $\boldsymbol{P}$, $\boldsymbol{P} = \boldsymbol{p}$, is called a context and $(\boldsymbol{D}|\boldsymbol{P} = \boldsymbol{p})$ is a context specific data set where $\boldsymbol{P} = \boldsymbol{p}$ holds for all records in $\boldsymbol{D}$.*

Definition 5 (context specific causal relationship) *Let $\boldsymbol{P} = \boldsymbol{p}$ be a context and Q be a predictor attribute and $\boldsymbol{P} \cap \{Q\} = \emptyset$. Q and the outcome attribute Y have a context specific causal relationship if $\mathrm{PAMH}(Q,Y)$ is greater than a threshold in the context specific data set $(\boldsymbol{D}|\boldsymbol{P} = \boldsymbol{p})$.*

A context specific causal relationship between the root node and the outcome attribute of a CDT is global or context free, i.e. the context attribute set $\mathbf{P}$ is empty, and a context specific causal relationships between a non-root node A and the outcome Y is a refinement of the causal relationship between A's parent and Y. For example, with the CDT in Figure 3, the causal relationship between the root 'age<30' and the outcome '>50K' is context free, while the causal relationship 'education-num>12' having with the outcome is in the context of 'age<30' being no, which is a refinement of the causal relationship 'age<30' (parent of 'education-num>12') having with the outcome and is a more specific relationship.

Definition 6 (Causal decision tree (CDT)) *In a causal decision tree, a non-leaf node Q represents a context specific causal relationship between Q and the outcome Y where the context is a series of value assignments of the attributes along the path from the root and to the parent of Q. A leaf node represents a value assignment of Y, which is the most probable value of Y in the context specific data set where the context is a series of value assignments of the attributes along the path from the root to the leaf.*

We use the following example to show that a CDT encodes causal relationships.

Example 2 From the data set shown in Figure 2, for path $(A = 1) \rightarrow (Y = 1)$ of the CDT, we have the following summaries of the strata in terms of attributes $\{B, C\}$:

$\{B,C\}$	Y	
$\{0,0\}$	1	0
$A=1$	10	0
$A=0$	10	0

$\{B,C\}$	Y	
$\{0,1\}$	1	0
$A=1$	0	5
$A=0$	20	0

$\{B,C\}$	Y	
$\{1,0\}$	1	0
$A=1$	10	0
$A=0$	0	10

$\{B,C\}$	Y	
$\{1,1\}$	1	0
$A=1$	15	0
$A=0$	0	20

We now calculate the Mantel-Haenszel test statistic (Equation (6)), $\mathrm{PAMH}(A,Y)$. The first table above (for the stratum $B = 0$, $C = 0$) does not contribute to the calculation of $\mathrm{PAMH}(A,Y)$ since it has one column of zero values.

In the stratum $B = 0$ and $C = 1$,

$$\frac{n_{11k}n_{22k} - n_{21k}n_{12k}}{n_{..k}} = \frac{0*0 - 20*5}{25} = -4$$

$$\frac{n_{1.k}n_{2.k}n_{.1k}n_{.2k}}{n_{..k}^2(n_{..k}-1)} = \frac{(20*5*5*20)}{25^2(25-1)} = 0.667$$

Similarly, we compute the intermediate results for strata ($B = 1$, $C = 0$) and ($B = 1$, $C = 1$), and obtain $\mathrm{PAMH}(A,Y) = 17.5$. For $\alpha = 0.05$ or $\chi^2_\alpha = 3.84$, since $17.5 > 3.84$, A and Y have a causal relationship based on the test.

In the context $A = 0$, we test if B and Y have a causal relationship, based on the following summaries of the data set:

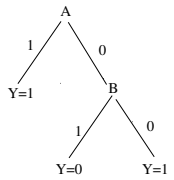
C	Y	
0	1	0
$B=1$	0	10
$B=0$	10	0

C	Y	
1	1	0
$B=1$	0	20
$B=0$	20	0

From the above tables, we have $\mathrm{PAMH}(B,Y){=}49{>}3.84$ in the context specific data set for $A{=}0$. So we can conclude that B and Y have a causal relationship in the context of $A{=}0$.

A	B	C	Y	count
0	0	0	1	10
0	0	1	1	20
0	1	0	0	10
0	1	1	0	20
1	0	0	1	10
1	0	1	0	5
1	1	0	1	10
1	1	1	1	15

(a)



(b)

Fig. 2. An example showing that a CDT represents causal relationships (a) An exemplar data set. (b) a CDT of the data

4.4 Dealing with high dimensional data

The key to causal effect estimation is to remove covariates' effects on the outcome. In the previous section, we use perfect stratification to remove the effect of other variables. However, for a high-dimensional data set, perfect stratification will produce too many strata, each of which has a small size. As a result, the statistical power for detecting dependency in data is lost, and many small strata will result in false negatives.

An alternative to perfect stratification is stratification or subclassification on propensity scores [29], [30], [31], which allocates individuals (samples) to different strata based on their propensity scores such that the propensity scores of the individuals in the same stratum are similar. An individual's propensity score represents the probability of the individual receiving the treatment conditioning on the observed values of the covariates. Suppose that X is a binary variable with values 1 or 0, representing an individual receiving and not receiving the treatment respectively, and $\boldsymbol{C}$ is the set of covariates, for an individual, given that $\boldsymbol{C}$=$\boldsymbol{c}$, the individual's propensity score, $e(\boldsymbol{c})$ is defined as [29]:

$$e(\boldsymbol{c}) = P(X = 1 \mid \boldsymbol{C} = \boldsymbol{c})$$

Since within a stratum the probability of each individual receiving the treatment is similar, we can still follow Observation 1 in Section 3 to estimate the causal effect of a predictor on the outcome by aggregating the causal effects across all the strata. Specifically, for high dimensional data, we can also use Equation (6) and the partial association test to detect a causal relationship, where the strata are those obtained by stratification on propensity scores, instead of those obtained by perfect stratification.

To apply stratification on propensity scores, we need solve two problems: choosing a method to estimate propensity scores; and deciding the granularity of the strata for the stratification.

Logistic regression is commonly used for estimating propensity scores [29], [32]. For the regression, the treatment variable is considered as the response variable and other variables as independent variables, while the outcome variable is ignored.

Fine-grained subclassification will lead to a larger number of smaller subgroups, resulting in a loss of the statistical power of stratification, while coarse-grained stratification will lead to bias in causal effect estimation. It has been shown that subclassification with 5 subgroups can remove at least 90% of the bias due to all the covariates in causal effect estimation [30], [33]. The use of 5 to 10 subgroups is a current convention [31].

4.5 Inference using causal decision trees

A causal decision tree represents context causal relationships and each non-leaf node Q is considered as a cause of Y given $\mathbf{P}$ denoting the precedent nodes of Q. Based on Equation (5), the average causal effect of Q on Y given $\mathbf{P}$ can be calculated as the following.

$$ACE(Q \rightarrow Y|\mathbf{P}{=}\mathbf{p}) = \sum_k \frac{n_k}{|(\mathbf{D}|\mathbf{P}{=}\mathbf{p})|}(\text{prob}(Y{=}1|Q{=}1,\mathbf{S}{=}\mathbf{s}_k) - \text{prob}(Y=1|Q=0,\mathbf{S}=\mathbf{s}_k))$$

where n_k is the size of the k-th stratum $\mathbf{s}_k$ in the context specific data set $(\mathbf{D}|\mathbf{P}=\mathbf{p})$.

5 The CDT Algorithms

Normal decision trees have the following advantages: (1) The divide and conquer strategy of decision tree induction is very efficient. A decision tree construction algorithm is scalable to the data set size and the number of attributes. This is a major advantage in exploring data; (2) Decision trees explore both global and context specific relationships, and the latter provides refined explanations for the former. They jointly provide comprehensive explanations for a data set.

Therefore in this paper we exploit these advantages for exploring causal relationships. However, the challenges for building a CDT include: (1) The criterion for choosing a branching attribute for a normal decision tree needs to be replaced by a causality based criterion. Based on the potential outcome model, we estimate the average causal effect of a treatment variable on the outcome variable by aggregating the causal effects of the treatment in all subgroups of the data set stratified based on the covariates. Specifically, we use the Mantel-Haenszel test for aggregating the causal effects in all the strata and for detecting significant causal effect, thus to determine the causal relationship between the treatment variable and the outcome variable. To obtain the stratified data, perfect stratification can be used, but for high dimensional data, (approximate) stratification on propensity scores should be applied; (2) The time complexity for the Mantel-Haenszel test using perfect stratification is quadratic to the size of a data set since all strata must be found in the first place. We propose to use quick sort to facilitate the stratification, which reduces the time complexity greatly. When data records are sorted by the values of relevant (control) variables, the records which have the same values of the control variables will be put together, forming the strata of the perfect stratification.

Based on the above discussions, in the following we present the CDT algorithm in two versions: the first version does perfect stratification (so it is called CDT-PS) and uses quick sort for improving the efficiency, and the second one does stratification on propensity scores (called CDT-SPS). The time complexity of the two CDT algorithms are analysed at the end of this section.

5.1 CDT by perfect stratification

As shown in Algorithm 1, CD-PS takes 3 inputs: the data set $\mathbf{D}$ for a set of predictor attributes $\mathbf{X}$ and one outcome attribute Y; user specified confidence level for Mantel-Haenszel test and correlation test (to find relevant or stratifying attributes); and the maximum height of the CDT. Having a maximum tree height makes the tree more interpretable. If we do not restrict the tree height, we can get a context which includes many attributes, and a causal relationship in such a context only explains a very specific scenario and has less interest to users.

Algorithm 1 firstly initiates the CDT (i.e. $\mathbf{T}$), and sets the count of the height of the tree (i.e. h) as zero in Line 1. Then the functions TreeConstruct and TreePruning are called subsequently. Finally, the CDT is returned.

The treeConstruct function uses a recursive procedure to construct a CDT and it takes 5 inputs: current node N to be expanded or terminated; the set of attributes $\mathbf{Z} \subseteq \mathbf{X}$ to expand the current subtree (whose root is N) and $\mathbf{Z}$ contains only the attributes that have not been used in the tree; the context specific data set $\mathbf{D}'$, where the context is the value assignments along the path from the root to N (inclusive); h, current height of tree up to N; and e, label of the edge from N to the next node to be expanded.

Lines 1 to 4 of function treeConstruct terminate N if no attribute is left in $\mathbf{Z}$ and/or the height of N reaches the maximum tree height. N is terminated by attaching to it a pair of leaves with edges of 1 and 0 respectively and labelling the leaves with the most probable values in $(\mathbf{D}'|1)$ and $(\mathbf{D}'|0)$ respectively.

If N is not to be terminated, Line 5 finds a set of attributes correlated with Y in the current context specific data $\mathbf{D}'$. The attributes found (called relevant attributes in this paper) are used to stratify $\mathbf{D}'$ for the Mantel-Haenszel test. The reason for choosing correlated attributes for stratification is discussed in Section 7.2.

In Lines 6, 8 and 9, the partial association between Y and each attribute in $\mathbf{Z}$ is tested. As shown in the algorithm, CD-PS does perfect stratification based on the values of the remaining correlated attributes (i.e. $\mathbf{Z} \setminus X_i$).

The attribute, W that has the most significant partial association with Y (i.e. has the largest Mantel-Haenszel test statistic) is selected in Line 11. If the partial association between W and Y is insignificant, in Lines 12 to 15 we terminate N by attaching a pair of leaves with edges of 1 and 0 respectively and labelling the leaves with the most probable values in data sets $(\mathbf{D}'|1)$ and $(\mathbf{D}'|0)$ correspondingly. If the partial association is significant, W is a context specific cause of Y and W is added to the tree in one of the following two ways. If $e = null$, W is set as the root of tree $\mathbf{T}$; otherwise, W is added as a child node of N and the edge between N and W is labelled as e. Line 21 removes W from $\mathbf{Z}$ so it will not be used in the subtree again. In Lines 22 to 24, TreeConstruct is called recursively for W with the context specific data sets $(\mathbf{D}'|W = w)$ where $w \in \{0, 1\}$.

The TreePruning function prunes leaves that do not have distinct labels. The function back traces the tree from the leaf nodes. When two sibling leaves of a parent node share the same label, their parent is converted to a leaf node and is labelled with the same label as their children in Line 3 of the function. Both leaves are then pruned in Line 4.

Algorithm 1 CDT with Perfect Stratification (**CDT-PS**) and CDT with Stratification on Propensity Scores (**CDT-SPS**)

Input: $\mathbf{D}$, a data set for the set of predictor attributes $\mathbf{X} = \{X_1, X_2, \ldots, X_m\}$ and the outcome attribute Y; $h_{\max}$, the maximum height the tree; α, significance level for the Mantel-Haenszel (partial association) test and correlation test; and r, the number of subgroups for stratification on propensity scores (for CDT-SPS only).

Output: $\mathbf{T}$, causal decision tree

1: let $\mathbf{T} = \emptyset$ and $h = 0$
2: TreeConstruct(T, $\mathbf{X}$, $\mathbf{D}$, h, null) // T is the root of $\mathbf{T}$
3: TreePruning($\mathbf{T}$)
4: return $\mathbf{T}$

TreeConstruct(N, $\mathbf{Z}$, $\mathbf{D}'$, h, e)

1: **if** $\mathbf{Z} == \emptyset$ OR $(++h) == h_{\max}$ **then**
2: add two leaf nodes to N with edges $e \in \{1, 0\}$ and label each with the most probable value of Y in $(\mathbf{D}'|N = e)$
3: return
4: **end if**
5: find a set of attributes in $\mathbf{Z}$ that are correlated with Y in $\mathbf{D}'$
6: **for** each correlated attribute X_i **do**
7: calculate propensity scores for each sample of $\mathbf{D}'$ given the stratifying attributes, $\mathbf{Z}\backslash\{X_i\}$ // for CDT-SPS only
8: Stratify samples via QuickSort and specific grouping criterion // for CDT-PS grouping according to the values of $\mathbf{Z}\backslash\{X_i\}$; for CDT-SPS, grouping according to propensity scores and the given number of subgroups, r
9: compute $\mathrm{PAMH}(X_i, Y)$ in stratified $\mathbf{D}'$
10: **end for**
11: find attribute W with the highest partial association test value
12: **if** partial association between W with Y is insignificant **then**
13: add two leaf nodes to N with edges $e \in \{1, 0\}$ and label each with the most probable value of Y in $(\mathbf{D}'|N = e)$
14: return
15: **end if**
16: **if** $e == null$ **then**
17: let node W be the root of $\mathbf{T}$
18: **else**
19: add node W as a child node of N and label the edge between N and W as e
20: **end if**
21: remove W from $\mathbf{Z}$
22: **for** each $w \in \{0, 1\}$ **do**
23: call TreeConstruct(W, $\mathbf{Z}$, $(\mathbf{D}'|W = w)$, h, w)
24: **end for**

TreePruning($\mathbf{T}$)

1: **for** each leaf in $\mathbf{T}$ **do**
2: **if** its sibling leaf has the same label of Y value as itself **then**
3: change their parent node as a leaf node and label itself with the common label
4: remove both leaves
5: **end if**
6: **end for**

5.2 CDT by stratification on propensity scores

As can be seen from Algorithm 1, CDT-SPS essentially follows the same steps as CDT-PS except that:

1) CDT-SPS requires the 4th input, *r*, the user specified number of subgroups of the stratification. As discussed previously, normally *r* is between 5 to 10.
2) While CDT-PS conducts perfection stratification, i.e. directly groups samples with the same values of the covariates ($\mathbf{Z}\backslash\{X_i\}$) into the same strata using QuickSort (see Line 8 of Algorithm 1), CDT-SPS firstly calculates the propensity scores of each sample (Line 7 of the TreeConstruct function) and then stratifies the data set into *r* subgroups based on the propensity scores of the samples. The propensity score of a sample is estimated by doing logistic regression of X_i using $\mathbf{Z} \setminus \{X_i\}$.

5.3 Time complexity

The time complexity of CDT-PS mainly attributes to 3 factors: tree construction, forming perfect strata, and causal tests.

For tree construction, at each split, firstly, we test the correlation of each (unused) attribute with Y, and the complexity is $O(mn)$ where m is the number of predictor attributes and n is the number of samples in the given data set. Then Mantel-Haenszel tests with Y are conducted for all (relevant) attributes, and this is the most expensive part of the algorithm. For each test, the context specific data set $\mathbf{D}'$ is sorted and strata are found in the data set, which has a complexity of $O(n \log n)$, and for all the tests at a split, the complexity is $O(mn \log n)$. At most we have $2^{h_{\max}}$ splits. When $h_{\max}$ is not big, it is a small number and let it be a constant n_s. Therefore the time complexity for tree construction

is $O(mn \log n)$. For tree pruning, the algorithm traverses the tree once and merge the leaves with the same labels under a branching node, which takes a constant time (proportional to n_s).

Overall the time complexity for building a CDT using CDT-PS is $O(mn \log n)$.

For CDT-SPS, calculating propensity scores is time consuming. The time complexity for logistic regression is $O(n^\alpha)$ where $2 < \alpha \leq 3$ depending on different optimisers [34]. The overall time complexity for CDT-SPS is $O(mn \log n + mn^\alpha) = O(mn^\alpha)$.

CDT-SPS does not scale well as CDT-PS does. The detailed comparisons regarding scalability are given in Section 6.4. However, the efficiency of CDT-SPS can be improved based on the research for fast estimation of propensity scores [35].

6 EXPERIMENTS

To evaluate the CDT algorithms, CDT-PS and CDT-SPS, firstly in Section 6.1 we experiment with 2 real world and 1 synthetic data sets to show that CDTs are able to identify more interpretable relationships when comparing to normal decision trees.

The normal decision trees are built using the C4.5 algorithm [3] implemented in Weka [36] with default parameters. It is difficult to evaluate discovered causal relationships as for most real world data sets we do not have the ground truths (true causal relationships). It is also impossible to use a method for evaluating classifiers to assess causal discovery results, because a model containing no causal relationships may give accurate classification. Thus we take a common sense approach to do the evaluation by using two data sets from which the results could make sense to ordinary people. We examine the results to see if they are reasonable, and contrast the CDTs to normal decision trees built based on the data.

In Section 6.2 experiments with synthetic data sets are done to demonstrate the ability of CDT-PS and CDT-SPS in finding causal relationships comparing to the PC algorithm [11], a commonly used Bayesian network learning algorithm.

Another set of experiments with 8 real world data sets are carried out in Section 6.3 to compare the classification accuracy of the CDTs and C4.5 trees.

In Section 6.4 the scalability of CDT-PS and CDT-SPS is evaluated and compared with the C4.5 and PC algorithms.

In all experiments the significance level for Mantel-Haenszel tests and association tests is 0.05, the number of subgroups for CDT-SPS is 5, and the maximum height of a CDT is 5 except in Section 6.3 where the heights of the CDTs are not limited.

6.1 CDTs find meaningful causal relationships

We use three data sets to illustrate that CDTs are able to identify meaningful causal relationships in data. We will focus our discussions on the differences in the results obtained by the CDTs and normal decision trees. We will see that due to the the different criteria used by the two types of trees for choosing branching attributes, the resulting CDTs and normal decision trees are different, and the CDTs represent justifiable causal relationships whereas the normal decision trees may not.

6.1.1 Adult data set - census income

The Adult data set (Table 1) was retrieved from the UCI Machine Learning Repository [37] and it is an extraction of 1994 USA census database. It is a well known classification data set used in predicting whether a person earns over 50K or not in a year.

TABLE 1
Summary of Adult and Ultra Short Stay Unit data sets

The Adult data set			
Attributes	yes	no	comment
age $<$ 30	14515	34327	young
age $>$ 60	3606	45236	old
private	33906	14936	private company employer
self-emp	5557	43285	self employment
gov	6549	42293	government employer
education-num$>$12	12110	36732	Bachelor or higher
education-num$<$9	6408	42434	education years
Prof	23874	24968	professional occupation
white	41762	7080	race
male	32650	16192	
hours $>$ 50	5435	43407	weekly working hours
hours $<$ 30	6151	42691	weekly working hours
US	43832	5010	nationality
$>$50K	11687	37155	annual income, outcome

The Ultra Short Stay Unit (USSU) data set			
Attributes	yes	no	comment
Triage Category 1	43	4269	most urgent
...	...	...	
Triage Category 5	42	4270	least urgent
Male	2158	2154	
Sunday	634	3678	
...	...	...	
Saturday	607	3705	
Diabetes	251	4061	
Asthma	174	4138	
Cardiovascular	142	4170	
Renal	185	4127	
Recent visit	650	3662	
Summer	2013	3289	
...	...	...	
Spring	1225	3087	
Live in the city	2691	1621	
Age 0-16	301	4011	
Age 17-35	2060	2252	
Age 36-64	2453	2859	
Age 65+	498	3814	
Hours in USSU $>$ 18	924	3388	
Hours in ED $>$ 3	1309	3003	
Admitted	799	3513	outcome

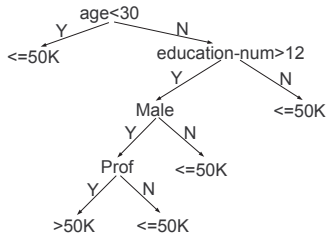


Fig. 3. CDTs of the Adult data set

We recoded the data set to make the causes for high/low income more clearly and easily understandable. The objective is to find the causal factors of high (or low) income.

CDT-PS and CDT-SPS obtain the same CDT (see Figure 3) with the Adult data set, and the normal decision tree built using the data is shown in Figure 4.

From Figure 4, a normal decision tree may be large for high classification accuracy, but a large tree has low interpretability. Although it is possible to reduce the size of a classification tree, its accuracy is sacrificed. The objectives of causal discovery and classification are not consistent. We should note that classification accuracy is not an objective of CDTs. Instead a CDT is built for better interpretation. Hence smaller CDTs are preferred.

The next observation is crucial to show the difference between

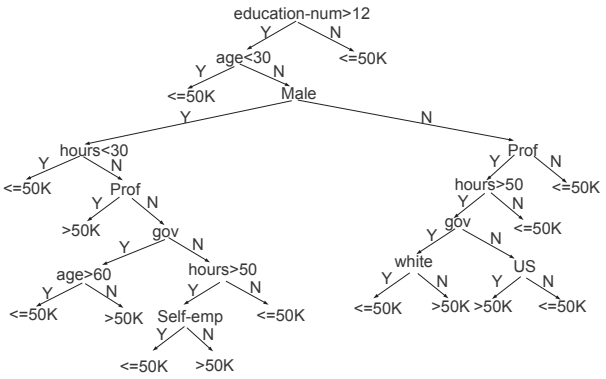
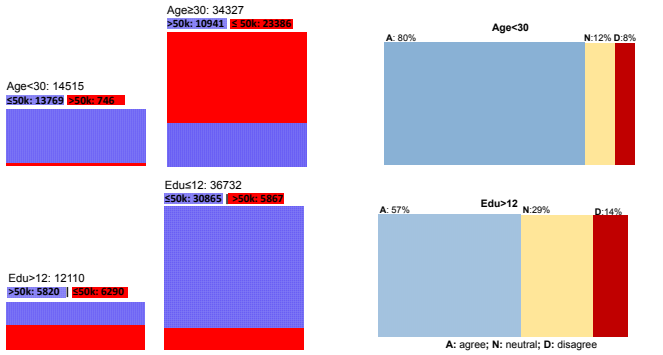


Fig. 4. A C4.5 decision tree of Adult data set



(a) Sample distribution (b) Strata distribution

Fig. 5. An illustration of the different choices of a splitting attribute between a normal decision tree and a causal decision tree

CDTs and normal decision trees. Causality based and classification based criteria do not make the same choice. The root (the first branching attribute) of the normal decision tree is 'education-num>12' and the root of the CDT is 'age<30'. In the following, we provide the justification for the choice by the CDT.

Firstly we look at the choice by the normal decision tree. A summary of the class distributions when the data set is partitioned respectively by 'education-num>12' and 'age<30' is given as:

counts	$> 50K$	$\leq 50K$
education-num>12	5820	6290
education-num≤12	5867	30865
counts	$> 50K$	$\leq 50K$
Age≥30	10941	23386
Age<30	746	13769

The classification error rates with the two attributes are 24.9% and 49.4% respectively, showing that 'education-num> 12' is a better classification attribute than 'age<30', but they do not imply that 'education-num>12' is a stronger causal factor of salary levels.

To make a causal conclusion, a fair comparison is required. We should not compare the salaries of different occupations or compare the salaries of part time workers with full time workers. Following this idea, to justify the choice made by the CDT, we stratify the data based on the values of the relevant attributes except for the candidate cause. We evaluate Claim I: "people with education-num>12 receive higher salary" (and Claim II "people younger than 30 receive lower salary") in these strata. Claim I receives support from 57% of the strata and Claim II receives support from 80% of the strata. So Claim II is more generally true than Claim I when other factors affecting salaries are eliminated. Therefore, 'age $<$ 30' is a stronger causal factor than 'education-num $>$ 12', and it is chosen by the CDT as the first branching attribute. This demonstrates that the criterion used by a CDT captures an important characteristics of causality, persistency [10], [38], which a classification criterion fails to capture.

A visual illustration of the above discussions is shown in Figure 5. To minimise classification errors (indicated by the red areas in Figure 5 (a), i.e. the portion of samples inconsistent with the tree labels), 'education-num $>$ 12' is chosen by C4.5 as the first branching attribute since it incurs significant less errors than 'age $<$ 30' (much smaller red areas for 'education-num $>$ 12'). However, from Figure 5 (b) the choice of 'education-num $>$ 12' leads to a significantly higher percentage of strata in which the causal relationship between 'education-num $>$ 12' and the outcome is disagreed (14% as indicated by the dark red area in the bottom diagram) than the case when 'age $<$ 30' is chosen (8%), therefore 'age $<$ 30' is preferred by CDT in order to achieve a higher percentage of strata agreeing on the causal relationship.

The causal influence of age on income can be seen in our real life too. Young workers normally receive low salaries in nearly all occupations regardless of their education levels, simply because their lack of experience. For older workers, their salaries are dependent on their education, professional occupations and so on as indicated by the CDT.

6.1.2 *The ultra short stay unit (USSU) data set*

The USSU (ultra short stay unit) data was collected from the emergency department of a regional hospital in Australia [39]. The data set records the information of patients who have used the USSUs of the emergency department. The objective here is to understand doctors' decisions for hospital admission following patients' stays in the USSUs. The CDT and normal decision tree built with the data set are quite different. We display and discuss the two trees up to level 3 to illustrate the difference between them.

Referring to Figure 6 (left), the CDT has captured the operational mechanisms of the emergency department. In justifying the root node of the CDT, when a patient stays in a USSU for 18 hours or longer (the maximum hours for a USSU stay are 20), doctors will get a strong indication of the seriousness of the patient's situation and thus the need of hospital admission. So the root node of the CDT reflects the possible logic behind the doctors' decision that longer stay in the USSUs leads to the need of hospital admission. Monday is a busy day since some patients should be discharged during the weekend are discharged on Monday. So, some patients to be admitted to the hospital wait longer than normal at the USSUs. As a result, it appears that Monday is a cause for a higher admission rate to the hospital for those having waited longer at the USSUs. 'Hours in ED>3' (time in the emergency department) also indicates the seriousness of a patient's condition (which will affect doctors' judgment), and is a causal factor of a patient being admitted. The paths of the CDT are associated with doctors reasoning, decisions, and practices, and hence they have causal interpretations.

The normal decision tree (Figure 6, right) picks up cardiovascular disease as its root. Cardiovascular disease seems to be related to hospital admission, but let us explain why it is not a causal factor. Patients with cardiovascular disease are mostly in the mature and senior groups (age 36-64 and 65+) and there are very few (or no) instances in the two other age groups. In other

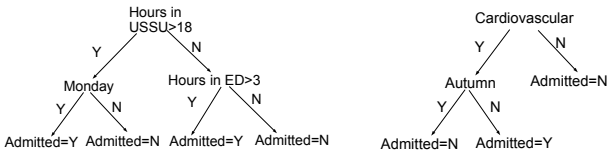


Fig. 6. CDT (left) and C4.5 tree (right) of the USSU data. The labels of a CDT leaf indicate the proportionally larger number of instances instead of the majority of instances. The distribution of the classes is so skewed and the majority of instances is "Admitted=N" since its domination

words, 'Cardiovascular disease $\rightarrow$ Admitted' does not represent a persistent relationship in the data set since it does not receive support from younger age groups. Hospital admission is a complex decision, and cardiovascular disease is a too simplistic indicator and may be misleading. For example, those senior patients with cardiovascular disease are very likely to suffer from other diseases, such as diabetes and renal disease, so their hospitalisation is a result of poor health due to a number of diseases. Given that a patient may have other diseases, cardiovascular disease does not necessarily result in a significant increase of the chance of hospital admission, and this indicates that cardiovascular disease is not a cause. On the other hand, 'Hours in USSU $>$ 18 $\rightarrow$ Admitted' does reflect the mechanism of the complex decisions made by doctors. It does not give a simple predictor for hospital admission as we wished, but it does show the fact that there is a need for doctors to make complex decisions.

Another interesting observation in this and the previous example is that a normal decision tree quickly leads to the tree nodes with small numbers of instances. For example the tree leaves of the normal decision tree in Figure 6 (right) have 23, 62, and 3428 instances respectively. In contrast, the leaves of the CDT in Figure 6 (left) have 95, 592, 780, and 2046 instances respectively. The leaves with small number of instances may lead to small classification errors but do not result in strong relationships.

6.1.3 A random data set

A CDT and a normal decision can be totally different. To demonstrate this point, we build a CDT and a normal decision tree with a randomised data set where there is no relationship at all. Values in each of 10 attributes are randomly drawn with 50% 1s and 50% 0s in the data set. When we try to learn a CDT from the data, no tree is returned and this is expected. However, C4.5 grows a decision tree as in Figure 7.

This result shows that the relationships in a normal decision tree may not be meaningful at all and a more interpretable decision tree, like a CDT, is necessary.

6.2 CDT identifies causal relationships

6.2.1 Finding global causal relationships

To show that CDTs are competent in discovering causal relationships, we use 5 groups of synthetic data sets, each group containing 10 data sets with the same number of attributes, to compare the findings of CDT-PS, CDT-SPS and the PC algorithm [11] from the data. In total 50 data sets are used, and each data set contains 10k samples. The data sets are generated using the TETRAD tool (http://www.phil.cmu.edu/tetrad/). To create a data set, in TETRAD we firstly generate randomly a causal Bayesian network structure with the specified number of attributes (20, 40, 60, 80, or 100), and randomly select a node with a specified degree (i.e. number of parent and children nodes, which is in the range of 3 to 7) as the outcome attribute for the data set. The conditional probability tables of the causal Bayesian network are also randomly assigned. The data set is then generated using the built-in Bayes Instantiated Model (Bayes IM) based on the conditional probability tables. The ground truth of the data is the set of nodes directly connected to the outcome attribute in the causal Bayesian network structure.

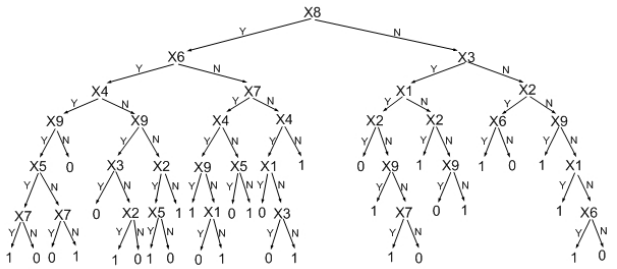


Fig. 7. A C4.5 decision tree of a randomly generated data set

We then apply CDT-PS, CDT-SPS and PC to each of the 50 data sets, and for each group of the data sets, the average recalls of the algorithms are shown in Table 2 (Part A). We do not use pruning for CDTs in this set of experiments. The data is generated by dependency and dependency may not produce distinct leafs as classification. To be consistent with the nature of the data and the criteria used by the methods for the comparison, pruning of CDTs is used and we set the maximum height of a CDT to 5.

It can be seen that in general both CDT-PS and CDT-SPS can detect similar percentages of causal relationships as PC does, indicating that both CDT algorithms have comparable ability and have obtained consistent results in discovering causal relationships as the commonly used approach. We are aware that the causal relationships identified by the CDT algorithms are context specific while those discovered by PC is global or context free. However, it is reasonable to assume that if a causal relationship exists with no context, it should appear in the context too, and these relationships have been mostly picked up by the CDTs.

TABLE 2
Average recalls of CDT and PC (95% confidence interval)

Part A: Average recall of global causal relationships					
Group	#D	#V	CDT-PS	CDT-SPS	PC
1	10	20	90.24%±0.07	87.37%±0.07	74.67%±0.17
2	10	40	89.17%±0.09	89.16%±0.09	78.83%±0.15
3	10	60	89.29%±0.09	89.28%±0.09	77.62%±0.12
4	10	80	83.62%±0.09	81.95%±0.11	90.05%±0.07
5	10	100	100.00%±0.00	97.14%±0.06	94.00%±0.08
Part B: Average recall of context specific causal relationships					
Group	#D	#V	CDT-PS	CDT-SPS	PC
6	10	20	80.83%±0.1	80.83%±0.1	n/a
7	10	40	74.67%±0.15	76.67%±0.16	n/a
8	10	60	72.23%±0.13	71.11%±0.14	n/a
9	10	80	66.03%±0.07	67.42%±0.07	n/a
10	10	100	56.15%±0.11	59.26%±0.11	n/a

#D: number of data sets in a group
#V: number of attributes in one data set

6.2.2 Finding context specific causal relationships

In order to test the performance of CDT-PS and CDT-SPS in finding context specific causal relationships, we also use 5 groups of synthetic data sets, each group containing 10 data sets with the same number of attributes (20, 40, 60, 80 or 100).

To create a data set, e.g. with 20 binary attributes, $\{v_1, v_2, \ldots, v_{20}\}$, we firstly create a causal Bayesian network structure with only one edge, e.g. from v_1 to v_{20} (all other nodes are isolated). With this structure, we use logistic regression to simulate the data set for the Bayesian network. One of the two causally related variables, e.g. v_{20} is chosen as the outcome, then v_1 in this example is the ground truth of the global cause of v_{20}. However, we do not know any context specific causal relationships around v_{20}. Our solution is to use v_1 as the context variable, and apply PC-select [15] (also known as PC-simple [14]) to the two partitions of the data set respectively, one partition containing all the samples with ($v_1 = 0$) and one containing all the samples with ($v_1 = 1$) (while the v_1 column is excluded). In this way, we identify the variables that are causally related to v_{20} within each of the two contexts, ($v_1 = 0$) and ($v_1 = 1$), and use the findings as the ground truth of the context specific causal relationships around v_{20}. PC-select is a simplified version of the PC algorithm for finding causal relationships around a given outcome variable.

We then apply CDT-PS and CDT-SPS to each of the 50 data sets. The CDT built from such a data set always has the node causally related to the output selected as its root, i.e. the CDT correctly finds the global causal relationship. Moreover, each of the CDTs also contains context specific causal relationships. We do not prune CDT trees in these data sets since some randomly generated data sets have skewed distribution, which makes the pruning too aggressive. We will design a pruning strategy for skewed data sets in future work.

Table 2 (Part B) summarises the average recalls of CDT-PS and CDT-SPS in finding the context specific causal relationships. From the table, both CDT-PS and CDT-SPS are able to discover the majority of the context specific causal relationships. PC, in contrast, does not find any context specific causal relationships in the data sets since it is not design for the purpose. If we want to use PC to find the context specific causal relationships, we have to run PC in each context specific data set, which is impractical. On the other hand, the CDT algorithms proposed in this paper can find context specific causal relationships in the complete data sets.

6.3 CDTs for classification

CDTs are designed for discovering and representing causal relationships, so they are not optimised for classification. However, as causal relationships imply the underlying mechanisms of the outcome variable taking different values due to the changes of the cause variables, it is expected that CDTs can be classifiers with good interpretability.

To validate the expectation, apart from the Adult and USSU data sets, we apply CDT-PS to another 8 commonly used UCI data sets (see Table 3), and compare the classification accuracy of the obtained CDTs with the normal decision trees built using the C4.5 implementation in Weka. Note that the Hypothyroid and Sick data sets are retrieved from the Thyroid Disease folder of the UCI Machine Learning Repository, discretised with the discretisation utility of MLC++ [40]. The Car Evaluation data set originally has 4 classes: acc, good, vgood, and unacc. In our experiment, samples of the acc, good and vgood classes are merged into one class. For these 8 UCI data sets, the attributes are nominal and they are converted to binary ones before being applied to CDT-PS. Since one nominal attribute is converted to multiple binary attributes, for the 8 data sets we increase the the maximum number of the relevant attributes (i.e. stratifying attributes) from 10 (the default value) to 15 when building the CDTs (see Section 7.2 for the discussions about limiting the number of stratifying attributes in practice). Moreover, we do not set height limit to the CDTs to make them comparable to C4.5 trees. Table 3 summarises the results of the comparison, where the accuracy is the average classification accuracy of the CDTs or C4.5 trees over the 10 runs of cross validation for a data set.

TABLE 3
A comparison of classification accuracy of CDTs and C4.5 trees

	Accuracy		Tree size	
Data set	C4.5	CDT	C4.5	CDT
Adult	80.80%	80.64%	29	9
USSU	81.17%	81.05%	51	39
BCW (orginal)	94.71%	91.7%	31	35
Car Evaluation	92.36%	93.98%	182	59
Congressional Voting	95.17%	94.71%	16	3
German Credit	72.10%	70.50%	96	17
Hypothyroid	99.21%	96.05%	14	21
K-R vs. K-P	99.44%	97.72%	59	57
Mushroom	100.00%	89.56%	28	9
Sick	98.00%	94.25%	27	15
Average	**91.30%**	**89.02%**	**53**	**26**

As expected, from Table 3, we see that overall the CDTs have achieved similar classification accuracy as the C4.5 trees, with an average accuracy of 89.02%, closely following the average accuracy of C4.5 trees (91.30%). At the same time, most of the CDTs are significantly smaller than the corresponding C4.5 trees, and on average the CDTs are half-sized of the decision trees.

Furthermore, given the causal semantic of CDTs, they have the potential to provide insight into the causal mechanisms of the occurrence and changes of the outcome, thus making more useful predictions or explanations and benefiting understanding and decision making.

6.4 Scalability of the CDT algorithms

We test the scalability of CDT-PS and CDT-SPS by comparing them with the C4.5 implemented in Weka [36] and the PC algorithm [11]. We use 12 synthetic data sets generated with the same procedure as described in Section 6.2.1. To be fair across the data sets, we choose the nodes with the same degree as the outcome attributes. The experiments are done using the desktop computer with a Quad core 3.4 GHz CPU and 16 GB of memory.

The comparison results are shown in Figure 8. The run time of CDT-PS is almost linear to the size of the data sets and the number of attributes. It is less efficient than C4.5 but more efficient than PC. CDT-SPS has good performance in terms of the number of attributes, while it does not scale well with the number of samples. A main reason for this observation is that in CDT-SPS, the logistic regressions are invoked to estimate propensity scores and their time complexity is polynomial to the size of a data set. The results have shown that CDT-PS is practical for both high dimensional and large data sets, while CDT-SPS is suitable for high dimensional but small or medium data sets.

7 Discussions

7.1 Difference from other causal trees

In this section, we differentiate CDTs from the conditional probability table tree (CPT-tree) [20] and causal explanation tree [21], the causal trees derived from causal Bayesian networks.

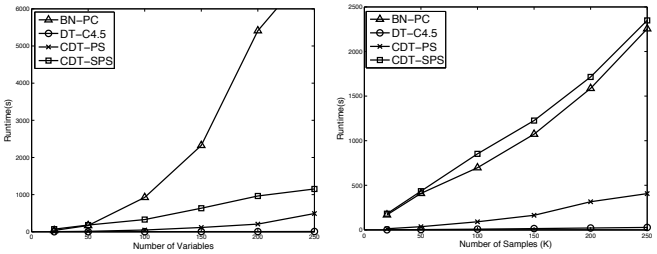


Fig. 8. The scalability of CDT in comparison to C4.5 and PC

A causal Bayesian network (CBN) [4] consists of a causal structure of a directed acyclic graph (DAG), with nodes and arcs representing random variables and causal relationships between the variables respectively, and a joint probability distribution of the variables. Given the DAG of a CBN, the joint probability distribution can be represented by a set of conditional probabilities attached to the corresponding nodes (given their parents). A CBN provides a graphical visualisation of causal relationships, a reasoning machinery for deriving new knowledge (effects) when evidence (changes of causes) is fed into the network; as well as a method for learning causal relationships in data. In recent decades, CBNs have emerged, especially in the area of machine learning, as a core methodology for causal discovery and inference in data.

A CBN depicts the relationships of all attributes under consideration, and it can be complex when the number of attributes is more than just a few. For example, it takes some effort to understand the CBN in Figure 9 learnt from the Adult data set, even though there are only 14 attributes in the data set. A CBN does not give a simple model to explain the causes of an outcome as our CDT does.

The conditional probability table tree (CPT-tree) [20] is designed to summarise the conditional probability tables of a CBN for concise presentation and fast inference. An example of CPT-trees is shown in Figure 10. The probabilistic dependence relationships among the outcome Y and its parent nodes X_1, X_2 and X_3 (causes of Y) are specified by a conditional probability table where the probabilities of Y given all value assignments of its parents are listed. The size of a conditional probability table is exponential to the number of parent nodes of Y and can be very large. For example, for 20 parent nodes, the conditional probability table will have 1,048,576 rows. This table will be difficult to display and the inference based on the table is inefficient too. Given a context, i.e. one or more parent nodes taking an assignment of a value, the probability of Y may be constant (without being affected by the values of other parents). So a conditional probability table can be represented clearly with a tree structure, called a conditional probability table tree (CPT-tree), as illustrated in Figure 10. In the CPT-tree, the causal semantics is naturally linked to the CBN where all parent nodes are direct causes of Y.

There are two major differences between a CPT-tree and a CDT. Firstly, CPT-trees are built from CBNs and CDTs are built from data sets directly. Before building the CPT-trees, we already know the causal relationships, and a CPT tree specifies how the assignments of some cause variables link to outcome values. This is impractical in many real world applications since we do not know the CBN or we could not build a CBN from a data set, particularly a large data set, as existing algorithms for learning CBNs cannot handle a large number of variables and they often only present a partially oriented CBN. Secondly, in a CBN, the parents of a node Y are all global causes of Y. As a CPT-tree

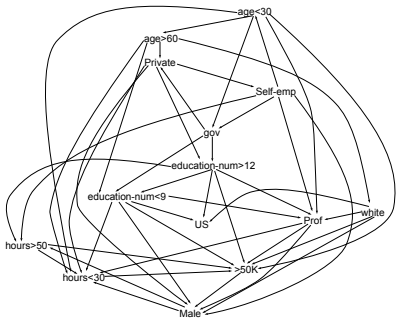


Fig. 9. A causal Bayesian network of the Adult data set

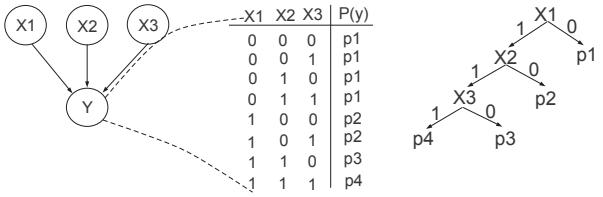


Fig. 10. An illustration of CPT-tree. (L): A Bayesian network; (M): Conditional probability table of Y; (R): CPT-tree.

is derived from a CBN, all the variables included in a CPT-tree are all global causes. However, it is possible that under a context, a variable becomes causally related to Y. So such causal relationships will not be discovered or represented by a CBN and thus not by the CPT-trees too, but they can be be revealed and represented by our CDTs.

A causal explanation tree [21] aims at explaining the outcome values using a series of value assignments of a subset of attributes in a CBN. A series of value assignments of attributes form a path of a causal explanation tree, and a path is determined by a causal information flow. The assignment of a set of attributes along a path represents an intervention in the causal inference in a CBN. The causal interpretation is based on the causal information flow criterion used for building a causal explanation tree. However this method is impractical since we do not have a CBN in most real world applications as explained previously. Similarly a causal explanation tree cannot capture the context specific causal relationships encoded in a CDT, because the explanation tree is obtained from a CBN, which only encodes global causes.

7.2 Practical considerations

The causal interpretation of a CDT is ensured by the evaluation in the stratified data set of the difference in the potential outcomes of a possible causal attribute X_i. In each stratum, the individuals are indistinguishable, or the effect of the attributes possibly affecting the estimation of the causal effect of X_i on Y is eliminated. Therefore, the causal effects estimated using the stratified data sets approach the true causal effects.

An assumption here is that the differences of individuals should be captured by the set of covariates used for stratification. This assumption implies causal sufficiency that all causes are measured and included in the data set. A naïve choice is to select all attributes other than the attribute being tested (X_i) and the outcome (Y), for stratification. However, this is not workable for high dimensional data sets since for perfect stratification many strata will contain very few or no samples when the number of

attributes is large and the estimation of propensity scores will become problematic if the number of samples is small [31]. As a result, the CDT algorithms may not find any causal relationship. For example, diverse information, such as demographic information, education, hobbies and liked movies, is collected as personal profile in a data set. However, if all the attributes are used for stratification, they reduce the chance of finding sizable or reliable strata for the causal discovery. In fact, it is unwise to use any irrelevant attributes, such as hobbies and liked movies, for stratification when the objective is to study, e.g. the causal effect of a treatment on a disease.

A reasonable and practical choice of stratifying attributes is the set of attributes that may affect the outcome, called relevant attributes in this paper. Differences in irrelevant attributes that do not affect the outcome should not impact the estimation of the causal effect of the studied attribute on the outcome. Therefore, only those relevant attributes should be used to stratify a data set, and this is what we have done in the CDT algorithms. In case there are many relevant variables, which may result in many small strata for perfect stratification or inaccurate estimation of propensity scores, we restrict the maximum number of relevant attributes to ten according the strength of correlations. The purpose of this work is to design a fast algorithm to find causal signals in data automatically without user interactions. We do tolerate certain false positives and expect that a real causal relationship will be refined by a dedicated follow-up observational study.

We limit the maximum number of relevant attributes for practical considerations. In many real world studies, the stratification may have to be based on a limited number of demographic attributes, e.g. gender, age group and residential areas. Thinking about a heath study, it is very difficult to recruit volunteers with the same background (age, diet, education, etc.), and stratification on more than a few attributes is just impractical. Nonetheless considering stratification with even only a small number of attributes is better than not using stratification.

CDTs help practitioners with the discovery of causal relationships in the following ways although it may not confirm causal relationships: (1) Because of stratification, many spurious relationships that are definitely not causal will be excluded from the resulting CDTs, so practitioners will have a smaller set of quality hypotheses for further studying; (2) Context specific causal relationships are more difficult to be observed than global causal relationships. CDTs are useful for practitioners to find hidden context specific causal hypotheses.

8 Conclusion

In this paper, we have proposed causal decision trees (CDTs), a novel model for representing and discovering causal relationships in data.

A CDT provides a compact and precise graphical representation of the causal relationships between a set of predictor attributes and an outcome attribute. The context specific causal relationships represented by a CDT are of great practical use and they are not encoded by existing causal models.

The algorithms developed for constructing a CDT utilises the divide and conquer strategy for building a normal decision tree and thus is fast and scalable to large data sets. The criterion used for selecting branching attributes of a CDT is based on the well established potential outcome model and partial association tests, ensuring the causal semantics of the tree.

Given the increasing availability of observational data, we believe that the proposed CDT method will be a promising tool for automated discovery of causal relationships in data, thus to support better decision making and action planning in various areas.

Acknowledgments

This work has been supported by Australian Research Council (ARC) Discovery Project Grant DP140103617.

References

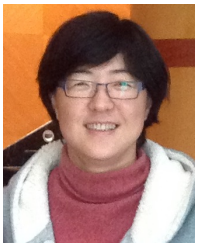
[1] N. Cartwright, "What are randomised controlled trials good for?" *Philosophical Studies*, vol. 147, no. 1, pp. 59–70, 2009.
[2] P. R. Rosenbaum, *Design of Observational Studies*, ser. Springer Series in Statistics. Springer, 2010.
[3] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson, 2006.
[4] P. Spirtes, "Introduction to causal inference," *Journal of Machine Learning Research*, vol. 11, pp. 1643–1662, 2010.
[5] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge University Press, 2009.
[6] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decision," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.
[7] G. W. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, April 2015.
[8] S. Greenland and B. Brumback, "An overview of relations among causal modelling methods," *International Journal of Epidemiology*, vol. 31, pp. 1030–1037, 2002.
[9] N. Mantel and W. Haenszel, "Statistical aspects of the analysis of data from retrospective studies of disease," *Journal of the National Cancer Institute*, vol. 22, no. 4, pp. 719–748, 1959.
[10] J. Li, L. Liu, and T. D. Le, *Practical Approaches to Causal Relationship Exploration*. Springer, 2015.
[11] P. Spirtes, C. C. Glymour, and R. Scheines, *Causation, Predication, and Search*, 2nd ed. The MIT Press, 2000.
[12] R. E. Neapolitan, *Learning Bayesian Networks*. Prentice Hall, 2003.
[13] D. Chickering, D. Heckerman, and C. Meek, "Large-sample learning of bayesian networks is np-hard," *Journal of Machine Learning Research*, vol. 5, pp. 1287–1330, 2004.
[14] M. K. P. Büehlmann and M. Maathuis, "Variable selection for high-dimensional linear models: partially faithful distributions and the PC-simple algorithm," *Biometrika*, vol. 97, pp. 261–278, 2010.
[15] D. Colombo, A. Hauser, M. Kalisch, and M. Maechler, "Package '*pcalg*'," 2014. [Online]. Available: http://cran.r-project.org/web/packages/pcalg/pcalg.pdf
[16] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 171–234, 2010.
[17] Z.Jin, J. Li, L. Liu, T. D. Le, B. Sun, and R. Wang, "Discovery of causal rules using partial association," in *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, Dec 2012, pp. 309–318.
[18] J. Li, T. Le, L. Liu, J. Liu, Z. Jin, and B. Sun, "Mining causal association rules," in *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*, Dec 2013, pp. 114–123.
[19] L. Frey, D. Fisher, I. Tsamardinos, C. Aliferis, and A. Statnikov, "Identifying markov blankets with decision tree induction," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, Nov 2003, pp. 59–66.
[20] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller, "Context-specific independence in bayesian networks," in *The 12th Conference on Uncertainty in Artificial Intelligence (UAI)*, 1996, pp. 115–123.
[21] U. H. Nielsen, J. philippe Pellet, and A. Elisseeff, "Explanation Trees for Causal Bayesian Networks," in *Uncertainty in Artificial Intelligence*, 2008, pp. 427–434.
[22] X. Su, C.-L. Tsai, H. Wang, , D. Nickerson, and B. Li, "Subgroup analysis via recursive partitioning," *Joural of Machine Learning Research*, vol. 10, pp. 141–158, Jun. 2009.
[23] J. Foster, J. Taylor, and S. Ruberg, "Subgroup Identification from Randomized Clinical Trial Data," *Statistics in medicine*, vol. 30, no. 24, pp. 2867–2880, 2011.

[24] M. Dudik, J. Langford, and L. Li, "Doubly robust policy evaluation and learning," in *Proceedings of ICML'11, the 28th International Conference on Machine Learning*, 2011.
[25] S. Athey and G. Imbens, "Machine Learning Methods for Estimating Heterogeneous Causal Effects," *ArXiv e-prints*, Apr. 2015.
[26] S. L. Morgan and D. J. Harding, "Matching estimators of causal effects: Prospects and pitfalls in theory and practice," *Sociological Methods & Research*, vol. 35, pp. 3–60, 2006.
[27] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical Methods for Rates and Proportions*, 3rd ed. Wiley, 2003.
[28] M. W. Birch, "The detection of partial association, I: The 2×2 Case," *Journal of the Royal Statistical Society*, vol. 26, no. 2, pp. 313–324, 1964.
[29] R. P. Rosenbaum and B. D. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.
[30] D. B. D. B. Rubin, "Estimating causal effects from large data sets using propensity scores," *Annals of Internal Medicine*, vol. 127, no. 8, pp. 757–763, 1997.
[31] E. A. Stuart, "Matching methods for causal inference: a review and a look forward," *Statistical Science*, vol. 25, no. 1, pp. 1–21, 2010.
[32] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate Behavioral Research*, vol. 46, no. 3, pp. 399–424, 2011.
[33] R. P. Rosenbaum and B. D. Rubin, "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 516–524, 1984.
[34] P. Komarek, "Logistic regression for data mining and high-dimensional classification," Ph.D. dissertation, School of Computer Science, Carnegie Mellon University, 2004.
[35] B. K. Lee, J. Lessler, and E. A. Stuart, "Improving propensity score weighting using machine learning," *Statistics in Medicine*, vol. 29, no. 3, pp. 337–46, 2010.
[36] Q. Hall and et. al, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.
[37] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml
[38] A. B. Hill, "The environment and disease: Association or causation?" *Proceedings of the Royal Society of Medicine*, vol. 58, pp. 295–300, 1965.
[39] J. Li, A. W.-C. Fu, and P. Fahey, "Efficient discovery of risk patterns in medical data," *Artificial Intelligence in Medicine*, vol. 45, pp. 77–89, 2009.
[40] R. Kohavi, D. Sommerfield, and J. Dougherty, "Data mining using MLC++: A machine learning library in C++," in *Tools with Artificial Intelligence*. IEEE Computer Society Press, 1996, pp. 234–245.

Jiuyong Li received his PhD degree in computer science from Griffith University in Australia. He is currently a professor at the University of South Australia. His main research interests are in data mining, privacy preservation and bioinformatics. His research work has been supported by 5 ARC Discovery pProjects, and he has published more than 100 research papers.

Saisai Ma received his bachelor (2010) and master (2013) degrees in Electronic Engineering from Hohai University, China. He is currently a PhD student at the University of South Australia. His research interests are causal discovery and its application to bioinformatics.

Thuc Le received his BSc (2002) and MSc (2006) degrees in pure Mathematics from the University of Pedagogy, Ho Chi Minh City, Vietnam. He received his BSc (2010) degree in Computer Science and PhD degree in Bioinformatics (2014) at the University of South Australia. He is currently a research fellow at the University of South Australia. His research interests are bioinformatics, data mining and machine learning.



Lin Liu received her bachelor and master degrees in Electronic Engineering from Xidian University, China, in 1991 and 1994 respectively, and her PhD degree in computer systems engineering from University of South Australia in 2006. She is currently a senior lecturer at the University of South Australia. Her research interests include data mining and bioinformatics, as well as protocol verification and network security analysis with Petri nets.

Jixue Liu received his PhD in computer science from the University of South Australia in 2001. He is currently a senior lecturer at the University of South Australia. His research interests include integrity constraint discovery, privacy in data publication, trust management on the internet, transformation of data, constraints, and queries among XML and relational sources, and view maintenance. Jixue Liu has published in the journals of TODS, JCSS, TKDE, Acta Informatica, etc. and in many database conferences.