A general framework for causal classification

Jiuyong Li*, Weijia Zhang*, Lin Liu*, Kui Yu+, Thuc Le* and Jixue Liu*

* School of Information Technology and Mathematical Sciences,
University of South Australia, Australia
+ School of Computer and Information,
Hefei University of Science and Technology, China

March 27, 2020

Abstract

In many applications, there is a need to predict the effect of an intervention on different individuals from data. For example, which customers are persuadable by a product promotion? which groups would benefit from a new policy? These are typical causal classification questions involving the effect or the *change* in outcomes made by an intervention. The questions cannot be answered with traditional classification methods as they only deal with static outcomes. In marketing research these questions are often answered with uplift modelling, using experimental data. Some machine learning methods have been proposed for heterogeneous causal effect estimation using either experimental or observational data. In principle these methods can be used for causal classification, but a limited number of methods, mainly tree based, on causal heterogeneity modelling, are inadequate for various real world applications. In this paper, we propose a general framework for causal classification, as a generalisation of both uplift modelling and causal heterogeneity modelling. When developing the framework, we have identified the conditions where causal classification in both observational and experimental data can be resolved by a naïve solution using off-the-shelf classification methods, which supports flexible implementations for various applications. Experiments have shown that our framework with off-the-shelf classification methods is as competitive as the tailor-designed uplift modelling and heterogeneous causal effect modelling methods.

1 Introduction

The objective of causal classification is to predict whether a treatment would change an individual's outcome [10]. In marketing applications, when the treatment is a promotional advertisement of a product, causal classification is to identify customers likely to purchase the product after having been shown the advertisement. In medical applications, causal classification is to predict if a treatment would improve a patient's outcome.

To differentiate causal classification from normal classification, we need to understand the difference between observed and potential outcomes. Following the potential outcome model [34, 18], for a treatment T, each individual has two potential outcomes, denoted as Y^1 and Y^0, for the outcomes of the person being treated $T=1$ and untreated $T=0$ respectively. At a time point, only one potential outcome can be observed for an individual. For example, if we observe a person buying the product after viewing the advertisement, then $Y = 1 \mid T = 1$ (Y denotes the observed outcome), the potential outcome when $T = 1$ is the same as the observed outcome, i.e. $Y^1 = 1$, but the other potential outcome, Y^0 when $T = 0$, indicating her purchase status without viewing the advertisement, is not observed.

Causal classification aims to predict the positive changes in potential outcomes (i.e. causal effect of a treatment) for an individual, whereas normal classification predicts whether an individual has the desired (observed) outcome or not.

Table 1 lists the four types of responses to a treatment (T is set to 1). A positive response means that an individual is positively influenced by the treatment, e.g. a person buys the product as a result of viewing the advertisement. A negative response means that an individual is negatively influenced by the treatment, e.g. a person having planned to buy the product does not buy it since she dislikes the advertisement. Nonresponse 0 and nonresponse 1 indicate that the treatment has no impact on an individual, e.g. after having viewed the advertisement, a person having no intention to buy the product still does not buy it (nonresponse 0) and a person having the plan to buy the product buys it (nonresponse 1).

It is difficult to differentiate the four types of responses based on observed outcomes. For example, in the observed buying group, we do not know if one has a positive response or a nonresponse 1 as the observed outcomes are $Y = 1$ in both cases, not helping with classifying the responses. This reflects the famous quote by John Wanamaker, the pioneer in marketing: “Half the money I spend on advertising is wasted; the trouble is I don't know which half.”

Table 1: Types of responses to a treatment.

Types of responses for an individual	Potential outcome if $T = 0$	Potential outcome if $T = 1$	Causal class label	Normal class label
Positive response	$Y^0 = 0$	$Y^1 = 1$	1	1
Nonresponse 1	$Y^0 = 1$	$Y^1 = 1$	0	1
Negative response	$Y^0 = 1$	$Y^1 = 0$	0	0
Nonresponse 0	$Y^0 = 0$	$Y^1 = 0$	0	0

Using potential outcomes, causal classification can distinguish the responses as indicated in Table 1. In the marketing example, a positive response is when a person would not buy the product if she did not see the advertisement ($Y^0 = 0$), and she has bought the product after viewing the advertisement ($Y^1 = 1$). Nonresponse 1 is when the person would still buy the product even if she did not see the advertisement ($Y^0 = 1$), and she has purchased the product by simply using the advertisement as a gateway ($Y^1 = 1$). In causal classification, only positive responses are labelled as 1, but in traditional classification, both positive responses and nonresponses 1 are labelled as 1.

It is challenging to make causal classification based on observational data as it involves counterfactual reasoning. When we observe a purchase by a customer after viewing an advertisement ($Y^1 = 1$), we need to infer her unobserved potential outcome Y^0, i.e. to answer the counterfactual question: "Would the customer purchase the product had she not viewed the advertisement?", to determine whether or not the purchase is a result of viewing the advertisement.

When data is collected from a randomised experiment, an uplift modelling method [23, 32, 14, 37] is commonly used in marketing research to model the causal effect of the treatment as the difference between the probabilities of the observed outcomes in the two groups, $P(Y \mid T = 1, \mathbf{X} = \mathbf{x}) - P(Y \mid T = 0, \mathbf{X} = \mathbf{x})$, where $\mathbf{X}$ is the set of features potentially influencing Y.

There is a lack of practical algorithms for causal classification in diverse applications. Causal classification can be achieved by causal heterogeneity discovery. Several machine learning methods have been developed recently to discover causal effect heterogeneity [4, 45, 21], i.e. to identify the subgroups across which the causal effects of a treatment are different and learn the models for predicting the heterogeneous causal effects across the subgroups. We can use such a model to predict the causal effect of a treatment on an individual's outcome with observational data, but there are a limited number of heterogeneity modelling methods, which are mostly tree based. They are not adequate for various applications for modelling complex causal relationships. Causal classification is a concept used by Fermandez and Provost [10], and authors reported a comparative theoretical analysis between normal classification and causal classification. However, in their analysis, it was assumed that there was no nonresponse type 1. The practical implication of the results is to be tested.

This paper will tackle the causal classification problem with observational data (and experimental data) by utilising causal graphical modelling theory and off-the-shelf machine learning methods, with the following contributions:

1. We have developed a theorem for causal classification under the causal sufficiency assumption to enable the use of naïve uplift modelling for causal classification with both experimental and observational data. The theorem has also established explicitly the condition when the existing uplift modelling methods can be used in observational data.

2. We have proposed an algorithmic framework for causal classification with observational data, by linking together normal classification, uplift modelling and causal graphical modelling seamlessly. To the best of our knowledge, this is the first work making such a theoretical link. The link opens the door for all off-the-shelf machine learning methods to be used for causal classification.

2 Classification, uplift modelling, and causal classification

We assume a data set with a treatment T, an outcome Y, a set of all other variables $\mathbf{X}$. For easy notation, we use y for $Y = 1$.

Before proceeding with the discussions, we introduce randomised experiments and differentiate experimental data from observational data. In a randomised experiment, samples are randomly assigned to the treated ($T = 1$) and control ($T = 0$) groups respectively. Due to the randomness, the difference in the outcomes between the two groups is an unbiased estimation of the causal effect of T on Y. In an observational data set, $T = 1$ and $T = 0$ are passively observed, not randomly assigned, and hence the

Table 2: Objective functions for classification, naïve uplift modelling and causal classification

Normal classification: maximise likelihood: $P(y \mid T = 1, \mathbf{X} = \mathbf{x})$
Naïve uplift model: maximise difference: $P(y \mid T = 1, \mathbf{X} = \mathbf{x}) - P(y \mid T = 0, \mathbf{X} = \mathbf{x})$
Causal classification: $P(y \mid do(T = 1), \mathbf{X} = \mathbf{x}) - P(y \mid do(T = 0), \mathbf{X} = \mathbf{x})$

difference in the outcomes between the observed treated and control groups may not indicate the average causal effect of T on Y.

A formal concept behind causal inference in randomised experiments is the unconfoundedness assumption [34, 33], which states that the assignment of treatment T is independent of the potential outcomes (Y^1, Y^0) given some covariates $\mathbf{C} \subseteq \mathbf{X}$. In a well designed experiment, the covariate set is chosen by experts and treatment assignment is random, thus the unconfoundedness assumption is satisfied. In an observational study treatment assignment may not be random, so the unconfoundedness assumption is unlikely to be satisfied. This becomes a main challenge for causal inference in observational data.

In the following, we differentiate the objective functions of classification, uplift modelling, and causal classification, as shown in Table 2.

Normal classification is to predict outcome y by maximising the likelihood $P(y \mid T = 1, \mathbf{X} = \mathbf{x})$. It makes static probabilistic predictions and does not model Y's change with the change of T.

The naïve uplift model referred to in this paper aims to maximise the difference in $\mathrm{P}(y|T = 1)$ and $\mathrm{P}(y|T = 1)$ for a given value of $\mathbf{X}$. The difference is modelled explicitly. The objective function of the naïve uplift modelling has the same form as the objective function of the true uplift modelling in literature [23, 32, 14], but the true uplift modelling assumes the use of experimental data while the naïve uplift modelling in this paper does not.

Causal classification is to predict the change of Y when an individual takes a treatment $T = 1$, and makes use of the conditional causal effect of T on Y, i.e. the degree of the change of Y as a result of changing or intervening on T under condition $\mathbf{X} = \mathbf{x}$. To represent this goal formally, we use Pearl's *do* operator [29], a notation commonly seen 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. $\mathrm{P}(y \mid do(T = 1))$, indicates the probability of $Y = 1$ when T is set to 1, which is different from

$\mathrm{P}(y \mid T=1)$, the probability of $Y=1$ when observing $T=1$. The objective function of causal classification is to estimate the difference, i.e. conditional causal effect of T on Y given $\mathbf{X}=\mathbf{x}$, $\mathrm{P}(y \mid do(T=1), \mathbf{X}=\mathbf{x}) - \mathrm{P}(y \mid do(T=0), \mathbf{X}=\mathbf{x})$.

Based on the objective function of causal classification, we can formally define the causal classification problem with observational data and experimental data as follows.

Definition 1 (Causal classification)**.** *Causal classification is to predict if a treatment T should be applied (i.e. $do(T=1)$) to an individual $\mathbf{X}=\mathbf{x}$ by using the test whether* $\mathrm{P}(y \mid do(T=1), \mathbf{X}=\mathbf{x}) - \mathrm{P}(y \mid do(T=0), \mathbf{X}=\mathbf{x}) \geq \theta$*, where θ is a user specified threshold. A model trained in data to estimate* $\mathrm{P}(y \mid do(T=1), \mathbf{X}=\mathbf{x}) - \mathrm{P}(y \mid do(T=0), \mathbf{X}=\mathbf{x})$ *is a causal classification model.*

The threshold θ is normally determined by the application. For example, in personalised advertising, θ can be determined by the budget of an advertisement campaign and the profit of each successful sale.

Causal classification is very challenging without a known causal graph (structure) and some necessary assumptions. In this paper, we identify the conditions when causal classification can be resolved efficiently using the naïve uplift modelling and existing classification methods. Our specific objectives in this paper are stated as the following.

Definition 2 (Problem statement)**.** *Given an observational data set with a binary treatment variable T, a binary or numerical outcome variable Y and a set of other variables $\mathbf{X}$ of any type, this paper aims to (1) identify conditions when causal classification can be achieved by the naïve uplift modelling, and (2) develop a framework for causal classification using off-the-shelf classification methods.*

Note that although the above problem statement and the following discussions are focused on observational data, the results are applicable to experimental data too. In the discussion, the outcome is binary, but the solutions can be extended to continuous outcome easily.

3 The conditions for causal classification in observational data

3.1 Preliminary

A DAG (directed acyclic graph) $G = (\mathbf{V}, \mathbf{E})$ is a directed graph with a set of nodes $\mathbf{V}$ and a set of directed edges $\mathbf{E}$, and no node has a sequence of directed edges pointing back to itself, i.e. there are no loops. If there exists an edge $P \to Q$ in G, P is a parent node of Q and Q is a child node of P. For a node $V \in \mathbf{V}$, we use $\mathrm{PA}(V)$ to denote the set of all its parents. 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 P is an ancestor of node Q if there is a directed path from P to Q, and equivalently Q is a descendant of P.

Definition 3 (Markov condition [29]). *In a DAG $G = (\mathbf{V}, \mathbf{E})$, $\forall V \in \mathbf{V}$, V is conditionally independent of all of its non-descendants given* $\mathrm{PA}(V)$*.*

Based on the Markov condition, the joint distribution of $\mathbf{V}$ is factorised as $\mathrm{P}(\mathbf{V}) = \prod_{V_i \in \mathbf{V}} \mathrm{P}(V_i | \mathrm{PA}(V_i))$.

Definition 4 (Faithfulness[40]). *If all the conditional independence relationships in $P(\mathbf{V})$ are entailed by the Markov condition applied to DAG $G = (\mathbf{V}, \mathbf{E})$, and vice versa, $P(\mathbf{V})$ and G are faithful to each other.*

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

When we carry out causal inference based on data, the following assumption is essential in addition to the Markov condition and causal faithfulness.

Definition 5 (Causal sufficiency [40]). *For every pair of variables observed in a data set, all their common causes also have observations in the data set.*

Causal sufficiency indicates that all causes of the outcome Y are observed and measured in the data set, which is required by our framework presented in the next sections.

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

d-Separation as defined below is an important concept to read dependencies from a causal DAG.

Definition 6 (d-Separation [29])**.** *A path p in a DAG is d-separated by a set of nodes $\mathbf{Z}$ if and only if*
(1) $\mathbf{Z}$ contains the middle node, V_k of a chain $V_i \rightarrow V_k \rightarrow V_j$ or $V_i \leftarrow V_k \leftarrow V_j$, a fork $V_i \leftarrow V_k \rightarrow V_j$ in p; and
(2) when p contains a collider V_k, none of V_k and its descendants is in $\mathbf{Z}$.

When nodes X and Y are d-separated by $\mathbf{Z}$ in a DAG, we have $(X \perp\!\!\!\perp Y \mid \mathbf{Z})$.

Pearl has invented the *do*-calculus [29] for estimating post-intervention probabilities using a causal DAG and data satisfying causal sufficiency.

Let G be a causal DAG, V_1 and V_2 be two variables in G. Let $G_{\overline{V_1}}$ represent the subgraph of G by removing all incoming edges of V_1, $G_{\underline{V_2}}$ the subgraph of G by removing all outgoing edges of V_2 and $G_{\overline{V_1},\overline{V_2}}$ the subgraph of G by removing all incoming edges of V_1 and V_2. V_1 and V_2 can be variable sets, the edge removals are then for each variable in the sets. The rules of *do*-calculus are presented as follows, where for a variable V, v represents $V = v$.

Theorem 1. *[The three rules of do-calculus [29]]*
Rule 1*: Insertion/Deletion of observation:*
$P(y|do(x), z, w) = P(y|do(x), w)$ *if* $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}}$;
Rule 2*: Action/Observation exchange*
$P(y|do(x), do(z), w) = P(y|do(x), z, w)$ *if* $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\underline{Z}}}$;
Rule 3*: Insertion/Deletion actions*
$P(y|do(x), do(z), w) = P(y|do(x), w)$ *if* $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X},\overline{Z(W)}}}$ *where $Z(W)$ is the set of Z nodes that are not ancestors of any W node in $G_{\overline{X}}$;*

Given a causal DAG and an expression of causal effect using *do* operations, if all *do* operations are reduced to *do* free operations by using the above rules one by one, the causal effect can be estimated in data [29].

Conditional causal effect is what we will use in causal classification, assuming that T is a cause of Y.

Definition 7 (Conditional causal effect)**.** *Let T be a parent node of Y in a causal DAG, and $\mathbf{X}$ be all other variables. The direct causal effect of T on Y is defined as* $(\mathrm{P}(y \mid do(T = 1), \mathbf{X} = \mathbf{x}) - \mathrm{P}(y \mid do(T = 0), \mathbf{X} = \mathbf{x}))$.

Conditional causal effect indicates the change of Y resulted from a change of T under condition $\mathbf{X} = \mathbf{x}$. To estimate direct causal effect, we firstly need to show that the above direct causal effect is identifiable, i.e. the probabilities with the *do* operations can be reduced to their do operation free forms so that they can be estimated using observational data.

In the following, we will develop our solution by identifying the conditions under which the conditional causal effect is identifiable.

3.2 Theorem for causal classification in observational data

We now derive the condition when causal classification can be resolved in observational data, under a realistic problem setting that all variables other than T and Y, denoted as $\mathbf{P}$, are pretreatment variables measured before manipulating T (i.e. they are ancestors of Y and T) and their values are kept unchanged when manipulating T. The variables in $\mathbf{P}$ affect the causal effect of T on Y as context since they do not change when the treatment T is manipulated. We also assume that Y does not have descendants, i.e. the effect variables of Y have not been included in the data set.

This problem setting is realistic as the variables other than T and Y often represent features describing individuals in a study, e.g. gender, age, and education, which are not affected by T or Y. This is often the case in many machine learning problems.

In order to infer causal effects in data, we should assume that there is no sample selection bias, i.e. all members of the target population have an equal chance to be included in the data set.

In our problem setting, the conditional causal effect can be reduced to a simple form as follows, given a causal DAG and Y's parents.

Lemma 1. *Given a data set containing a set of pretreatment variables* $\mathbf{P}$*, the outcome* Y*, and the treatment variable* $T \in \mathrm{PA}(Y)$*, and let* $\mathrm{PA}'(Y) = \mathrm{PA}(Y)\backslash\{T\}$*, conditional causal effect of* T *on* Y *is* $(\mathrm{P}(y \mid do(T=1), \mathrm{PA}'(Y) = \mathbf{p}') - \mathrm{P}(y \mid do(T=0), \mathrm{PA}'(Y) = \mathbf{p}'))$.

Proof. Let $\mathbf{P} = \{\mathrm{PA}'(Y), \mathbf{Z}\}$. $\mathrm{P}(y \mid do(T=1), \mathbf{P} = \mathbf{p}) = \mathrm{P}(y \mid \mathrm{PA}'(Y) = \mathbf{p}', \mathbf{Z} = \mathbf{z})$, In DAG $G_{\overline{T}}$ where the incoming edges of node T have been removed, $\mathbf{Z}$ and Y are d-separated by $\mathrm{PA}'(Y)$, and hence $Y \perp\!\!\!\perp \mathbf{Z} \mid \mathrm{PA}'(Y)$. Therefore, $\mathrm{P}(y \mid \mathrm{PA}'(Y) = \mathbf{p}', \mathbf{Z} = \mathbf{z}) = \mathrm{P}(y \mid \mathrm{PA}'(Y) = \mathbf{p}')$ according to Rule 1 in Theorem 1.

Similarly, $\mathrm{P}(y \mid do(T=0), \mathbf{P}) = \mathrm{P}(y \mid do(T=0), \mathrm{PA}'(Y) = \mathbf{p}')$.

Therefore, the lemma is proved. □

Now we present the following theorem on the condition for estimating the conditional causal effect of T on Y with observational data in our problem setting.

Theorem 2. *Given a data set containing a set of pretreatment variables* $\mathbf{P}$*, the outcome* Y*, and the treatment variable* $T \in \mathrm{PA}(Y)$*, and assume that* *the data set satisfies causal sufficiency. Conditional causal effect of T on Y given $\mathbf{P} = \mathbf{p}$, i.e.* $\mathrm{P}(y \mid do(T = 1), \mathbf{P} = \mathbf{p}) - \mathrm{P}(y \mid do(T = 0), \mathbf{P} = \mathbf{p})$ *is equal to* $\mathrm{P}(y \mid T = 1, \mathrm{PA}'(Y) = \mathbf{p}') - \mathrm{P}(y \mid T = 0, \mathrm{PA}'(Y) = \mathbf{p}')$.

Proof. Firstly, $\mathrm{P}(y \mid do(T = 1), \mathbf{P} = \mathbf{p}) = \mathrm{P}(y \mid do(T = 1), \mathrm{PA}'(Y) = \mathbf{p}')$ according to Lemma 1.

In DAG $G_{\underline{T}}$ where outgoing edges of T have been removed, $\mathrm{PA}'(Y)$ d-separate nodes T and Y, and hence $Y \perp\!\!\!\perp T \mid \mathrm{PA}'(Y)$. Therefore, $\mathrm{P}(y \mid do(T = 1), \mathrm{PA}'(Y) = \mathbf{p}') = \mathrm{P}(y \mid T = 1, \mathrm{PA}'(Y) = \mathbf{p}')$ according to Rule 2 in Theorem 1. Therefore, $\mathrm{P}(y \mid do(T = 1), \mathbf{P} = \mathbf{p}) = \mathrm{P}(y \mid T = 1, \mathrm{PA}'(Y) = \mathbf{p}')$.

Similarly, $\mathrm{P}(y \mid do(T = 0), \mathbf{P} = \mathbf{p}) = \mathrm{P}(y \mid T = 0, \mathrm{PA}'(Y) = \mathbf{p}')$.

Therefore, the theorem is proved.

□

Theorem 2 indicates that given a causal DAG and a data set which are faithful to each other and assuming causal sufficiency, conditional causal effect can be estimated as conditional probability of Y given its parents using observational data.

Now, we can derive a main result of this paper.

Corollary 1. *In the same setting as in Theorem 2, the causal classification problem can be resolved by the naïve uplift modelling on the projected data set containing* $(T, \mathrm{PA}'(Y), Y)$.

Proof. Based on Theorem 2, $\mathrm{P}(y \mid do(T = 1), \mathbf{P} = \mathbf{p}) - \mathrm{P}(y \mid do(T = 0), \mathbf{P} = \mathbf{p}) = \mathrm{P}(y \mid T = 1, \mathrm{PA}'(Y) = \mathbf{p}') - \mathrm{P}(y \mid T = 0, \mathrm{PA}'(Y) = \mathbf{p}')$. Referring to Table 2, in this case the objective functions of causal classification and the naïve lift modelling are the same. Therefore, causal classification can be resolved by the naïve uplift modelling on the projected data set containing $(T, \mathrm{PA}'(Y), Y)$. □

Corollary 1 links causal classification with normal classification. When we have a true causal DAG representing causal relationships, we can conduct causal classification in observational data following Corollary 1. When the true DAG or PA(Y) is unknown, we can learn it from observational data when assuming causal sufficiency.

In some applications, domain experts have the knowledge about the direct causes of Y, the following corollary supports causal classification without learning a DAG (parents of Y) from data.

Corollary 2. *Let* $\{T\} \cup \mathbf{P}'$ *be the set of all direct causes of* Y*, causal classification can be achieved by the naïve uplift modelling on data set* $(T, \mathbf{P}', Y)$.

Proof. When $\{T\} \cup \mathbf{P}'$ are direct causes of Y, they must be parents of Y in the causal DAG. So, $\text{PA}'(Y) = \mathbf{P}'$. According to Corollary 1, causal classification can be achieved by the naïve uplift modelling on data set $(T, \mathbf{P}', Y)$. $\square$

4 Framework and Algorithm

There are two key components in our causal classification framework, finding PA(Y) and building classification models on the projected data set. We can obtain PA(Y) from a given causal DAG or domain knowledge, or learn it from data. A causal classification models is then built on the projected data set $(T, \text{PA}'(Y), Y)$. Based on the results from the previous section, the objective functions of naïve uplift modelling and causal classification on the projected data set are consistent when $\mathbf{P}$ contains pretreatment variables.

In this paper, we present a framework where users can assemble their own causal classification system using off-the-shelf machine learning methods.

4.1 Finding parents of Y in data

When we do not know the causes of Y, finding PA(Y) from data is a major step for causal classification. One straightforward way is to learn an entire causal DAG from data and then to read PA(Y) from the DAG, However, learning an entire DAG is computationally expensive or intractable with high dimensional data. Furthermore, it is often unnecessary and wasteful to find the entire DAG when we are only interested in the local structure around Y.

Local structure discovery [2] fits our purpose better. Currently there are mainly two types of local structure discovery methods, one for identifying PC(Y), the set of Parents (direct causes) and Children (direct effects) of the target Y; and one for discovering $MB(Y)$, the Markov Blanket of Y, i.e. the parents, children and spouses of Y. Discovering PC(Y) is sufficient for our work as in our problem setting, Y does not have descendants, i.e. $\text{PC}(Y) = \text{PA}(Y)$. Several algorithms have been developed for discovering PC(Y), such as MMPC (Max-Min Parents and Children) [43] and HITION-PC [1]. These algorithms use the framework of constraint-based Bayesian network learning and employ conditional independence tests for discovering PC(Y).

4.2 Two Model approach

Our framework builds a causal classification model and conducts classification using the following Two Model approach.

Definition 8 (Two Model approach)**.** *Given a data set D and assume causal sufficiency and $T \in \mathrm{PA}(Y)$. Let $M_{T=1}$ and $M_{T=0}$ be two classifiers built with $D_{\Pi(\mathrm{PA}'(Y))\wedge(T=1)}$ and $D_{\Pi(\mathrm{PA}'(Y))\wedge(T=0)}$ respectively, where $D_{\Pi(\mathrm{PA}'(Y))\wedge((T=1)}$ Or $D_{\Pi(\mathrm{PA}'(Y))\wedge((T=0)}$ is a projected data set from D to $\mathrm{PA}'(Y)$ and selected by $T = 1$ or $T = 0$. The test for causal classification in Definition 1 can be achieved by $\mathrm{P}(y \mid M_{T=1}(\mathbf{P}' = \mathbf{p}')) - \mathrm{P}(y \mid M_{T=0}(\mathbf{P}' = \mathbf{p}')) > \theta$ where $\mathbf{P}' = \mathrm{PA}'(Y)$.*

Based on the proposed framework, we present the Causal Classification by the Two Model approach (CCTM) algorithm in Algorithm 1. The training phase of CCTM is to build two classifiers using features in $\mathrm{PA}'(Y)$ in the two sub datasets containing $T = 1$ and $T = 0$ respectively. Any classification method, such as decision tree or SVM can be plugged in to build the classifiers. In the prediction phase, the trained classifier pairs $M_{T=1}$ and $M_{T=0}$ predict whether a treatment will lead to a positive response (effect) or not. Line 1 of the prediction phase projects the test data set to contain the same features in $\mathrm{PA}'(Y)$ only in order to use the two classifiers to estimate $\mathrm{P}(y \mid T = 1, \mathbf{P}' = \mathbf{p}')$ and $\mathrm{P}(y \mid T = 0, \mathbf{P}' = \mathbf{p}')$ respectively for an individual. If the difference in the probabilities (estimated conditional causal effect) is larger than θ, the individual is predicted to have a positive response and should be treated. Otherwise, the treatment should not be applied to the individual.

5 Experiments

5.1 Parent discovery in data

We first evaluate the performance of the local structure learning algorithms MMPC and HITON-PC. Their implementations are from the Causal Explorer package [41], and G^2 test (with significance level 0.01) is used for conditional independence test. For the conditional independence tests, the maximum size of a conditioning variable set is 3 for both algorithms. The experiments are done on a PC with Intel(R) i5-8400 and 16GB memory.

Four benchmark Bayesian networks (BNs), CHILD [6], ALARM [5], PIGS [20], and GENE [39] (`www.bnlearn.com/bnrepository`), are used to

ALGORITHM 1: Causal Classification by the Two Model approach (CCTM)

```
/*—Training—*/
Input: Data set $D$ containing treatment variable $T$, pretreatment variables $\mathbf{P}$ and outcome variable $Y$.
Output: Two models $(M_{T=1}, M_{T=0})$ .
call a local PC algorithm to find $\mathrm{PA}(Y)$
let $\mathrm{PA}'(Y) = \mathrm{PA}(Y) \backslash T$
project data set $D$:$(T, \mathbf{P}, Y)$ to $D'$:$(T, \mathrm{PA}'(Y), Y)$
split data set $D'$ to $D_1 \mid T = 1$ and $D_0 \mid T = 0$
call a classification method to build a classifier $M_{T=1}$ on $D_1$
call a classification method to build a classifier $M_{T=0}$ on $D_0$
output $(M_{T=1}, M_{T=0})$

/*—Prediction—*/
Input: Model pair $(M_{T=1}, M_{T=0})$, $\mathrm{PA}'(Y)$, test data set $D_T$ without treatment assignment and outcome, and a user specified threshold $\theta$
Output: $D_T$:$(\hat{T}, \mathbf{P}, CE)$ where $\hat{T}$ contains treatment assignment and $CE$ contains estimated conditional causal effects.
project data set $D_T$:$(\mathbf{P})$ to $D'_T$:$(\mathrm{PA}'(Y))$
for each $r \in D'_T$ do
    let $\mathrm{P}(y \mid T = 1, r) = M_{T=1}(r)$
    let $\mathrm{P}(y \mid T = 0, r) = M_{T=0}(r)$
    if $((\delta = \mathrm{P}(y \mid T = 1, r) - \mathrm{P}(y \mid T = 0, r)) > \theta$ then
        let $t = 1$
    else
        let $t = 0$
    end if
    add record $(\hat{T} = t, r, CE = \delta)$
end for
output $D_T$:$(\hat{T}, \mathbf{P}, CE)$
```

generate the evaluation data sets:.They contain 20, 37, 441 and 801 variables respectively. For each BN, we generate data sets with 500, 1000, and 5000 samples respectively. For each sample size, we generate a group of 10 datasets, so in total 120 datasets are generated for 4 BNs. We make use of nodes having no descendants to be consistent with our problem setting, i.e. T and Y have no descendants and all other variables are pretreatment

Table 3: Quality of parent discovery

BN	Size	Alg	Precision	Recall	F1 score
CHILD	500	MMPC	97.0±0.2	90.0±0.1	92.0±0.1
		HITON	97.0±0.2	90.0±0.1	92.0±0.2
	1000	MMPC	100±0.0	100±0.0	1.00±0.0
		HITON	100±0.0	100±0.0	100±0.0
	5000	MMPC	**100±0.0**	**100±0.0**	**100±0.0**
		HITON	**100±0.0**	**100±0.0**	**100±0.0**
ALARM	500	MMPC	60.0±0.3	90.0±0.3	71.0±0.2
		HITON	60.0±0.3	90.0±0.2	71.0±0.2
	1000	MMPC	92.0±0.0	100±0.1	95.0±0.1
		HITON	92.0±0.0	100±0.1	95.0±0.0
	5000	MMPC	**100±0.0**	**100±0.0**	**1.00±0.0**
		HITON	**100±0.0**	**100±0.0**	**100±0.0**
PIGS	500	MMPC	91.0±0.0	100±0.1	95.0±0.0
		HITON	92.0±0.0	100±0.1	95.0±0.0
	1000	MMPC	100±0.0	100±0.0	100±0.0
		HITON	100±0.0	100±0.0	100±0.0
	5000	MMPC	**100±0.0**	**100±0.0**	**100±0.0**
		HITON	**100±0.0**	**100±0.0**	**100±0.0**
GENE	500	MMPC	76.0±0.1	95.0+0.2	82.0+0.1
		HITON	76.0±0.1	95.0±0.2	83.0±0.1
	1000	MMPC	72.0±0.0	100±0.2	82.0±0.1
		HITON	83.0±0.0	100±0.2	89.0±0.1
	5000	MMPC	**100±0.0**	**100±0.0**	**100±0.0**
		HITON	**100±0.0**	**100±0.0**	**100±0.0**

variables.

To demonstrate the quality of parent discovery, the average precision, recall, and F1 score are reported in Table 3. In most cases, the algorithms produce accurate results.

We then generate data sets with 5K, 15K, 25K, 35K and 50K samples respectively to evaluate the scalability of the algorithms. As shown in Figure 1, both MMPC and HITON-PC are scalable to the size of data sets.

In the proposed CCTM algorithm, MMPC is used.

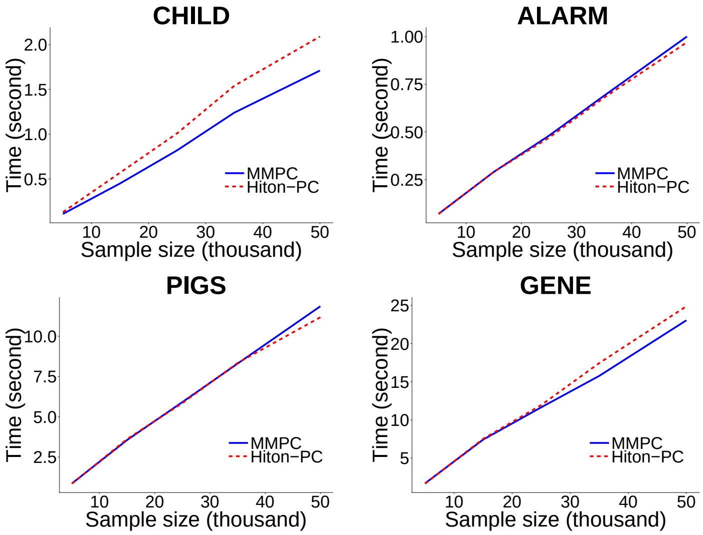


Figure 1: Scalability of MMPC and HITON-PC

5.2 Causal effect estimation

We compare CCTM algorithms with the state-of-the-art uplift modelling methods and causal heterogeneity modelling methods, including Uplift Random Forests (Uplift RF) [13], Uplift Causal Conditional Inference Forests (Uplift CCIF) [12], t-Statistics Tree [42], CausalTree [4], and the X-Learner [21]. Methods' implementations are from authors' or commonly used packages: Uplift RF and Uplift CCIF from `https://cran.r-project.org/web/packages/uplift/index.html`, t-Stats Tree and Causal Tree from `https://github.com/susanathey/causalTree`, and X-Learner `https://github.com/soerenkuenzel/causalToolbox`. Default parameters are used.

We use two popular classifiers, SVM and Random Forest (RF) to instantiate our proposed causal classification framework to two algorithms, denoted as CCTM-SVM and CCTM-RF respectively. The implementations of RF and SVM are from `https://cran.r-project.org/web/packages/randomForest/index.html` and `https://www.csie.ntu.edu.tw/~cjlin/libsvm/respectively`. Default parameters are used.

Two groups of simulation data sets (Group 1 and Group 2) are generated following work [15]. The generation program is at `https://cran.r-project.org/web/packages/CovSelHigh/index.html`. The causal DAGs are shown Figure 2. A group contains 10 datasets, each with 10,000 samples and 102 variables. T and Y are binary. Apart from X_1 to X_{10} in the

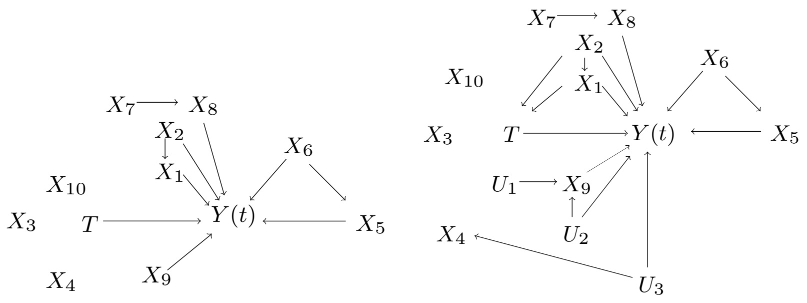


Figure 2: Causal DAGs for synthetic data generation. Left: for Group 1 data sets; Right: for Group 2 data sets.

DAGs, other 90 variables which are irrelevant to T and Y are included to simulate real world situations. !00 variables are drawn from a mixture of continuous and binary distributions. When generating a dataset in Group 2, after obtaining the dataset based on the right structure in Figure 2, we remove the columns for variables U_1,U_2 and U_3 from the dataset to simulate latent variables. A half of data set is used for training models and another half is used to test the accuracy. Threshold θ is set to 0. A prediction is correct if the treatment assignment is the same as the assignment based on the ground truth causal effect for data generation.

For each group of datasets, we report the average performance of the algorithms in Table 4. Using the variables in $\mathrm{PA}'(Y)$ consistently achieves higher causal effect estimation accuracy than using all the variables for all methods, except X-Learner RF in Group 2 data, where the results of using all variables and using $\mathrm{PA}'(Y)$ are very close. This means that our theoretical findings, i.e. Theorem 2, improves other uplift modelling methods and causal heterogeneity modelling methods. Additionally, the proposed CCTM algorithms achieve competitive accuracy in comparison with other methods. The accuracies of the both CCTM algorithms are very close to the highest accuracies.

5.3 Evaluation with real-world data sets

Two real world data sets, the Hillstrom's email campaign data set [17] (Hillstrom for shot) and the Twins birth data set [3] are used to further evaluate CCTM. We compare CCTM-RF and CCTM-SVM (for which $\mathrm{PA}'(Y)$ are used) with all the other methods.

Hillstrom contains 42613 customer records from an email marketing campaign collected for an uplift modelling challenge [17]. Half of these cus-

Table 4: Accuracy of causal effect estimation of T on Y. The highest accuracy in each group is marked in bold. Parent nodes improve the accuracies. CCTM methods perform competitively with other methods.

Method	Strategy	Group 1	Group 2
Causal Tree	All	80.0±0.5	73.4±1.3
	PA′(Y)	**81.2±0.5**	**79.3±0.9**
t-Stats Tree	All	34.5±2.1	11.6±1.3
	PA′(Y)	**74.6±0.8**	**81.7±5.1**
Uplift CCIF	All	77.4±1.0	89.0±1.3
	PA′(Y)	**78.8±1.3**	**89.2±0.6**
Uplift RF	All	77.3±1.4	89.2±1.0
	PA′(Y)	**78.9±1.6**	**89.3±1.0**
Two Model RF	All	71.2±1.2	78.1±1.3
CCTM RF	PA′(Y)	**84.1±0.8**	**88.3±0.6**
Two Model SVM	All	84.4±0.9	87.9±1.2
CCTM SVM	PA′(Y)	**84.8±1.5**	**88.6±1.1**
X-Learner RF	All	84.4±0.8	90.4±0.8
X-Learner RF	PA′(Y)	**85.0±0.8**	**90.6±0.7**

tomers were randomly chosen to receive an advertisement email targeting male users, and the other half of the customers served as a control group. There are 7 pretreatment variables describing customers. The outcome is whether a customer visits the website. MMPC finds three parent variables for the outcome.

Twins birth is a benchmark data set for causal inference [46]. It consists of 4821 samples of twin births (with birth weight $<$ 2kg and having no missing values) in the USA between 1989 and 1991 [3]. Each record contains 40 pretreatment variables. Treatment $T=1$ indicates the heavier one in the twins and $T=0$ indicates the lighter one. The outcome is the mortality of a child after one year. MMPC finds 4 parent variables for the outcome.

Since there are not ground truth conditional causal effects, we use the Qini curve [30], a widely used metric for uplift modelling to compare the algorithms. Qini curve shows the cumulative number of the uplift as a function of the number of individuals treated. The larger the area a curve covers, the better the corresponding method is.

From the Qini curves in Figure 3, the proposed algorithms, CCTM-SVM and CCTM-RF achieve competitive performance with other compared methods, ranked the second and third with the Hillstrom data set, and the

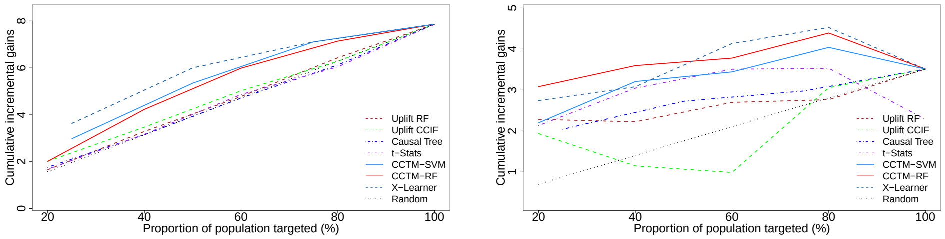


Figure 3: The Qini curves of different methods on the two real world data sets (Left: Hillstrom, Right: Twins). Blue and red solid lines: the proposed methods; Dashed lines: other methods; Black dotted line: a random classifier.

third and first with the Twins data set. Both algorithms are in the high performing groups respectively with the two data sets.

5.4 Discussions

We will discuss the strengths and weaknesses of CCTM in this section.

The strengths of CCTM are its simplicity, and ease of implementation with off-the-shelf classification methods. It is principally correct when assuming causal sufficiency. The proposed CCTM is more advanced than the previous Two Model methods in that it has chosen adjustment variables for causal effect estimation.

We will discuss the weaknesses of CCTM based on the criticisms on the general Two Model approach [31], as summarised in the following: **1)** The Two Model approach builds two independent models, and it does not explicitly fit the differences in two populations; **2)** it is unable to detect weak uplifting signals; **3)** the feature used for outcome prediction (classification or regression) might be different from the features used for uplifting prediction, and hence a Two Model method may fail to use key predictors for uplift; and **4)** the independent fitting errors in two models may be significantly larger than the error in one model.

We now address the above criticisms with respect to CCTM point-by-point. Some weaknesses are not unique to CCTM algorithms and some have been resolved by the design of CCTM.

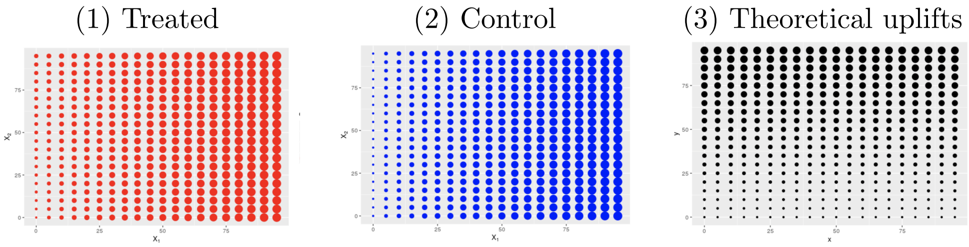
Criticism 1). As shown in Table 2, the objective of causal classification is not to fit the difference in two populations but to unbiasedly estimate causal effect in data. To obtain an unbiased causal effect estimate, the right set of confounders need to be adjusted [18, 29]. Theorem 2 and Corollary 2 in the paper are sound for finding the right set of confounders with causal sufficiency assumption. Given the right set of confounding variables, the Two Model approach is a principled approach for causal effect estimation [4, 21]. Some researchers ague that two potential outcomes follow different mechanisms [25] and hence it makes sense to use two distinct models for two potential outcomes. It is arguable that one model approach can estimate causal effect better than Two Model approach since it involves only one fitting error instead of two, and we will discuss this when we analysing Criticism 4) below.

Criticism 2). We now show using the simulation data set provided by the authors in [31] that in fact Two Model methods and One Model methods do not have observable differences in their performance in detecting weak uplifting signals. The data set was designed to show the incapability of a Two Model method for detecting weak uplifting signals. The data set contains 64,000 data instances, one treatment variable T, two covariates (X_1, X_2) and one outcome variables Y are shown in Figure 4. Theoretical uplifts are also shown in the figure.

We have repeated the experiments on this data set using all previous methods except Uplift CCIF as the available implementation of the algorithm works on binary outcome only. The results are shown in Figure 5. There is not a clear winner on this data set. Causal tree, Up-lift RF, and CCTM SVM perform similarly. They capture the trend of uplift increasing with X_2, but they have not modelled the gradient of changes in X_2 very well. CCTM RF and T-Stats Tree seemingly capture the trend of change with X_1, but poorly. X-Learner consistently underestimates the causal effects.

Criticism 3). This paper resolves the problem based on the graphical causal modelling framework. Features used for the outcome prediction (classification or regression) are different from variables for causal effect estimation. A set of adjustment variable set is used for causal effect estimation and causal classification. We have also shown that parent discovery as a "feature selection" method improves both Two Model methods and other methods in causal effect/uplift estimation.

Criticism 4). This criticism is based on the assumption that a model has a constant fitting error. One counter argument is that there are a large pool of methods to choose from for modelling various relationships for a Two Model method since it can use any supervised machine learning method, and hence there is a better chance to find the most suitable modelling method for an application. Therefore, the fitting error of a Two Model method can be small. This flexibility overcomes the weakness of having two fitting errors. Furthermore, when the two potential outcomes follow different mechanisms,


(1) Treated

(2) Control

(3) Theoretical uplifts

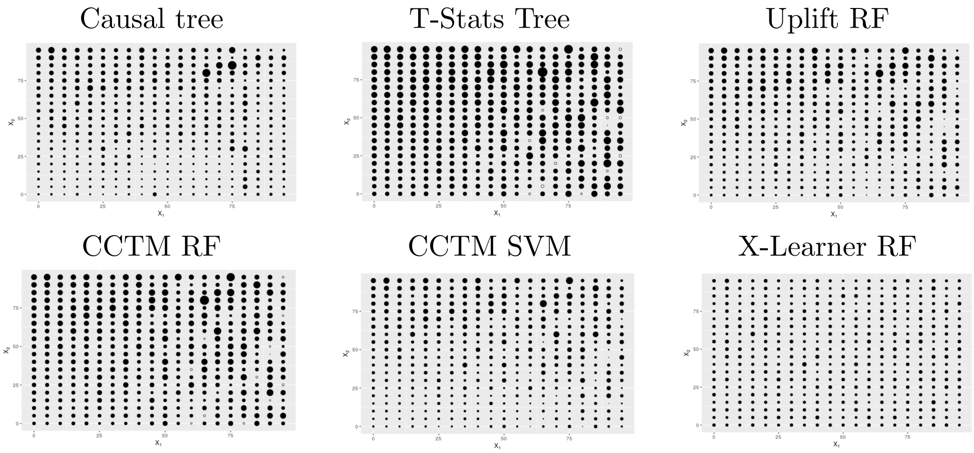
Figure 4: An illustration of the data set used in [31] (1) A visualisation of data instances in the treated group. Values of covariates X_1 and X_2 are grouped by bins of size 5. The area of a disc represents the average outcome in a group. (2) A visualisation of data instances in the control group. (3) Theoretical uplifts in the data set. Note that, the areas of discs in this subfigure is in a different scale from the previous two, and is roughly 6 times larger than that of the previous two since the causal effects are weak. The uplift increases with X_2, and slightly increases with X_1 given the same X_2 value. The disc at the top-right corner is the largest one.

using two distinct models is necessary. In practice, a Two Model method is not evidently more inaccurate than a One Model method. In our experiment, CCTM performs as well as other methods. Another evaluation on a simulation data set in [14] shows that a Two Model method achieves better estimations than other One Model methods compared.

6 Related work

Causal classification is closely related to causal effect estimation and causal effect heterogeneity. The potential outcome model [18] and causal graphical models [29] are two major frameworks for causal effect estimation. Great efforts have been made on the research of average causal effect estimation.

Causal effect heterogeneity is modelled by conditional average causal effects as the causal effects vary in subpopulations. Su et al. [42] used recursive partitioning to construct the interaction tree for causal effect estimation in subgroups. Foster et al. [11] introduced the virtual twins method to define subgroups with enhanced causal effects. In [45], random forest was used to predict the probability of an outcome given a set of covariates and CART was used to find a small set of covariates strongly correlated with the treatment to define the subgroups. Dudik et al. [8] developed an optimal decision making approach via the technique of Doubly Robust estimation. Athey et



Causal tree

T-Stats Tree

Uplift RF

CCTM RF

CCTM SVM

X-Learner RF

Figure 5: Illustrations of estimated uplifts by various methods. The areas of discs in all subfigures are in the same scale as the theoretical uplifts plot in Figure 4. There is not a clear winner among the methods. The data set is challenging for all methods.

al. [4] built the Causal Tree to find the subpopulations with heterogeneous causal effects. An X-Learner method [21] was proposed for causal heterogeneity modelling with unbalanced treated and untreated samples. All the methods assume a data set with a known covariate set. Recently, several algorithms have also been proposed to estimate conditional average causal effects using neural networks [38, 24, 22].

Covariate selection is essential for causal effect estimation. Covariate set renders the treatment and the outcome to satisfy the ignorability [18] or unfoundedness assumption. Unlike in an experiment where covariates are normally selected by domain experts, data driven covariate selection is very channelling since ignorability is impossible to be tested in data. Data driven methods use the backdoor criterion [29] to identify a covariate set, either based on a causal graph created using domain knowledge or learned from data. VanderWeele and Shpitser [44] linked the conditional ignorability with the backdoor criterion. de Luna et al. [7] and Entner et al. [9] have proposed methods to find covariate sets using conditional independence test. Maathuis and Colombo [26] generalised the backdoor criterion for data without causal sufficiency. All these methods are inefficient and could not work on high dimensional data and their effectiveness has not been tested in real world data sets.

Uplift modelling is another line of work for predicting conditional causal effects, mainly in marketing research where data collection is through some experimental designs. Causal effect has not been mentioned in uplift modelling, but fundamentally, uplift modelling is a type of causal inference [14, 10]. The first proposal of uplift modelling is by Radcliffe and Surry [32], Hansotia [16] and Lo [23]. In the well designed experimental data set, Rzepakowski and Jaroszewicz adapted decision trees for uplift modelling [35, 36]. Similar adaptions have extended to Bayesian networks [28] and SVMs [27]. In a similar fashion to the CATE estimation literature, ensemble methods have been introduced to model uplift using a forest of uplift modeling trees [13]. A special case of transformed outcome method has also been introduced to uplift modeling using off-the-shelf estimators directly on the transformed outcomes [19]. Uplift modelling has recently been linked to casual effect heterogeneity modelling [14, 10], but no unified algorithmic framework has been presented.

7 Conclusion

This paper presents a general framework for causal classification, which generalises both uplift and causal heterogeneity models. We have presented a theorem which identifies the conditions for causal classification in observational data and links causal classification and normal classification. The theorem enables a general framework for causal classification using off-the-shelf machine learning methods. We have shown that our theorem improves existing uplift modelling and causal effect heterogeneity modelling methods for better causal effect estimation and our algorithms have competitive performance comparing to other uplift modelling and causal heterogeneity modelling methods in synthetic and real world data sets. We have discussed strengths and weaknesses of the proposed framework in depth and the discussions potentially guide choosing suitable causal classification methods in various applications.

References

[1] Aliferis, C., Tsamardinos, I., Statnikov, A.: Hiton: a novel markov blanket algorithm for optimal variable selection. In: AMIA Annual Symposium Proceedings, vol. 2003, pp. 21–25. American Medical Informatics Association (2003)

[2] Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D.: Local causal and Markov blanket induction for causal discovery and feature selection for classification Part I: Algorithms and empirical evaluation. Journal of Machine Learning Research **11**, 171–234 (2010)

[3] Almond, D., Chay, K.Y., Lee, D.S.: The costs of low birth weight. The Quarterly Journal of Economics **120**(3), 1031–1083 (2005)

[4] Athey, S., Imbens, G.: Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences **113**(27), 7353–7360 (2016)

[5] Beinlich, I.A., Suermondt, H.J., Chavez, R.M., Cooper, G.F.: The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In: The Second European Conference in Artificial Intelligence in Medicine, pp. 247–256. Springer (1989)

[6] Cowell, R.G., Dawid, P., Lauritzen, S.L., Spiegelhalter, D.J.: Probabilistic networks and expert systems: Exact computational methods for Bayesian networks. Springer (2006)

[7] De Luna, X., Waernbaum, I., Richardson, T.S.: Covariate selection for the nonparametric estimation of an average treatment effect. Biometrika **98**(4), 861–875 (2011)

[8] Dudik, M., Langford, J., Li, L.: Doubly robust policy evaluation and learning. In: Proceedings of the 28th International Conference on Machine Learning, pp. 1097–1104 (2011)

[9] Entner, D., Hoyer, P., Spirtes, P.: Data-driven covariate selection for nonparametric estimation of causal effects. In: Artificial Intelligence and Statistics, pp. 256–264 (2013)

[10] Fernandez, C., Provost, F.: Causal classification: Treatment effect vs. outcome estimation (2018). URL `http://www.misrc.umn.edu/workshops/2018/spring/Causal_Targeting_Feb_2018b.pdf`

[11] Foster, J.C., Taylor, J.M.G., Ruberg, S.J.: Subgroup Identification from Randomized Clinical Trial Data. Statistics in medicine **30**(24), 2867–2880 (2011)

[12] Guelman, L., Guillén, M., Marín, A.M.P.: Optimal personalized treatment rules for marketing interventions: A review of methods, a new

proposal, and an insurance case study. UB Riskcenter Working Paper Series, 2014/06 (2014)

[13] Guelman, L., Guillén, M., Pérez-Marín, A.M.: Uplift random forests. Cybernetics and Systems - Intelligent Systems in Business and Economics **46**(3-4), 230–248 (2015)

[14] Gutierrez, P., Gérardy, J.Y.: Causal inference and uplift modelling: A review of the literature. In: Proceedings of the 3rd International Conference on Predictive Applications and APIs, Proceedings of Machine Learning Research, Volume 67, pp. 1–13 (2017)

[15] Häggström, J.: Data driven confounder selection via Markov and Bayesian networks. Biometrics **74**, 389–398 (2018)

[16] Hansotia, B., Rukstales, B.: Incremental value modeling. Journal of Interactive Marketing **16**(3), 35–46 (2002). DOI 10.1002/dir.10035

[17] Hillstrom, K.: The minethatdata e-mail analytics and data mining challenge (2008)

[18] Imbens, G.W., Rubin, D.B.: Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge University Press (2015)

[19] Jaskowski, M., Jaroszewicz, S.: Uplift modeling for clinical trial data. In: Workshop on Clinical Data Analysis (2012)

[20] Jensen, C.S.: Blocking Gibbs sampling for inference in large and complex Bayesian networks with applications in genetics. Ph.D. thesis, Aalborg University (1997)

[21] K´unzel, S.R., Sekhon, J.S., Bickel, P.J., Yu, B.: Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of National Academy of Sciences **116 (10)**, 4156–4165 (2019)

[22] Künzel, S.R., Stadie, B.C., Vemuri, N., Ramakrishnan, V., Sekhon, J.S., Abbeel, P.: Transfer learning for estimating causal effects using neural networks. Tech. rep., arXiv (2018)

[23] Lo, S.Y.V.: The true lift model: A novel data mining approach to response modeling in database marketing. ACM SIGKDD Explorations Newsletter **4**(2), 78–86 (2002)

[24] Louizos, C., Shalit, U., Mooij, J.M., Sontag, D., Zemel, R., Welling, M.: Causal effect inference with deep latent-variable models. In: Advances in Neural Information Processing Systems, pp. 6446–6456 (2017)

[25] Luna, X.D., Waernbaum, I., Richardson, T.S.: Covariate selection for the nonparametric estimation of an average treatment effect. Biometrika **98**(4), 861–875 (2011)

[26] Maathuis, M.H., Colombo, D., et al.: A generalized back-door criterion. The Annals of Statistics **43**(3), 1060–1088 (2015)

[27] Nassif, H., Kuusisto, F., Burnside, E., Page, D., Shavlik, J., Costa, V.: Score as you lift (sayl): A statistical relational learning approach to uplift modeling. In: Joint European conference on machine learning and knowledge discovery in databases, pp. 595–611 (2013)

[28] Nassif, H., Wu, Y., Page, D., Burnside, E.: Logical differential prediction bayes net, improving breast cancer diagnosis for older women. In: American Medical Informatics Association Annual Symposium Proceedings, vol. 2012, pp. 1330–1339 (2012)

[29] Pearl, J.: Causality: Models, Reasoning, and Inference, 2nd edn. Cambridge University Press (2009)

[30] Radcliffe, N.: Using control groups to target on predicted lift: Building and assessing uplift model. Direct Marketing Analytics Journal pp. 14–21 (2007)

[31] Radcliffe, N., Surry, P.: Real-world uplift modelling with significance-based uplift trees. Tech. rep., White Paper TR-2011-1, Stochastic Solutions (2011)

[32] Radcliffe, N.J., Surry, P.D.: Differential response analysis: Modeling true responses by isolating the effect of a single action. Credit Scoring and Credit Control IV (1999)

[33] Rosenbaum, R.P., Rubin, B.D.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)

[34] Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology **66**(5), 688–701 (1974)

[35] Rzepakowski, P., Jaroszewicz, S.: Decision trees for uplift modeling. In: 2010 IEEE International Conference on Data Mining, pp. 441–450 (2010)

[36] Rzepakowski, P., Jaroszewicz, S.: Decision trees for uplift modeling with single and multiple treatments. Knowledge and Information Systems **32**(2), 303–327 (2012)

[37] Rzepakowski, P., Jaroszewicz, S.: Uplift modeling in direct marketing. Journal of Telecommunications and Information Technology pp. 43–50 (2012)

[38] Shalit, U., Johansson, F.D., Sontag, D.: Estimating individual treatment effect: generalization bounds and algorithms. Tech. rep., arXiv (2016)

[39] Spellman, P.T., et al.: Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular biology of the cell **9**(12), 3273–3297 (1998)

[40] Spirtes, P., Glymour, C.C., Scheines, R.: Causation, Predication, and Search, 2nd edn. The MIT Press (2000)

[41] Statnikov, A., Tsamardinos, I., Brown, L.E., Aliferis, C.F.: Causal explorer: A matlab library of algorithms for causal discovery and variable selection for classification. Causation and Prediction Challenge Challenges in Machine Learning, Volume 2 p. 267 (2010)

[42] Su, X., Tsai, C.L., Wang, H., , Nickerson, D., Li, B.: Subgroup analysis via recursive partitioning. Joural of Machine Learning Research **10**, 141–158 (2009)

[43] Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing Bayesian network structure learning algorithm. Machine Learning **65**(1), 31–78 (2006)

[44] VanderWeele, T.J., Shpitser, I.: A new criterion for confounder selection. Biometrics **67(4)**, 1406–1413 (2011)

[45] Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association **113**(523), 1228–1242 (2018)

[46] Yao, L., Li, S., Li, Y., Huai, M., Gao, J., Zhang, A.: Representation learning for treatment effect estimation from observational data. In: Advances in Neural Information Processing Systems 31 (NIPS 2018), pp. 2638–2648 (2018)