Developing data mining methods for finding causal relationships involving multiple factors

Saisai Ma
Master of Engineering in Signal and Information Processing

A thesis submitted for the degree of
PhD in Computer and Information Science



University of
South Australia

School of Information Technology & Mathematical Sciences
Division of Information Technology, Engineering and the Environment
University of South Australia

June 2017

Contents

List of Figures **v**

List of Tables **viii**

1 Introduction **1**

2 Literature review **12**

2.1 Discovering causal relationships involving single factors . . . 13

2.1.1 Learning a global causal network 14

2.1.2 Learning a partial causal structure 17

2.1.3 Learning causal relationships around a fixed target variable 18

2.2 Discovering causal relationships involving multiple factors . . 19

2.2.1 Discovering the combination of causal factors 20

2.2.2 Discovering cause interactions 22

2.2.3 Discovering context specific causes 24

2.3 Conclusion 29

3 Background **30**

3.1 Causal Bayesian networks and HITON-PC 30

3.2 The potential outcome model 35

3.3 Matching methods 39

3.4 Conclusion 41

4 Mining combined causes in large data sets **42**
4.1 Problem statement and definitions 43
4.2 Uncovering combined causes 45
4.2.1 The naïve approach 45
4.2.2 Basic idea of the proposed method 46
4.2.3 Multi-level HITON-PC 47
4.2.4 False discoveries of multi-level HITON-PC 50
4.3 Experiments 50
4.3.1 Data sets 51
4.3.2 Performance evaluation 54
4.3.3 Efficiency and scalability 58
4.4 Conclusion 60

5 Data-driven discovery of cause interactions **62**
5.1 Problem statement and definitions 64
5.1.1 Notation 64
5.1.2 The sufficient-component cause model 65
5.1.3 Problem definition 67
5.2 Cause interactions 69
5.2.1 2-way cause interactions 69
5.2.2 m-way cause interactions 71
5.2.3 Algorithm 74
5.2.3.1 Candidate variable generation 74
5.2.3.2 Stratification 76
5.2.3.3 Cause interaction discovery 78
5.2.3.4 Pruning schema and complexity analysis . . 80
5.3 Experiments 81
5.3.1 Synthetic data 82
5.3.2 Benchmark data 86
5.3.3 Clinical data 89
5.3.4 Genomic data 90
5.3.5 Scalability 93
5.4 Conclusion 95

6 A general framework for context specific causal relationship discovery 97
6.1 Problem statement and definitions . . . 98
6.2 Context specific causal relationship discovery . . . 100
6.2.1 Overview of CSCR . . . 100
6.2.2 Selecting rules with strong causal signals . . . 101
6.2.3 Causal rule tests . . . 102
6.2.4 Instantiations of CSCR . . . 107
6.2.4.1 CSCR with a single decision tree . . . 107
6.2.4.2 CSCR with multiple decision trees . . . 109
6.2.4.3 CSCR with association rule mining . . . 110
6.3 Experiments . . . 111
6.3.1 Synthetic data generation . . . 111
6.3.2 Performance evaluation . . . 115
6.3.3 Scalability . . . 118
6.3.4 Applications to real world data sets . . . 119
6.3.5 Comparison with the existing method . . . 122
6.4 Conclusion . . . 126

7 Causal decision trees 127
7.1 Problem statement and definitions . . . 128
7.1.1 Why a decision tree may not encode causality? . . . 128
7.1.2 Problem statement . . . 132
7.2 Causal decision trees . . . 134
7.2.1 A measure for causal effect . . . 135
7.2.2 CDT by perfect stratification . . . 138
7.2.3 CDT by stratification on propensity score . . . 142
7.2.4 Time complexity . . . 143
7.3 Experiments . . . 144
7.3.1 CDTs find meaningful causal relationships . . . 146
7.3.1.1 Adult data set - census income . . . 146
7.3.1.2 The Ultra Short Stay Unit (USSU) data set 150
7.3.1.3 A random data set . . . 153

7.3.2 CDT identifies causal relationships 154
7.3.2.1 Finding global causal relationships 154
7.3.2.2 Finding context specific causal relationships 156
7.3.3 CDTs for classification 157
7.3.4 Scalability of the CDT algorithms 159
7.4 Conclusion 161

8 Conclusions **162**

References **168**

List of Figures

1.1 Multiple individual causes (the left figure) vs. the combined cause (the right one), where solid arrows denote causal relationships and the diamond represents that the combination of two variables causes an outcome. 5

1.2 Cause interaction between genetic and environmental variables on a disease. (The dashed lines represent the interaction between two variables.) Example: $Gene = Pro12Ala\ allele$, $Environmental\ exposure = dietary\ fat\ intake$ and $Disease = obesity$ [1]. 7

1.3 A simple decision tree. 9

2.1 Y-structure. 18

2.2 An illustration of CPT-tree. (L): A Bayesian network; (M): Conditional probability table of Y ; (R): CPT-tree. 25

2.3 An example of SGM to present the context specific independence. 26

4.1 A random Bayesian network. 44

4.2 A toy example to show the data generation (The solid arrows denote causal relationships and the diamond represents that the combination of two variables causes an outcome.). 53

4.3 Scalability with number of variables - Census data 59

4.4 Scalability with number of variables - Synthetic data 59

4.5 Scalability with umber of records - Census data 60

5.1 Examples of sufficient causes of an immature death. 63

5.2 The sufficient-component cause model for the target Y, which has three sufficient causes. 65
5.3 An illustration of 3-way cause interactions. 72
5.4 Cause interactions found from sufficient causes, consisting of three component causes. 85
5.5 Venn diagram of the number of cause interactions discovered by the four algorithms from the Car Evaluation data set. . . 87
5.6 The cause interactions discovered by at least three proposed algorithms, which are associated with the KEGG pathway, pathways in cancer, and have the experimentally confirmed regulatory relationships. 92
5.7 Scalability with the number of variables. 93
5.8 Scalability with the number of samples. 94

6.1 A causal graphical model 104
6.2 A decision tree of the toy data set. 108
6.3 Diversified multiple trees of the toy data set. 109
6.4 The causal Bayesian network of synthetic data in the context $X_c = 0$ (data part I). 113
6.5 The causal Bayesian network of synthetic data in the context $X_c = 1$ (data part II). 114
6.6 Scalability evaluation with CSCR-S, CSCR-3, CSCR-5 and CSCR-Ass, when # context variables = 1, i.e. $|\boldsymbol{X_c}| = 1$ . . . 118
6.7 Scalability evaluation with CSCR-S, CSCR-3, CSCR-5 and CSCR-Ass, when # context variables = 2, i.e. $|\boldsymbol{X_c}| = 2$ . . . 119
6.8 A CRT of the Car Evaluation data set, where $safety(med)$ is set as the treatment. The rule in the path from the root to the leaf with dashed box has a low causal effect. 123
6.9 A CRT of the Adult data set, where $sex(male)$ is the treatment. The rule in the path from the root to the leaf with dashed box has a low causal effect. ($work$: working class, $self.emp$: self employment.) 125

7.1 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 131
7.2 An example showing that a CDT represents causal relationships (a) An exemplar data set. (b) a CDT of the data . . . 137
7.3 CDTs of the Adult data set 147
7.4 A C4.5 decision tree of Adult data set 148
7.5 An illustration of the different choices of a splitting attribute between a normal decision tree and a causal decision tree . . 149
7.6 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 152
7.7 A C4.5 decision tree of a randomly generated data set 154
7.8 The scalability of CDT with the number of variables in comparison to C4.5 and PC 160
7.9 The scalability of CDT with the number of samples in comparison to C4.5 and PC 160

List of Tables

1.1 A toy example of the Simpson's paradox. 2

4.1 A Brief Description of Data Sets 51
4.2 Comparison of the proposed algorithms with the naïve methods 55
4.3 Comparison of combined causes discovered by CR-CS, CR-PA, MH-PC-F and MH-PC-B with small synthetic data sets 56
4.4 Comparison of combined causes discovered by CR-CS, CR-PA, MH-PC-F and MH-PC-B with larger synthetic data sets 57
4.5 Number of (single and combined) causes discovered by MH-PC-F and MH-PC-B in real world data sets 57
4.6 Examples of combined causes identified from Sick and German data sets 58

5.1 Two 2×2 contingency tables 79
5.2 The difference between the four algorithms. 80
5.3 A summary of synthetic data sets 82
5.4 Cause interactions discovered by PS-OR, PS-Prob, SGPS-OR and SGPS-Prob from sufficient causes, each of which composes of two component causes. 83
5.5 Description of the Car Evaluation data set. 86
5.6 Some cause interactions identified by the proposed algorithms, from the Car Evaluation data set. 88

5.7 Some cause interactions identified by four algorithms, from the METABRIC data set. (*diag_age*: age at diagnosis of the disease, *size*: size of tumor in cm, *grade*: grade of disease, *stage*: composite of size and # positive nodes, *pos_lymph*: # positive lymph nodes, *removed_lymph*: # lymph nodes removed, *ER.Expr*: estrogen receptor expression, *chemo*: whether patient had chemotherapy, *radiation*: whether patient had radiation therapy, *hormone*: whether patient had hormone therapy.) 90
5.8 Statistical significance of validated miRNA-mRNA-miRNA cause interaction triplets. 92

6.1 Causal rules discovered by CSCR-S on Syn-10. 115
6.2 The accuracy of CSCR with synthetic data sets, where P, R and F_1 represent *precision*, *recall* and F_1 *score*. 117
6.3 Performance of CSCR with Syn-100. 117
6.4 Comparison of causal rules discovered by CSCR-S, CSCR-3 and CSCR-Ass from the Car Evaluation data set. 120
6.5 Comparison of causal rules discovered by CSCR-S, CSCR-3 and CSCR-Ass from the Adult data set. 121
6.6 Comparison of causal rules discovered by CRT and CRCS-S from the Car Evaluation data set. 124
6.7 Comparison of causal rules discovered by CRT and CRCS-S from the Adult data set. 125

7.1 Summary of Adult data set 146
7.2 Summary of Ultra Short Stay Unit data set 151
7.3 Average recalls of CDT and PC (95% confidence interval) . . 155
7.4 A comparison of classification accuracy of CDTs and C4.5 trees 158

Abstract

The central goal of the research in many fields is the discovery of causal relationships. Randomized Controlled Trials (RCTs) have been used as the gold standard for causal discovery, but RCTs are often infeasible due to practical and ethical limits. Given the restrictions, observational studies are considered as the alternatives to RCTs and it has been shown that well-designed observational studies can obtain comparative results as RCTs. In many applications, an observational study is adopted to validate a hypothesized causal relationship which is usually generated based on prior knowledge. However, it may be difficult to generate such causal hypotheses with domain knowledge only, especially when there are a large number of variables under study.

Significant progress has been made in inferring causal relationships directly from observational data. Many methods have been proposed both to build causal models, and to improve the efficiency of causal discovery from observational data. A critical limitation of the methods is that they are impractical for detecting causal relationships involving multiple factors. For example, many common diseases arise due to the joint influence of genetic and environmental factors, instead of just one individual factor. Whereas the joint influence of multiple factors is normally hard to identify, even with domain knowledge, a small but increasing attention has been paid to detecting such multiple causal factors from data. The existing work, however, typically assumes that the joint causal relationships or factors are known in advance and their focus is to validate and/or quantify the hypothesized relationships.

In this thesis, we study the problem of discovering the causal relationships involving multiple factors and building predictive model with causal interpretations directly from observational data. First, we propose a multi-level approach based on an efficient local causal discovery method to discover combined causes from data. A combined cause studied in this thesis refers to a combined variable that is a cause of an outcome, but none of its components is a cause by itself. This definition is different from the generally discussed multiple causes of an outcome. Second, we develop a general framework to explore whether the component variables of a combined cause interact with each other to produce the outcome, or they only have an additive effect on the outcome. Then, to uncover context specific causal relationships, a novel method is proposed to identify causes and their contexts simultaneously. Finally, we develop a predictive model which can provide causally interpretable predictions for decision making. We have applied the proposed methods to synthetic data and several commonly used public data sets for evaluation.

The main contribution of this thesis is that we have designed a number of data mining methods to efficiently discover causal relationships involving multiple factors from observational data. These methods have the potential to reveal more complex causal relationships behind the phenomena and thus provide useful techniques for examining the mechanisms of complex systems. The causal prediction model built using the proposed method can assist with evidence based decision making.

Declaration

This thesis presents work carried out by myself and does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; to the best of my knowledge it does not contain any materials previously published or written by another person except where due reference is made in the text; and all substantive contributions by other to the work presented, including jointly authored publications, are clearly acknowledged.

Name: Saisai Ma

Signed: 

Date: 29/06/2017

Acknowledgements

This thesis is the outcome of a PhD candidature at the School of Information Technology and Mathematical Sciences, the University of South Australia (UniSA). This would not have been possible without the guidance of my supervisors, assistance from collaborators, and support from friends and family.

Firstly, I reserve my special gratitude for my family. I thank my parents, Jing Ma and Zuxiang Zhou, who raised me up with immense sacrifice and tremendous love. My special thanks go to my wife, Qiaoling, for her understanding and encouragement during my PhD. I also thank my little boy who brings so much laughs and giggles to our life.

I would like to express my deep gratitude to my principal supervisor, Dr. Lin Liu. Throughout my candidature, I have been fortunate to have Dr. Liu as an exemplary mentor, a serious collaborator, and an enthusiastic coach. I particularly appreciate her meticulous effort and patience in helping me to achieve a scientific writing style to the highest standard.

I am deeply grateful to my co-supervisor, Professor Jiuyong Li, for the invaluable advice and detailed feedbacks he provided to my work during the candidature. His perspective and devotion for research was particularly contagious and motivational for me in the pursuit of my doctoral degree and a future research career. A special thank extends to my co-supervisor, Dr. Thuc Duy

Le, whose brilliant suggestions always make ideas significantly improved.

I thank my all colleagues in the Data Analytics Group of UniSA. In the last few years, we had many activities with good memories. I particularly thank Dr. Kui Yu, as I have benefited from our insightful discussions and collaborations. I thank Dr. Jie Chen, Dr. Wei Kang, Sarowar Sattar, Masud Karim, Weijia Zhang, Michael Bewong, Tao Hoang, Chen Zhan, Xuying Yao and Jeffery Ansah for their continuous encouragements.

My research work at UniSA would not have been possible without the generous financial support of scholarship from the school and the university (UPS). I thank professor Andy Koronios, head of the school of Information Technology and Mathematical Sciences. I would like to thank Dr. Jixue Liu and Dr. Ivan Lee for their support during my PhD journey.

List of publications

1. **Saisai Ma**, Jiuyong Li, Lin Liu, and Thuc Duy Le, Mining combined causes in large data sets, *Knowledge-Based Systems*, vol. 92, pp. 104-111, 2016. (Impact factor: 3.325)

2. Jiuyong Li*, **Saisai Ma***, Thuc Duy Le, Lin Liu, Jixue Liu, Causal decision trees, *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 2, pp. 257-271, 2016. (Impact factor: 2.476) (* contributed equally to this work. Jiuyong Li initiates the first idea and the first draft, and Saisai Ma significantly improves the idea and implementation to the current version.)

3. **Saisai Ma**, Lin Liu, Jiuyong Li, Thuc Duy Le, and Zeng-Hua Lu, Data-driven discovery of cause interactions, *Data Mining and Knowledge Discovery*, under review.

4. **Saisai Ma**, Jiuyong Li, Lin Liu, and Thuc Duy Le, Efficient exploration of context specific causal rules, *IEEE Transactions on Knowledge and Data Engineering*, under review.

5. Jiuyong Li, Thuc Duy Le, Lin Liu, Jixue Liu, Zhou Jin, Bingyu Sun, and **Saisai Ma**, From observational studies to causal rule mining. *ACM Transactions on Intelligent Systems*

and Technology, vol. 7, no. 2, article 14, 2015. (Impact Factor: 2.414)

To my parents Jing Ma and Zuxiang Zhou, my parents-in-law,
Ping Sun and Qiaolin Wang, my wife, Qiaoling Sun and
my little boy, Yao.

"*To know truly is to know by causes.*"

Francis Bacon (1561-1626)

Chapter 1

Introduction

A variety of observational data has been available in many fields, enabling researchers to detect the relationships between variables based on the data. Many data mining and machine learning methods have been developed to build computer models from data to assist predictions and decision making.

Association based approaches have been widely studied to discover interesting relationships hidden in large data sets [2]. For example, dietary habits may be considered as a potential factor of a certain disease, if an association between the habits and the disease is found in the data. The associations found in supermarket transactions, e.g. the association between bread and milk, may be used to improve the sales of supermarkets [3].

However, associations do not necessarily mean causality. For instance, increased ice cream sales may have a strong association with deaths by drowning. Obviously we will never conclude that ice cream consumption causes drowning or vice verse. This is just because that in summer people tend to buy more ice creams, and also go swimming more frequently, and thus more deaths by drowning in summer too.

Table 1.1: A toy example of the Simpson's paradox.

	Recovered	Not recovered	Sum	Recover rate
Drug	20	20	40	50%
No drug	16	24	40	40%
Sum	36	44	80	
Female	Recovered	Not recovered	Sum	Recover rate
Drug	2	8	10	20%
No drug	9	21	30	30%
Sum	11	29	40	
Male	Recovered	Not recovered	Sum	Recover rate
Drug	18	12	30	60%
No drug	7	3	10	70%
Sum	25	15	40	

Actually associations between variables may not be consistent in different subgroups. Table 1.1 shows a toy example of the Simpson's paradox [4]. We see that at the whole population level people taking the drug have a higher recovery rate than those not taking the drug (50% vs. 40%). However, when we examine the recovery rate in each gender group, we have an opposite conclusion: in either the female or male group, the drug leads to a lower recovery rate (20% vs. 30% and 60% vs. 70%). Therefore, associations identified from data may not lead to a convincing conclusion.

In contrast, causal relationships imply the mechanisms of outcome variables taking their values and how the changes of cause variables would lead to the changes of the outcome variables. For example, as the climate (e.g. humidity and temperature) is a cause of bloom date of flowers, a change of the climate has a significant impact on the bloom date. Therefore, with identified causal relationships, we can predict potential effects before actually carrying out any actions, which is extremely useful in preventing erroneous decision making. For example, in medical science, causal relationships help us understand the causes of diseases, and thus deliver better diagnosis and

cures for diseases. Due to the characteristics of causality, causal discovery [4, 5] is a central goal of many scientific studies in various areas.

The major difference between causal relationship discovery methods and association based methods is that the former assesses the relationship between a (potential) cause variable and an outcome variable by removing the effects of its covariates on the outcome but the latter only considers the relationship between the potential cause variable and the outcome in isolation.

The gold standard of causal discovery has typically been Randomized Controlled Trials (RCTs) [5, 6]. RCTs are a type of scientific experiments aiming to assess the effect of a new treatment. Participants are allocated at random (by chance alone) to either the group receiving the treatment or to a group receiving standard treatment (or placebo treatment) as the control. The outcomes of two groups of participants are then compared to obtain an estimate of the treatment effect. Nevertheless, RCTs often may not be feasible, particularly with human participants, due to practical and ethical limits. For example, it is not permitted to instruct patients to take a new drug with serious or unknown side effects. In addition, it may be difficult and costly to apply RCTs to high dimensional problems with large numbers of potential causes.

Given these restrictions, identifying causal relationships via observational studies [7, 8, 9] is considered as the alternative to RCTs, and well-designed observational studies can obtain comparative results as RCTs [10]. As suggested by the name, observational studies are based on passively observed data, and investigators observe subjects and measure variables of interest without assigning treatments to the subjects. Observational studies are the commonly used methods in epidemiology and statistics to investigate causal relationships [11, 12]. However, generally, an observational study is designed to validate a hypothesized causal relationship and the hypothesis needs to

be provided based on the prior knowledge or an anticipation of the relationship between an exposure and the outcome. Thus observational studies may not be a feasible approach to exploring possible causal relationships in data, especially when the number of variables is large.

It has been shown that causal relationships between random variables can be learned directly from observational data [4]. Many data-driven methods have been developed for automated causal discovery from data. Notably Causal Bayesian Networks (CBNs) [4, 13] are considered as a main framework to represent and learn causal dependency among a set of random variables. In a CBN, each node represents a random variable and each arrow is a cause-effect relationship from the parent to the child. To alleviate the efficiency issues of learning a CBN from high dimensional data, some constraint based [13] and local causal discovery methods [14] have been developed to reduce the computational cost. However, the efficiency of these methods is still an open issue, and thus practical applications of these methods are restricted.

Another main limitation of these data-driven causal discovery methods is that they are mainly designed to discover causal relationship between a single factor and the outcome. However, a single causal factor is often insufficient for reasoning about the cause of a particular effect [15]. For example, many diseases are resulted from the joint effect of genetic defects and environmental exposures [1]. In practice, the mechanisms behind many phenomena are usually complex, with multiple factors, rather than just a single factor involved [15].

A typical type of causal relationships involving multiple factors is the case that the combination of multiple factors is a cause of an outcome, while each of the factors has weak (or even no) causal signals on the outcome. We define such a combination of multiple factors as a ***combined cause*** in this thesis. It is different from the generally discussed multiple causes of an outcome, as

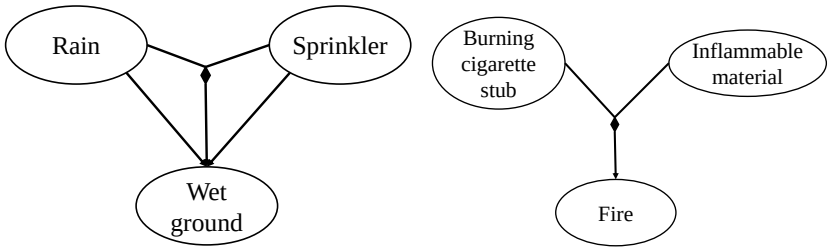


Figure 1.1: Multiple individual causes (the left figure) vs. the combined cause (the right one), where solid arrows denote causal relationships and the diamond represents that the combination of two variables causes an outcome.

for the latter each factor is a cause of the outcome already. Figure 1.1 shows the comparison between the two concepts. Sprinkler causes wet ground, and so does rain. Sprinkler and rain together cause wetter ground. Therefore, sprinkler and rain each is a cause of the wet ground. However, there is no causal link from burning cigarette stub or inflammable material to a fire, but the combination of these two factors leads to a fire. Actually combined causes commonly exist in various areas. For example, with gene regulation, it was found that the expression level of a gene might be co-regulated by a group of other genes, which could lead to a disease [16, 17].

Another important type of causal relationships involving multiple factors is the ***context specific causal relationships*** [18]. One factor, say a treatment A, does not cure a disease Y in general, but it will under certain conditions or contexts, e.g. for a group of patients with certain genetic profile [19]. Context specific causal relationships have important applications in various areas, including personalized treatment and online recommendation systems.

However, both combined causes and context specific causes are hidden and difficult to be discovered from observational data, as a single potential causal factor may not show causal signal in data. Specifically, to identify combined causes in data, one critical challenge is the computational complexity with large data sets, as the number of combined variables is exponential to the

number of individual variables. To the best of our knowledge, the problem of learning combined causes from data has received little attention, except the work in [20, 21, 22, 23]. Although a growing attention has been paid to the discovery of the context and heterogeneity of causal relationships [18, 24, 25], most existing methods assume a known cause and focus on identifying the contexts of heterogeneous effects of the cause from data.

In this thesis, we focus on studying these two major types of causal relationships that involve multiple factors. Our goal is to design efficient and automated methods to discover such causal relationships directly from observational data.

Firstly, a method is developed to mine combined causes from data. As mentioned above, the combined causes considered here are different from the generally discussed multiple causes. The main objective of the developed method is to find previously unobserved patterns and relationships in data, i.e. each component of a combined cause is not a cause but the combination is. Such combined causes are much more difficult to be detected even based on domain knowledge, especially when the number of variables under study is large (e.g. genetic data). To tackle the challenge, we propose a Multi-level Hiton-PC (MH-PC) method [26], based on an efficient local causal discovery method, HITON-PC [14], to discover the combined causes from data. Experiments with synthetic and real world data are conducted to assess its performance and efficiency.

Secondly, we further explore whether the component variables of a combined cause interact with each other to produce the outcome, instead of only having an additive effect on the outcome. Such an interaction between causal factors is called a ***cause interaction*** in the thesis. Figure 1.2 shows a typical example in real life, where genetic and environmental variables interact with each other in the process leading to a disease [1]. In medicine and pharmacy, it is important to perceive potential drug-drug interactions

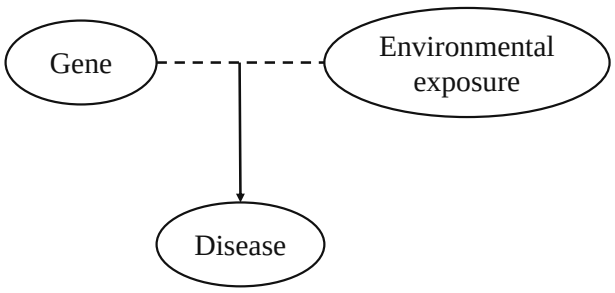


Figure 1.2: Cause interaction between genetic and environmental variables on a disease. (The dashed lines represent the interaction between two variables.) Example: $Gene = Pro12Ala\ allele$, $Environmental\ exposure = dietary\ fat\ intake$ and $Disease = obesity$ [1].

in advance [27], as interactions may increase or reduce the effect of drugs, or produce an adverse effect. The identification of cause interactions themselves is indeed very useful, as the knowledge of cause interactions has many real world applications. For example, if we know that *alcohol intake* and gene $ADH1C$ interact with each other to produce $myocardial\ infarction$ [1], reducing alcohol intake will reduce the risk of myocardial infarction.

It is challenging to detect interactions between multiple variables based on observational data only. Previous methods mostly use odds ratio and risk difference to measure the interactions between two variables [28, 29], but this type of interactions do not necessarily mean causality. VandeWeele et al. [30, 31] have made progress in detecting sufficient cause interactions under the sufficient-component cause model [32]. However, the work largely stays at theoretical level. We present a general framework, Data-driven Approach to Cause Interaction Discovery (DACID), and develop multiple instantiations of the framework to discover cause interactions. They are the first data mining algorithms for discovering cause interactions between causal factors directly from data.

Thirdly, a novel method, Context Specific Causal Rule discovery (CSCR), is proposed to explore context specific causal rules from data. Context

specific causal relationships are difficult to be discovered since the overall causal effect may be averaged out to be marginal in the whole population. For example, for a treatment, some patients respond positively and some respond negatively, and hence the overall effect among all patients is marginal. Furthermore, effective methods are required to detect contexts and causal relationships under the contexts at the same time, given the large number of possible combinations of contexts and causal factors.

Most existing methods assume a known cause and focus on identifying the contexts of heterogeneous effects of the cause from data. In contrast, our goal is to design a true data mining method to discover context specific causal relationships without knowing or assuming a cause, that is, to find both contexts and the causal relationships under the contexts simultaneously. In order to address the efficiency issue, we directly make use of the highly efficient and mature data mining techniques (e.g. decision tree and association rule mining) to select tentative causes and context variables with respect to a given target. Then within the much reduced search space, we employ the potential outcome model [33, 34] to assess the relationships to identify causes and their contexts.

Finally, we develop a causality based classification model, Causal Decision Tree (CDT) [35]. Each node of a CDT has a causal relationship with the outcome variable, under certain contexts. Specifically, the main objective of CDTs is to achieve a good predictive model with causal interpretations. Decision tree [36] has been widely used in many areas, including social and medical data analysis, due to its good classification performance, efficiency and graphical (tree) representation. However, it is not designed with causal discovery in mind. For example, in Figure 1.3, the decision tree built from a hypothesized data set shows that the use of Tinder (a matchmaking mobile application) is able to help cure the disease. Obviously, it is misleading since the majority of people using Tinder are young whereas most people not using Tinder are old. Young people will recover from the disease any-

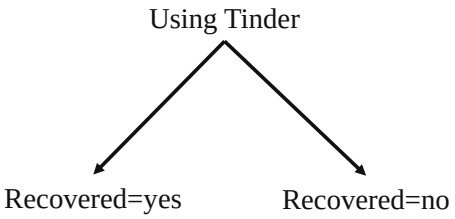


Figure 1.3: A simple decision tree.

way and old people have a lower chance of recovery. This misleading results from the non-causal characteristics of classification methods (e.g. decision tree). When building a CDT, we follow a well established causal inference framework and make use of a classic statistical test to establish the causal relationship between a predictor variable and the outcome variable. Therefore, a CDT has the potential to provide better insights into the causal mechanisms of the occurrence and changes of the outcome, and thus helps make more useful predictions and decision making.

To summarize, our goal is to develop efficient and automated methods for discovering causal relationships with multiple factors involved, and the main contributions of the thesis are:

1. We formally define the concepts of causal relationships involving multiple factors, i.e. combined causes, interactions among causal factors and context specific causes, and discuss their practical meaning and applications.

2. We propose two novel methods to mine combined and context specific causal relationships, respectively. To examine the interactions among causal factors, we develop a general framework and multiple algorithms instantiated from the framework. The major difference of the proposed methods, compared to existing methods, is that we detect causal relationships from data directly, while most existing methods hypothesize the causal relationships between variables in advance and

validate the hypothesized relationships. We also propose a causality based classification method for prediction and causal interpretations. The experiments with synthetic data sets are performed to demonstrate the effectiveness and efficiency of the proposed methods. The experiments with real world data sets show that the proposed methods are capable of providing insights and better understanding of data for predictions and decision making.

The rest of the thesis is organized as follows:

- In Chapter 2, we review the literature related to the work in this thesis, including methods to discover causal relationships with single causal factors and those involving multiple factors. We also identify gaps in the existing research work and the difference between existing work and our research.
- In Chapter 3, we introduce the background for understanding the following chapters.
- In Chapter 4, we propose a novel approach to mining the combined causes in large data sets, based on a well-established local causal discovery method.
- In Chapter 5, we introduce a general framework and develop four algorithms instantiated from the framework to detect cause interactions.
- In Chapter 6, we propose an efficient and practical method for exploring context specific causal relationships in data. With the proposed method, we capture both contexts and the causal relationships under the contexts in one go.
- In Chapter 7, we present causal decision tree, including its definition and the algorithm for learning a causal decision tree from data.

- Finally, in Chapter 8, we conclude the thesis and discuss the potential further work.

Chapter 2

Literature review

In this chapter, we review the existing research in the area of causal discovery with observational data, focusing on the work related to the research of the thesis.

As discussed in Chapter 1, Randomized Controlled Trials [5, 6] have been used as the gold standard for causal discovery, but they are often infeasible due to practical and ethical limits. Observational studies [7, 8, 9] are considered as the alternatives to RCTs, where investigators observe subjects and measure variables of interest without assigning treatments to the subjects.

For RCTs and observational studies, counterfactual inference [37], or the potential outcome model [33, 34] has been widely used for the investigation of causal effects. In general, counterfactual inference is to answer counterfactual questions [38, 39], e.g. "Would this patient recover from the disease, if the one received a different medication?". As suggested by the name of counterfactual inference, the inference model estimates the counterfactual/unobserved outcomes, where the treatment state is inconsistent with the actual state, and compares the difference between actual and counter-

factual outcomes to acquire the causal effects. Section 3.2 will present the complete idea of counterfactual inference in detail.

However, in both RCTs and observational studies, domain knowledge is required to generate the hypotheses about the cause-effect relationships to be tested, hence they can only be applied to problems with available domain knowledge and a handful of variables. In recent decades, many methods [4, 13] have been developed to learn causal relationships from data directly, in the fields of machine learning and data mining. Most of the methods are designed for finding a cause involving a single factor, which fall into two main categories: those for learning a global causal structure containing all the variables of interest; and those for learning a partial causal structure or a local causal structure around a fixed variable. Section 2.1 will review these two categories of causal discovery methods.

For detecting causes involving multiple factors, there has been less work done comparing to the effort of finding single-factor causes, and most of them are application oriented or stay at the theoretical level. In Section 2.2, we review existing research that addresses the problem of multi-factor causal relationship discovery, including the work related to combined causes, cause interactions and context specific causes.

2.1 Discovering causal relationships involving single factors

Among all the methods for finding causes involving single factors, the most notable method is the Causal Bayesian Network (CBN) framework. The structure of a CBN is represented as a Directed Acyclic Graph (DAG) [4], where a node represents a variable and an arrow is interpreted as a cause-

effect relationship between the parent node (cause) and the child node (effect). Many algorithms, including the score and search based methods and constraint based methods, have been developed for CBN strucutre learning. However, learning a CBN is still of high computational cost and may even be infeasible with high dimensional data, and thus much attention has been paid to the discovery of a partial or local causal structure. In Section 2.1.1, we review existing methods learning a global CBN from data. Sections 2.1.2 and 2.1.3 review the methods learning a partial and local causal structure, respectively.

2.1.1 Learning a global causal network

Over the years, many algorithms have been developed for learning a complete CBN structure from data to discover causal relationships. These methods are generally fall in two categories: the score and search based methods and constraint based methods, as discussed below.

A score and search method starts with a given search space of the possible structures (of the set of variables), and uses a scoring function and a search strategy to search the space for one or a set of structures that maximize the score based on the given data set. Given the large number of possible structures, exhaustive search over the search space is not feasible. Many heuristic search algorithms have been developed to build a CBN structure using all kinds of scoring criteria, such as the $K2$ score criterion [40] and the scoring criterion based on the minimum description length principle [41, 42].

In recent years, Genetic Algorithms (GAs) have become a popular approach to learning Bayesian network structures. Larranaga et al. [43] used GAs to search the best ordering of variables and adopted the $K2$ structure learning algorithm to evaluate the orderings. A closely related approach comes

from Hsu et al. [44], who had the same basic idea, but held back training data to produce a score for an ordering using importance sampling. Many researchers have investigated combining GAs with other techniques, e.g. machine learning approaches. For example, Wong et al. [45] developed a hybrid approach by combining the conditional independence tests and the search phases. Blanco et al. [46] incorporated a matrix transpose mutation operator into the Estimation of Distribution Algorithms (EDAs) [47] for Bayesian structure learning.

A constraint based method learns a CBN structure based on conditional independence tests [48] with a given data set. In the early 1990s, Geiger et al. [49] developed an algorithm to recover polytrees from an oracle, which indicates whether two variables are independent given other variables. At the same time, Fung and Crawford [50] developed a constructor system to learn undirected graphical models, also known as Markov networks, which encode conditional independence in a manner similar to Bayesian networks. The χ^2 statistical test was used in the method for testing conditional independence between variables.

One of the first systems that recovered a DAG from data is the SGS (Spirtes-Glymour-Scheines) algorithm developed by Spirtes et al. [13]. SGS starts with a complete undirected graph of random variables, and then performs conditional independence tests for edge elimination, collider detection and edge orientations. However, this is very inefficient, as each pair of variables required tests involving every subset of the remaining variables (all the variables except the pair of variables tested), and the number of the subsets increases exponentially as the number of variables.

An important advance, the PC (Peter-Clark) algorithm [13], was made by Peter Spirtes and Clark Glymour to improve the computational efficiency. Similar to SGS, the PC algorithm also starts with a complete undirected graph and then modifies the structure and orients the edges by using condi-

tional independent tests. The major difference is in the edge removal step: PC performs the tests in certain order and skips unnecessary tests. While SGS does conditional independence tests for each pair of variables X and Y conditioning on each set of other variables $\boldsymbol{S}$ ($\boldsymbol{S} = \boldsymbol{V} \setminus \{X, Y\}$, where $\boldsymbol{V}$ denotes the complete variable set), PC does the tests for X and Y in a level-by-level manner. The PC algorithm adopts the adjacency search to improve efficiency. Specifically it firstly tests conditioning on each single variable $S \in \boldsymbol{S}$. If $X \perp\!\!\!\perp Y|S$, it removes the edge between X and Y, otherwise tests will be performed conditioning on two or more variables. However, the adjacency search makes the output of PC order-dependent, which becomes very problematic in large (or high dimensional) data. Colombo et al. [51] proposed a PC-stable algorithm by removing part or all of this order-dependence.

Both PC and PC-stable fail to deal with the problem of unmeasured confounders. Inductive Causation (IC) [52, 53] was developed to take latent variables into account, and IC draws a causal graph with undirected, unidirected and bi-directed arrows, where a bi-directed arrow indicates that there exists a possible latent variable between two nodes. Inspired by the approach of IC, several faster algorithms, such as Fast Causal Inference (FCI) [13] and Real Fast Causal Inference (RFCI) [54], were proposed. They make less conditional independence tests and the tests condition on a relatively small number of variables.

Although huge efforts have been made and a lot of algorithms have been developed, the efficiency of learning a complete causal network is still an open issue. More recently, instead of taking a global view and constructing a complete causal network, many researchers have proposed methods that only learn a partial or local causal structure.

2.1.2 Learning a partial causal structure

Aiming at achieving high computational efficiency, some methods have been developed to find a sub-graph of the complete causal structures. Normally they focused on finding causal relationships with some fixed structures, e.g. the V-structures [55] and Y-structures [56].

The LCD algorithm [57] is a ploynomial time constraint based algorithm. With LCD, if variables A, B and C are pairwise dependent, and that A and C become independent given B, then the relationships between them can be in one of the following three causal structures:

$$A \to B \to C, A \leftarrow B \leftarrow C, or\ A \leftarrow B \to C.$$

Silverstein et al. [55] proposed a CCU rule to detect the V-structure. The rule states that if A and B are dependent, A and C are dependent, and B and C are independent, then if B and C become dependent given A, we can conclude that B and C each is a cause A, i.e.

$$B \to A \leftarrow C.$$

The LCD and CCU rules mostly depend on the correctness of the statistical tests. Once one test wrongly indicates the dependence relationship between variables, the structure obtained from these methods will be invalid [55].

Based on V-structure, Main et al. [56] developed a computationally feasible method for Y-structure discovery. Figure 2.1 shows an example of the Y-structure, where A, B and C forms a V-structure. A score and search based method was developed to learn a Y-structure from data. Moreover, this method has the potential to address possible latent variables, i.e., it does not assume no hidden common causes.

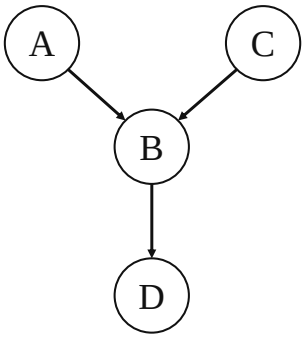


Figure 2.1: Y-structure.

2.1.3 Learning causal relationships around a fixed target variable

Recently, much attention has been paid to learning the causal relationships with respect to a fixed target. This type of methods have important applications in various areas, since people may be more concerned about the causes of some specific outcomes, e.g. lung cancer.

The first local-learning algorithm, Max Min Parents and Children (MMPC) [58, 59], was developed for identifying the parents and children of a given target in a CBN. MMPC can scale up to data sets with thousands of variables. The authors argued that local algorithms could be used with no degradation in quality to selectively learn parts of the CBN, when learning the full CBN is impractical. This method is tractable because it operates in much smaller variable set than the full set.

A related approach was discussed in [60], named HITON-PC, which introduced an efficient algorithm for causal feature selection. Given a fixed target variable Y, HITON-PC picks up variables one-by-one to perform conditional independence tests with respect to Y. A variable cannot be included in the parents and children set, if it is (conditional) independent with Y conditioning on other variables. The output of HITON-PC is the set of parents and

children of Y. Another version, semi-interleaved HITON-PC, is introduced in [14] to achieve higher computational efficiency.

Margaritis et al. [61] proposed the Grow-Shrink (GS) algorithm to discover causality from the Markov Blanket (MB). The algorithm firstly acquires the MB of the target node, and then tests conditional independence around the target to infer the local structure. Then it uses a heuristics algorithm to remove the possible spouse links and orient the arrows. Based on the GS algorithm, Pellet and Elisseeff proposed the TC algorithm [62, 63], which combines removal of spouses and orientation into a single pass and thus is faster. HITON-MB [14] is also a MB based method to do local learning.

PC-Select (also known as PC-Simple) [64] is a simplified version of the PC algorithm to learn local network structure. PC-Select was originally proposed to select variables (feature selection) from high dimensional data. It finds the variables that are strongly influential on a given target. PC-Select follows the idea of the PC algorithm to perform conditional independence tests, i.e. detect the conditional sets of a pair of nodes. Instead of learning a complete CBN, PC-Select is to learn the local causal relationships around a given target, i.e. to find the parents and children of the target.

2.2 Discovering causal relationships involving multiple factors

As mentioned in previous section, most existing causal discovery methods are designed to detect causal relationships involving single causal factors, instead of those involving multiple factors. However, given the importance of such causal relationships, some work, although at small scale, has been done to tackle the problem of multi-factor causal discovery, which will be

reviewed in the following.

2.2.1 Discovering the combination of causal factors

Some researchers have taken the approach of observational studies to investigate the effects of the combinations of multiple causal factors.

Kvaavik et al. [65] studied the individual and combined effects of smoking, alcohol, diet and some physical activities on an immature death, by using a cohort study [66]. In the study, 4886 individuals were involved in the study and records were collected in a follow-up period of 20 years. Regression is used to evaluate the relationships between health behaviors and mortality. The results have confirmed their hypothesis that the combination of poor health behaviors has a substantial effect on mortality.

Luo et al. [67] studied combined effects of smoking and Peripheral Arterial Disease (PAD) on all-cause and Cardiovascular Disease (CVD) mortality. Note that all-cause mortality refers to the total number of deaths due to any cause. Based on the status of smoking and PAD, a total of 1979 males 35 years of age or older were divided into four subgroups, and Cox proportional regression models were employed to estimate relative risks of all-cause and CVD mortality across different groups. This study concluded that the combined effect of smoking and PAD on mortality is slightly higher than the PAD's effect.

Cheng et al. [68] introduced a probabilistic contrast model to detect conjunction of factors on the outcome. Authors specified the analogy between causal induction and the ANalysis Of VAriance (ANOVA), and then developed an explicit model of causal induction. The probabilistic contrast model considered both single causes and conjunctive causes, where conjunc-

tive causes are defined as the causes involving multiple necessary factors to produce the outcome. 91 students in University of California, Los Angeles were involved in the study for the evaluation of their proposed model.

Although observational studies can be effective in examining the combined effect of multiple causal factors, they highly rely on domain knowledge (i.e. hypothesized causal relationships), which may not be available in practice. On the other hand, with the rapidly increased accumulation of observational data, a growing interest has been paid on developing data-driven discovery methods for mining combined causal factors.

Rehder [69] extended graphical models to capture independent (single) and conjunctive (combined) causes from data. However, as it is expensive to build a graphical model including both independent and conjunctive causal relationships from data, it is not practical to apply the method to solve the real world problems.

Jin et al. [20] proposed the Causal Rule discovery using Partial Association (CR-PA) algorithm for extracting causal rules from association rules. For an association rule $(\boldsymbol{X} = \boldsymbol{x}) \rightarrow (Y = y)$, where $\boldsymbol{X}$ is a predictor variable set (consisting of single or multiple variables) and Y is the outcome, CR-PA performs the partial association test to determine if $(\boldsymbol{X} = \boldsymbol{x}) \rightarrow (Y = y)$ is causal. Association rule mining was utilized to reduce the search space of causal rules and the partial association test was for the causal relationship test. CR-PA is capable of finding both single and combined causal factors. Meanwhile QuickSort was employed to improve the computational efficiency.

Li et al. [21, 22] proposed an algorithm, Causal Rule discovery with Cohort Study (CR-CS), to mine potential causal rules from observational data. The basic idea is to utilize the synergy of observational studies and association rule mining to develop an efficient method for automated discovery of causal

relationships in large data sets. Authors firstly used association rule mining to find the association rules regarding an outcome, which were treated as hypothesized cause-effect relationships. Then for each of the hypotheses, they conducted an observational study (i.e a respective cohort study) to test if the exposure is a real cause, i.e. to identify if the association rule is a causal rule.

Yadav et al. [23] introduced a similar work to mine frequent causal patterns from observational data. They first introduced the concept of closed intervention sets, which extends single interventions to concurrent interventions. Then association rule mining is used to find frequent patterns and the potential outcome model for the causal effect estimation. They also discussed how to correct for a number of common biases.

While the work has initiated the concept of the combined causes, their focus was on integrating association rule mining with observational studies or traditional statistical analysis for causal discovery. Different from their work, we aim to develop an automated and causality based method for efficiently discovering combined causes from observational data.

2.2.2 Discovering cause interactions

In many areas, such as medicine and psychology, detecting the interactions between multiple factors has attracted many research efforts in the past decades.

Cheng et al. [15] introduced a theory to examine the interactions between multiple causal factors. The overall idea is to compare the effects in different status of causal factors (presence or absence), to estimate the effect of the interactions between factors. Similarly, White [70] employed contingency

information to judge interactions between two causal factors. Specifically, contingency information is shown in a two-cause contingency table (4×2 table for binary variables), based on the status (presence or absence) of two factors and the status of the effect occurring or not.

These methods normally were designed to validate hypothesized cause interactions, which may be difficult to be generated even based on domain knowledge. Meanwhile, the number of potential predictor variables under their study is quite limited, and thus they may not be workable with real world problems.

More recently, data mining and machine learning strategies are applied to learn interactions from large data sets. Hahn et al. [71] proposed a maximum entropy probability model to search for genomic interactions on disease risks. Miller et al. [72] developed a Multi-factor Dimensionality Reduction (MDR) method for collapsing high-dimensional genetic data into a single dimension and then detected interactions in relatively small sample sizes. By employing Bayesian network learning and information gain, Jiang et al. [73] have developed multiple methods to discover interacting single nucleotide polymorphism, and successfully detected some insights from real world data. However, all these methods may not work well to find the interactions between the factors, which individually do not show strong causal signal.

Many efforts have been made in the field of epidemiology to assess whether causal factors interact with each other to produce an effect. An important progress has been made in the past decade with the proposal of the sufficient cause interactions [30], which are defined under the sufficient-component cause model [74]. VanderWeele et al. [30, 31] developed a general condition to detect sufficient cause interactions, and a looser condition by introducing a monotonic effect assumption. Extensive discussions about the interactions are presented in [75], including multiple scales and different

mechanistic forms of interactions. An incorporation model of sufficient-component causes and causal DAGs was developed in [76]. This model allows researchers to represent the sufficient cause interactions using a causal DAG.

However, the existing work on cause interaction discovery largely stays at theoretical level. More importantly there is no concrete computational methods for discovering the interactions between component causes, especially from large data sets.

2.2.3 Discovering context specific causes

Now we review the existing research on representing and discovering another major categories of multi-factor causal relationships, the context specific causal relationships.

A stream of research that is related to the topic is about deriving Context Specific Independence (CSI) based on a known Bayesian network [77, 78, 79], which enables the decomposition of a joint probability to a finer-grained factorization than conditional independence, and thus some researchers intended to capture CSI to speed up the Bayesian network inference algorithms.

Boutilier et al. [80] firstly formalized the concept of CSI and developed a Conditional Probability Table tree (CPT-tree) to capture CSI based on a Bayesian network. An example of CPT-tree is shown in Figure 2.2. 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 CPT, where the probabilities of Y given all value assignments of its parents are listed. Given a context, i.e. one or more parent nodes taking an assignment of a value, the

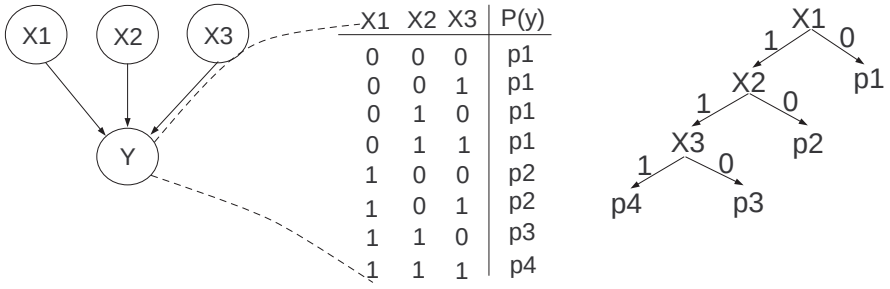


Figure 2.2: An illustration of CPT-tree. (L): A Bayesian network; (M): Conditional probability table of Y ; (R): CPT-tree.

probability of Y may be constant (without being affected by the values of other parents). So a CPT can be represented clearly with a tree structure. However, the size of a CPT is exponential to the number of parent nodes of Y, and the table will be difficult to display.

Zhang et al. [81] studied the role of CSI in general and discussed the computational leverages afforded by CSI without inference algorithms specified. Poole et al. [82] presented a compact representation that exploits conditional probabilities of variables under their parents (i.e. contexts). Authors proposed contextual variable elimination, i.e. opportunistic elimination of variables in contexts, to improve the computational efficiency without affecting correctness.

Instead of aiming at fast inference with Bayesian networks, some others focused on extending a Bayesian network by adding special notations, e.g. labels on the edge, such that the extended Bayesian network can explicitly present the context specific causal relationships. Milch et al. [83] introduced contingent Bayesian networks, where each edge was labelled with a condition and the edge was active only under the condition. Structural conditions were discussed to guarantee a contingent Bayesian network define a unique distribution. Similar diagrammatic notations are introduced for graphical representation model to describe context specific dependencies

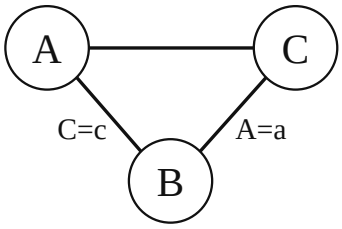


Figure 2.3: An example of SGM to present the context specific independence.

[84, 85]. Gates [86] are another notation to explicitly represent context-sensitive independence in a graphical model. The dependence within a gate only exists when the variable having specific values. Gates also can work as a data structure to perform inference.

Nyman et al. [87] proposed a general class of Stratified Graphical Models (SGMs) based on looser restrictions. Figure 2.3 shows an example of SGM, where a label on an edge denotes that the context specific independence holds with the stratum. For example, the label on the edge between B and C indicates that B and C are independent conditioning on the stratum $A = a$, i.e. $B \perp\!\!\!\perp C|A = a$. This model enables us to display the context specific independence between distinct pairs of variables under different strata. Authors also extended this model to the multivariate Gaussian family, Stratified Gaussian Graphical Models (SGGMs) [88]. The Euclidean space is employed to specify certain segments, under which the context specific independence holds. SGGMs are potential to develop efficient inference methods for continuous systems.

These graphical model based methods typically assume a known graphical model, such as a Bayesian network and then exploit CSI based on the networks. However, this is impractical in many real world applications since we do not know the graphical model or we could not build it from a data set, particularly for a large data set. Secondly, in a Bayesian network, the parents of a node Y are all global causes of Y. 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 Bayesian network and thus not by these methods either.

Another main stream of research that is related to context specific causal discovery is subgroup analysis. Subgroup analysis is commonly used to evaluate the treatment effects in a specific sub-population defined by some context variables.

Bonetti et al. [24] proposed a method to test the survival time with different treatments, on the subset of patients who have higher (or smaller) values of a covariate of interest. An alternative approach, Multivariable Fractional Polynomial Interaction (MFPI) [89], was developed to compare treatment effects in different subgroups defined by one particular covariate.

There has been an increasing interest in applying machine learning methods to subgroup analysis, for estimating the heterogeneity of causal effects in different sub-populations.

Su et al. [90] adapted the idea of recursive partitioning to construct an Interaction Tree (IT) for causal effect estimation in different subgroups. The IT procedure can define a variety of sub-populations automatically, where modified Classification And Regression Tree (CART) is employed. Foster et al. [91] introduced the virtual twins method to define subgroups with enhanced treatment effect, where random forest [92] was employed 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. [93] developed an approach to get the optimal decision making via the technique of Doubly Robust (DR) estimation, which is a statistical estimation approach for incomplete data. Athey et al. [18] utilized a popular data mining technique to find the contexts (sub-populations) across which heterogeneous effects of a treatment

can be observed. Specifically, it applied regression tree with a modified MSE (Mean Squared Error) criterion to partition the population into subsets, and then in each sub-population the causal effect of a known cause was estimated, so that the differences of the effects across the sub-populations can be investigated.

Tian et al. [94] developed a simple approach to estimate treatment-covariate interactions. An interaction between a treatment and covariates refers to that the treatment has different effects on the outcome when the covariates have different values. They coded the treatment as ± 1 and included the modified covariates (the products of original covariates and the coded treatment) in a standard model (e.g. LASSO [95]). Imai et al. [25] expressed the heterogeneity of effect estimation as a feature selection problem, where different LASSO penalty terms are used over treatment variable and covariates. Then they can estimate the conditional treatment effect by the predicted potential outcomes from the fitted model. Green et al. [96] introduced a Bayesian Additive Regression Trees (BART) to estimate the heterogeneity of treatment effects. BART detected the interactions between the treatment and covariates automatically, and thus reducing the researchers' subjective selection bias. Chen et al. [97] focused on identifying qualitative treatment-covariate interactions. A Bayesian approach was proposed to search for qualitative treatment-covariate interactions with a multiple regression setting.

These subgroup analysis methods assume a known cause or treatment, and aim to test hypotheses about the heterogenous causal effects of the specified treatment in different sub-populations. Although these techniques can be straightforwardly extended to search over many possible causes and corresponding sub-populations, their computational costs are normally very high and it may not be feasible in practice.

2.3 Conclusion

Over the last decades, great efforts have been made in discovering and inferring causal relationships. In this chapter, we have reviewed the existing research work, focusing on the causal relationship discovery from observational data.

One of the most advanced theories of causal discovery is the causal Bayesian network and many algorithms are proposed to learn CBNs directly from data. To achieve higher computational efficiency, researchers have studied to learn a partial causal structure or a local causal structure around a fixed target variable. A major limitation of these methods is that they are designed to capture causal relationships between a single factor and an outcome.

A growing attention has been paid to studying complex causal relationships from observational data. However, some existing methods largely stay at theoretical level and cannot be applied to real world problems, while some others have made assumptions about the causal relationships in advance and focused on validating these hypotheses based on the observational data. There are very few concrete computational methods for discovering complex causal relationships involving multiple factors, especially from large data sets.

Motivated by the gaps in current literatures, we aim to study complex causal relationships with multiple factors involved and develop efficient and automated methods to discover these relationships from observational data.

Chapter 3

Background

In this chapter, we firstly introduce Causal Bayesian Networks [13], a main framework for representing and learning causal relationships. We also present a well-known local causal discovery method, HITON-PC [60], which is the basis of the proposed method in Chapter 4. Then we present a well established framework for causal inference, the potential outcome model, and the techniques involved in the model. The potential outcome model will be utilized in Chapters 5, 6 and 7 for causal relationship discovery.

3.1 Causal Bayesian networks and HITON-PC

A Bayesian network (BN) [98] provides a graphical representation of the joint probability distribution of a set of random variables, by encoding the conditional independence relationships between the variables with the graph. A BN can be formally defined as:

Definition 3.1 (Bayesian network) *[98] A Bayesian network* $BN = \langle G, \boldsymbol{P} \rangle$ *consists of a directed acyclic graph* $G = \langle \boldsymbol{V}, \boldsymbol{E} \rangle$ *and a set of conditional probability distributions* $\boldsymbol{P} = \{Pr(X|pa(X))|X \in \boldsymbol{V}\}$. *The conditional probability distribution* $Pr(X|pa(X))$ *is associated with the variable* $X \in \boldsymbol{V}$ *and* $pa(X)$ *is the set of parents of* X.

For a Bayesian network, the joint probability of the variables can be expressed as the product of conditional probability distributions:

$$Pr(\boldsymbol{V}) = \prod_{X \in \boldsymbol{V}} Pr(X|pa(X)) \tag{3.1}$$

A causal Bayesian network is a Bayesian network where each edge is interpreted as a cause-effect relationship from the parent node to the child node. It can be learned from data under certain causal assumptions, commonly including the causal sufficiency, Markov condition and the faithfulness assumption [13].

Definition 3.2 (Causal sufficiency [13]) *A set of variables* $\boldsymbol{V}$ *is said to be causally sufficient if and only if any common cause of two or more variables in* $\boldsymbol{V}$ *is also in* $\boldsymbol{V}$.

Definition 3.3 (Causal Markov condition [13]) *A causal Bayesian network* $\langle G, \boldsymbol{P} \rangle$ *satisfies causal Markov condition, if and only if for each variable* $X \in \boldsymbol{V}$, X *is conditionally independent of the set of all its nondescendents given* $pa(X)$.

Definition 3.4 (Faithfulness [13]) *A causal Bayesian network* $\langle G, \boldsymbol{P} \rangle$ *satisfies the faithfulness assumption, if and only if every conditional independence relation in* $\boldsymbol{P}$ *is entailed by the causal Markov condition applied to* G.

As mentioned in Chapter 2, there are two types of algorithms to learn a CBN, the score and search based methods and constraint based methods. In the following, we introduce the PC algorithm, a well-known constraint based method. The HITON-PC algorithm used in our method (Chapter 4) also conducts conditional independence tests as the PC algorithm.

The PC algorithm includes two main steps. Firstly, it creates a fully connected undirected graph containing all variables and employs Conditional Independence (CI) tests to determine if an edge is to be removed. The outcome of this step is a skeleton graph learned from data. Specifically for each pair of variables (X, Y), the PC algorithm tests if X and Y are independent given other variables $S \subset \boldsymbol{V} \backslash \{X, Y\}$ (S can be an empty set), i.e. $X \perp\!\!\!\perp Y|S$. If any variable set exists s.t. $X \perp\!\!\!\perp Y|S$, then the edge between X and Y is removed. In the second step, the PC algorithm orients the skeleton graph to learn a Completed Partially Directed Acyclic Graph (CPDAG), where edges are partially directed. Compared with an earlier version of causal Bayesian network learning, i.e. the SGS algorithm [13], the efficiency of the PC algorithm derives that it performs the conditional independence tests in some order and dynamically update the structure. Thus the computational efficiency is significantly improved when the true graph is sparse.

HITON-PC is a local causal discovery algorithm to find the Parents and Children (PC) set of a fixed target variable. Similar to the PC algorithm, HITON-PC also applies CI tests to identify causal relationships between variables, but it utilizes a different approach to reduce the search space for the tests.

The semi-interleaved HITON-PC is a variant of the HITON-PC algorithm to achieve higher computational efficiency. Here we introduce the semi-interleaved HITON-PC algorithm in detail, as it is used as the basis of the proposed algorithms in Chapter 4. For the simplicity of the presentation,

ALGORITHM 3.1 The Semi-interleaved HITON-PC Algorithm [60]

Input: A data set $\boldsymbol{D}$ for predictor variable set $\boldsymbol{V}$ and the target Y

Output: $TPC(Y)$, the set of parents and children of Y

1: $TPC(Y) \leftarrow \emptyset$
2: Let *OPEN* contain all variables associated with Y, sorted in descending order of strength of associations.
3: **repeat**
Phase I: Inclusion (line 4)
4: Move the first variable X from *OPEN* to the end of $TPC(Y)$
Phase II: Elimination (lines 5-16)
// Forward stage
5: **if** $OPEN \neq \emptyset$ **then**
6: $X \leftarrow$ the variable last added to $TPC(Y)$
7: **if** $\exists \boldsymbol{S} \subset TPC(Y) \backslash \{X\}$, s.t. $I(X, T \mid \boldsymbol{S})$ **then**
8: Remove X from $TPC(Y)$
9: **end if**
// Backward stage
10: **else**
11: **for** each $X \in TPC(Y)$ **do**
12: **if** $\exists \boldsymbol{S} \subset TPC(Y) \backslash \{X\}$, s.t. $I(X, T \mid \boldsymbol{S})$ **then**
13: Remove X from $TPC(Y)$
14: **end if**
15: **end for**
16: **end if**
17: **until** $OPEN = \emptyset$
18: Output $TPC(Y)$

we use HITON-PC to refer to the semi-interleaved HITON-PC in the rest of this thesis.

Referring to ALGORITHM 3.1, the input of the algorithm includes a data set $\boldsymbol{D}$ for the predictor variable set $\boldsymbol{V}$ and the target Y. The output is the set of parents and children of Y. In the algorithm, HITON-PC uses two data structures, a priority queue *OPEN* and a list $TPC(Y)$. Initially *OPEN* contains all predictors associated with Y and $TPC(Y)$ is empty. It then iterates between the two phases, inclusion and elimination, until *OPEN* becomes empty.

In the inclusion phase, the variable having the strongest association with Y is removed from *OPEN* and added to $TPC(Y)$ (line 4). In the elimination phase, if *OPEN* is not empty, the forward stage (lines 5-9) is executed. The variable newly added to $TPC(Y)$, X, is eliminated from $TPC(Y)$ if it is independent of Y given a subset of current $TPC(Y)$, otherwise it is kept (still tentatively) in $TPC(Y)$. If *OPEN* is empty, the backward stage (lines 10-16) is activated, and each variable X in current $TPC(Y)$ is tested, and if a subset of $TPC(Y)$ is found such that X is independent of Y given the subset, X is removed from $TPC(Y)$.

HITON-PC uses several heuristics to improve efficiency. At the forward stage, CI tests are conducted only on the newly added variable, instead of performing a full variable elimination. To compensate for possible false discoveries caused by this heuristic, HITON-PC uses the backward stage to "tighten up" $TPC(Y)$ by testing the conditional independence of each variable with Y given the other variables in $TPC(Y)$. Moreover, the use of the priority queue, *OPEN*, allows variables having stronger associations with Y to be included and evaluated first. As these variables are more likely to be the true parents or children of the target, once they are in $TPC(Y)$, it is expected that given these variables, other variables that should not be in $TPC(Y)$ are quickly identified and removed so that in the forward stage

TPC(Y) will not be over expanded, thus reducing the number of CI tests. Additionally, in practice, HITON-PC restricts the maximum order of CI tests to a given threshold *max-k*, i.e. the maximum size of the conditioning set ($\boldsymbol{S}$, see ALGORITHM 3.1) is *max-k*.

The time complexity of HITON-PC mainly depends on the number of CI tests. Each variable needs to be tested on all subsets of *TPC*(Y). Thus the complexity regarding each variable is $O(2^{|TPC(Y)|})$, and the total time complexity is $O(|\boldsymbol{V}|2^{|TPC(Y)|})$. When *max-k* is specified, the complexity becomes polynomial, i.e. $O(|\boldsymbol{V}||\,TPC(Y)|^{max-k})$. Extensive experiments have shown that HITON-PC is able to cope with thousands of variables with low rate of false discoveries [14].

3.2 The potential outcome model

Counterfactual inference [37, 99] has become increasingly standard for causal inference in many areas, e.g. epidemiological and medical studies. Rubin [38] has firstly applied the counterfactual inference to statistical inference in observational studies and developed the potential outcome model [33, 34], which is a well-established framework for estimating causal effects in observational or experimental studies. Here we introduce the basic concepts of the model and a principle for estimating the average causal effect, mainly following the introduction in [100].

Let X be a predictor variable and Y the outcome attribute where $x \in \{0, 1\}$ and $y \in \mathbb{R}_{\geq 0}$. For easy discussion, we consider that $X = 1$ is a treatment and $Y = 1$ the recovery. In 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. Note that we also use Y_1 and Y_0 to denote Y_i^1 and Y_i^0 in the thesis, when we focus on the population level potential outcomes. Then we have the following definition.

Definition 3.5 (Individual level Causal Effect (ICE)) *The individual level causal effect is defined as the difference of two potential outcomes of an individual, i.e.* $\delta_i = Y_i^1 - Y_i^0$.

The causal effects of individuals in a population are normally aggregated to get the average causal effect as defined below, where $E[.]$ stands for the expectation operator in probability theory.

Definition 3.6 (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.* $E[\delta_i] = E[Y_i^1] - E[Y_i^0]$.

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 $\boldsymbol{D}$, the ACE, $E[\delta_i]$ can be estimated as:

$$\begin{aligned} E_{\boldsymbol{D}}[\delta_i] =& \pi(E_{\boldsymbol{D}}[Y_i^1|X_i=1] - E_{\boldsymbol{D}}[Y_i^0|X_i=1]) + \\ & (1-\pi)(E_{\boldsymbol{D}}[Y_i^1|X_i=0] - E_{\boldsymbol{D}}[Y_i^0|X_i=0]) \end{aligned} \tag{3.2}$$

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, $(Y_i^0|X_i = 0)$[1]. However, we cannot observe the counterfactual outcomes, $(Y_i^0|X_i = 1)$ and $(Y_i^1|X_i = 0)$, since the investigated state did not really happen. This observation is described as the "fundamental problem of causal inference" [101] or a missing data problem [102]. We have to estimate the counterfactual outcomes: 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 $\boldsymbol{D}$ we can obtain the "naïve" estimation of the ACE:

$$E_{\boldsymbol{D}}^{naive}[\delta_i] = E_{\boldsymbol{D}}[Y_i^1|X_i = 1] - E_{\boldsymbol{D}}[Y_i^0|X_i = 0] \tag{3.3}$$

The question is when the naïve estimation (Equation 3.3) will approach the true estimation (Equation 3.2). In other words, how can we unbiasedly estimate the counterfactual outcomes? The solution is to balance the covariate distributions in the treatment and control groups, such that individuals have the same characteristics apart from being treated or not. In a randomized controlled trail, it can be achieved via randomized treatment assignment mechanism. In an observational data set, however, the random assignment is not possible.

Matching methods are a commonly used approach to mimicking randomized treatment assignment in non-controlled studies. We will detail the techniques related to matching methods in the next section. Here we use

[1]For individual i, the notation Y_i^t represents the potential outcome in the treatment state $t = 1$ or in the control state $t = 0$. I also use $x_i = 1$ *or* 0 to indicate the potential state that individual i actually receives. Obviously, if $t = x_i$ the potential outcome is observed and named as an observed outcome, a counterfactual outcome otherwise. For example, $Y_i^1|X_i = 1$ means the potential outcome in treatment state for the one actually treated, and thus it can be observed. $Y_i^1|X_i = 0$ is counterfactual instead.

perfect stratification to present the idea how the potential outcome model can estimate the counterfactual outcomes and causal effects. Let the differences of individuals in a data set be characterised by a set of attributes $\boldsymbol{C}$ (excluding X and Y) and let the data set be perfectly stratified by $\boldsymbol{C}$. 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:

$$E[Y_i^1|X_i = 0, \boldsymbol{C} = \boldsymbol{c}_k] = E[Y_i^1|X_i = 1, \boldsymbol{C} = \boldsymbol{c}_k] \tag{3.4}$$

$$E[Y_i^0|X_i = 1, \boldsymbol{C} = \boldsymbol{c}_k] = E[Y_i^0|X_i = 0, \boldsymbol{C} = \boldsymbol{c}_k] \tag{3.5}$$

where $\boldsymbol{C} = \boldsymbol{c}_k$ indicates a stratum of perfect stratification. Since individuals are indistinguishable in the stratum, unobserved potential outcomes can be estimated by observed ones directly. 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 3.4), 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 3.5). By replacing Equation 3.2 with Equations 3.4 and 3.5, we have:

$$\begin{aligned}
&E_{\boldsymbol{D}}[\delta_i|\boldsymbol{C} = \boldsymbol{c}_k] \\
=&\pi(E_{\boldsymbol{D}}[Y_i^1|X_i = 1, \boldsymbol{C} = \boldsymbol{c}_k] - E_{\boldsymbol{D}}[Y_i^0|X_i = 1, \boldsymbol{C} = \boldsymbol{c}_k])+ \\
&(1-\pi)(E_{\boldsymbol{D}}[Y_i^1|X_i = 0, \boldsymbol{C} = \boldsymbol{c}_k] - E_{\boldsymbol{D}}[Y_i^0|X_i = 0, \boldsymbol{C} = \boldsymbol{c}_k]) \\
=&\pi(E_{\boldsymbol{D}}[Y_i^1|X_i = 1, \boldsymbol{C} = \boldsymbol{c}_k] - E_{\boldsymbol{D}}[Y_i^0|X_i = 0, \boldsymbol{C} = \boldsymbol{c}_k])+ \\
&(1-\pi)(E_{\boldsymbol{D}}[Y_i^1|X_i = 1, \boldsymbol{C} = \boldsymbol{c}_k] - E_{\boldsymbol{D}}[Y_i^0|X_i = 0, \boldsymbol{C} = \boldsymbol{c}_k]) \\
=&E_{\boldsymbol{D}}[Y_i^1|X_i = 1, \boldsymbol{C} = \boldsymbol{c}_k] - E_{\boldsymbol{D}}[Y_i^0|X_i = 0, \boldsymbol{C} = \boldsymbol{c}_k] \\
=&E_{\boldsymbol{D}}^{naive}[\delta_i|\boldsymbol{C} = \boldsymbol{c}_k]
\end{aligned} \tag{3.6}$$

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

Observation 3.1 (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.

3.3 Matching methods

When randomization of treatment assignment is not possible, matching methods are normally used to reduce the bias due to the covariates $\boldsymbol{C}$. Two key components are involved in the matching: the distance measure and a method to do the matching.

Distance measures are utilized to measure the similarity of two individuals i and j in terms of their covariate values. The simplest measure is that two individuals i and j have exactly the same values of covariates $\boldsymbol{C}$. This exact measure may have a low statistical power for detecting dependency in data. For example, Chapin [103] only find 23 pairs that exactly match on six variables from a pool of 671 treated and 523 controls.

Mahalanobis distance and Jaccard distance [2] are also commonly used to measure the distance between individuals. Although these measures are better than exact distance to deal with multiple covariates, it is still hard to do matching in high dimensional data.

The proposal of propensity score [33] made an important advance to define the similarity of individuals. The propensity score is defined as the proba-

bility of the individual receiving the treatment $X = 1$ conditioning on the covariates $\boldsymbol{C}$:

$$e(\boldsymbol{C}) = Pr(X = 1|\boldsymbol{C} = \boldsymbol{c}). \tag{3.7}$$

where propensity score are mostly estimated by using logistic regression [104], boosted CART [105] and Generalized Boosted Models (GBM) [106].

Propensity score facilitates the balance of distributions of the covariates. Rubin et al. [104, 107] analyzed the approximate reduction in bias with a linear combination of normally distributed covariates. We can determine in advance the possible bias reduction with some information known, e.g. original sample sizes and the correlation between the covariates and the outcome [107].

However, normal propensity score method cannot handle the problem involving multiple (> 2) treatments. Imbens [108] proposed the Generalized Propensity Score (GPS) to extend binary treatment to multiple treatments. The GPS is defined as the conditional probability of receiving a particular treatment $X = x$ given covariates $\boldsymbol{C}$:

$$g(x, \boldsymbol{C}) = Pr(X = x|\boldsymbol{C} = \boldsymbol{c}) \tag{3.8}$$

Thus each individual obtains a vector of generalized propensity scores $G(\boldsymbol{C}) = (g(x_1, \boldsymbol{c}_1)), \ldots, g(x_Z, \boldsymbol{c}_Z)$, the conditional probabilities of receiving Z different treatments, respectively. Normally multinomial logistic regression is used to obtain the GPS vector [109].

With distance measure selected, we can choose different matching methods to do matching. Apart from exact matching mentioned above, nearest neighbour matching is the most commonly used method [110]. We can select optimal matched control or ratio matched (i.e. multiple) controls for

each treated individual, and with or without replacement based on the size of control and treated groups.

Subclassification [111] divides individuals into multiple subclasses, within each of which individuals have exact or similar distributions of covariates. use of 5 to 10 subgroups is a current convention [111]. Coarsened Exact Matching (CEM) [112] is a matching method based on monotonoic imbalance bounding. With CEM, the balance between two groups is selected by the user ex ante, and thus the imbalance will never be larger than the ex ante user choice. More recently, cardinality matching methods have been developed to maximize the size of the matched samples subject to constraints on covariate balance [37, 113, 114]. Cardinality matching directly balances the original/selected covariates, and thus its quality does not rely on the specification of correct model on the treatment assignment [115].

3.4 Conclusion

In this chapter, we firstly introduce causal Bayesian networks and the PC algorithm. Then we present a well-known local causal discovery method, HITON-PC, which is employed to build a novel method in Chapter 4 for finding combined causes from observational data. We also present the potential outcome model that is designed for causal inference. The model is used in Chapters 5, 6 and 7 to examine the causal relationships.

Chapter 4

Mining combined causes in large data sets

In this chapter, we study the discovery of combined causes from observational data. As illustrated in Figure 1.1, the combined cause is different from the generally discussed multiple causes. We study the case when multiple variables each alone is not sufficient to cause an effect, but their combination is.

As discussed in Chapter 1, most causal discovery methods are developed to find a cause consisting of a single variable, whereas the combined causes cannot be discovered with these methods, such as Causal Bayesian Networks. In a CBN, an edge is drawn from A to C only when A is a cause of C. Meanwhile, an edge between A and C in a CBN implies that A and C are not conditional independent. In contrast, if A and B each alone is not a cause of C, then each of A and B is (conditional) independent with C. Thus no edge is drawn from A or B to C, and it is impossible to examine the combined causal effect of A and B on C. This limitation of CBNs was discussed in [13] (page 48) as follows:

> "Suppose drugs A and B both reduce symptoms C, but the effect of A without B is quite trivial, while the effect of B alone is not. The directed graph representations we have considered in this chapter offer no means to represent this interaction and to distinguish it from other circumstances in which A and B alone each have an effect on C."

To identify combined causes in data, one critical challenge is the computational complexity with large data sets, as the number of combined variables is exponential to the number of individual variables.

To tackle this challenge, we employ an efficient local causal discovery method, HITON-PC [14], in our design of an automated method for discovering combined causes from observational data. Moreover, to find combined causes in data, in fact, it is not necessary to consider all combinations of variables. Particularly, we are not interested in a combined variable, e.g. $W = BG$ in Figure 4.1, of which a component B is a cause already, as it is reasonable to assume that the causal relationship between W and the target Y is due to the relationship between B and Y. To improve efficiency, we can exclude such combined variables when finding combined causes of Y, and only consider the combined variables whose components are not causes of Y. In the following, we firstly define the combined cause, state the problem and then present our proposed method.

This chapter is based on the work published in [26].

4.1 Problem statement and definitions

Before the problem statement, we firstly define the notation to be used in this chapter. We use upper case letters, e.g. A and B, to represent random

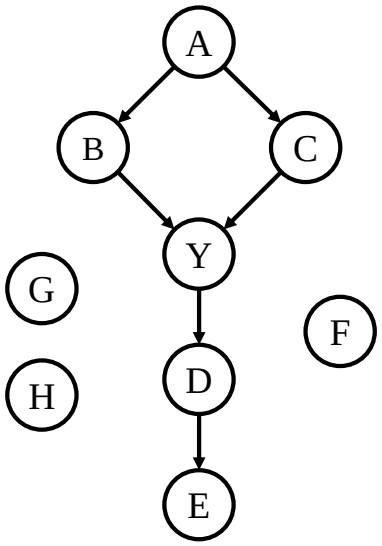


Figure 4.1: A random Bayesian network.

variables, and multiple upper case letters, e.g. AB (or $A\&B$), to denote the combined variable consisting of A and B. Bold-faced upper case letters, e.g. $\boldsymbol{A}$ and $\boldsymbol{B}$, represent a set of variables. Particularly, we denote the set of predictor variables and the target variable with $\boldsymbol{V}$ and Y respectively. The conditional independence between A and Y given $\boldsymbol{S}$ is represented as $A \perp\!\!\!\perp Y \mid \boldsymbol{S}$.

We deal with binary variables only in this chapter, i.e. each variable has two possible values, 1 or 0. The value of a combined (binary) variable AB is 1 if and only if each of its component (binary) variables is equal to 1 (i.e. $A = 1$ and $B = 1$).

The combined causes studied in this chapter can be defined as follows.

Definition 4.1 (Combined Cause) *Let W be a combination of multiple variables. W is a combined cause of Y if W is a cause of Y and any of its component variables, $X \in W$, is not a cause of Y.*

The research problem in this chapter is to discover combined causes from data, which can be stated as follows.

Problem 4.1 *Given a data set* $\boldsymbol{D}$ *with a set of predictor variables* $\boldsymbol{V}$ *and the target variable* Y*, find the combined variables* $W \subset \boldsymbol{V}$*, such that* W *is a combined cause of* Y *based on Definition 4.1.*

4.2 Uncovering combined causes

In this section, we firstly introduce the naïve approach to discovering combined causes and discuss the issues with the approach, which leads to the basic idea of our proposed method. Then we develop and evaluate the multi-level HITON-PC algorithm followed by the discussion of its time complexity.

4.2.1 The naïve approach

The naïve or straightforward way to detect the combined causes from data is as follows. Firstly, we generate a new variable set with combined variables using the original variable set. For example, for $\boldsymbol{V} = \{A, B, C, D, E, F, G, H\}$ (see Figure 4.1), the new variable set (with $2^8 - 1$ variables) is $\boldsymbol{V'} = \{A, ..., H,\ AB, ..., GH,\ ABC, ..., FGH,\ ..., ABCDEFGH\}$. Then we run a local causal discovery algorithm, such as HITON-PC, to find both single and combined causes using the data set created for $\boldsymbol{V'}$.

The naïve approach, however, is not feasible because the number of combined variables is exponential to the number of individual variables. In the following, we discuss how our proposed method tackles the problem.

4.2.2 Basic idea of the proposed method

Based on Definition 4.1, we can design an algorithm to find such combined causes (and single causes) in a level by level manner. We firstly, at the $(k-1)^{th}$ *level* ($k \geq 2$), obtain $TPC_{k-1}(Y)$, the set of parents and children of Y, each consisting of $k-1$ individual variables. Then at the k^{th} *level*, combined variables are generated based on $nTPC(Y)$, the set of non-cause variables at all the lower levels, i.e. $nTPC(Y) = nTPC_1(Y) \cup \ldots \cup nTPC_{k-1}(Y)$, where $nTPC_i(Y)$ ($i \in \{1, \ldots, k-1\}$) is the set of non-cause variables each containing i individual variables. Specifically, a k^{th} *level* combined variable can be generated by combining an i^{th} *level* ($i \in \{1, \ldots, k-1\}$) non-cause variable and a $(k-i)^{th}$ *level* non-cause variable.

For non-cause variables A and B, if the combination AB is a combined cause, then it is reasonable to assume that A positively contributes to the relationship between B and the target Y and vice versa. For example, the combustible dust suspended in the air (even at a high concentration) has no causal effect on a dust explosion, but an ignition source will improve their relationship significantly and thus the combination of the two factors can result in a dust explosion. This observation leads to the following definition.

Definition 4.2 (Redundant Combined Variable) *For $\forall A, B \in \boldsymbol{V}$, the combination AB is a redundant combined variable, if either $A \perp\!\!\!\perp Y \mid B = 1$ or $B \perp\!\!\!\perp Y \mid A = 1$, where A and B are not individual causes of the target Y.*

By excluding redundant combined variables, we can further improve the efficiency of the causal discovery.

Based on the above discussion and HITON-PC, we propose the Multi-level HITON-PC (MH-PC) method for finding both single and combined causes

of a given target. In the following section, we present the details of the method.

4.2.3 Multi-level HITON-PC

Referring to ALGORITHM 4.1, at the first level, MH-PC invokes HITON-PC to find the single causes of Y, $TPC_1(Y)$ (line 1) and initiates $TPC(Y)$ as $TPC_1(Y)$ (line 2). The single non-cause variables are put in $nTPC_1(Y)$ and $nTPC(Y)$ (non-cause variables identified at all the lower levels) is initially empty (line 2).

At level k ($k \geq 2$), MH-PC firstly updates $nTPC(Y)$ so that it contains the non-causes from levels 1 to $k-1$ (line 4). Next the algorithm generates combined variables containing k individual variables by combining the variables in $nTPC(Y)$ (line 5). Redundant combined variables are then removed (line 6) and the new data set D_k for level k combined variables ($\boldsymbol{V_k}$) is created too (line 7). From lines 8 to 23, we identify level k combined causes from $\boldsymbol{V_k}$. Initially *OPEN* contains all the combined variables in $\boldsymbol{V_k}$ which are associated with Y and the variables are sorted in descending order of the strength of associations (line 8). Similar to HITON-PC, the inclusion and elimination phases are carried out iteratively till *OPEN* is empty. At the end of the iteration (line 23), $TPC_k(Y)$ includes the discovered combined causes consisting of k variables, and $TPC(Y)$ includes all the causes from level 1 to level k. Note that in line 17, to improve the efficiency further, the backward stage only checks the level k candidates in $TPC_k(Y)$, instead of all candidates in $TPC(Y)$, as all the lower level parents and children have been confirmed at previous levels.

In line 24, the set of level k non-causes is updated before completing the work at level k. Finally MH-PC outputs $TPC(Y)$ until $k = k_{max}$ (the

ALGORITHM 4.1 **M**ulti-level **H**ITON-**PC** (**MH-PC**)

Input: A data set $\boldsymbol{D}$ for predictor variable set $\boldsymbol{V}$ and target Y;
k_{max}, the maximum level of causal discovery
Output: $TPC(Y)$, the set of (single and combined) parents and children of Y

1: $TPC_1(Y)$ = HITON-PC($\boldsymbol{D}, \boldsymbol{V}, Y$) // Call HITON-PC
2: $TPC(Y) = TPC_1(Y)$; $nTPC_1(Y) = \boldsymbol{V} \setminus TPC_1(Y)$; $nTPC(Y) = \emptyset$
3: **for** $k = 2$ to k_{max} **do**
4: $nTPC(Y) = nTPC(Y) \cup nTPC_{k-1}(Y)$
5: Generate k^{th} *level* combined variable set $\boldsymbol{V'}$ based on $nTPC(Y)$
6: $\boldsymbol{V_k}$ = redundancyTest($\boldsymbol{V'}$)
7: Generate corresponding data set $\boldsymbol{D_k}$ for $\boldsymbol{V_k}$
8: Let *OPEN* contain all variables (in $\boldsymbol{V_k}$) associated with Y, sorted in descending order of strength of associations.
9: **repeat**
Phase I: Inclusion (line 10)
10: Move the first variable from *OPEN*, add it to the end of $TPC(Y)$ and $TPC_k(Y)$
Phase II: Elimination (lines 11-22)
// *Forward stage*
11: **if** *OPEN* $\neq \emptyset$ **then**
12: $X \leftarrow$ the variable last added to $TPC(Y)$
13: **if** $\exists \boldsymbol{S} \subset TPC(Y)$, s.t. $X \perp\!\!\!\perp Y \mid \boldsymbol{S}$ **then**
14: Remove X from $TPC(Y)$ and $TPC_k(Y)$
15: **end if**
// *Backward stage*
16: **else**
17: **for** each $X \in TPC_k(Y)$ **do**
18: **if** $\exists \boldsymbol{S} \subset TPC(Y) \setminus \{X\}$, s.t. $X \perp\!\!\!\perp Y \mid \boldsymbol{S}$ **then**
19: Remove X from $TPC(Y)$ and $TPC_k(Y)$
20: **end if**
21: **end for**
22: **end if**
23: **until** *OPEN* $= \emptyset$
24: $nTPC_k(Y) = \boldsymbol{V_k} \setminus TPC_k(Y)$
25: **end for**
26: Output $TPC(Y)$

maximum level of causal discovery).

In the forward stage (i.e. the case when $OPEN \neq \emptyset$), as with HITON-PC, MH-PC searches the current $TPC(Y)$ for a subset $\boldsymbol{S}$ to test whether the combined variable X is independent of Y given $\boldsymbol{S}$ (line 13). Since combined variables in $TPC(Y)$ are combinations of individual variables, MH-PC may have conducted some redundant conditional independence tests. For example, the CI test between X and Y given a combined variable ST (i.e. $X \perp\!\!\!\perp Y \mid ST$) may be unnecessary if the test given the two individual variables S and T (i.e. $X \perp\!\!\!\perp Y \mid \{S, T\}$) has been done. To address this issue, we propose a variant of MH-PC, called **MH-PC-B** (B version of MH-PC). To avoid confusion, we call the MH-PC algorithm shown in ALGORITHM 4.1, **MH-PC-F** (Full version of MH-PC). In the forward stage, when conducting the level k test with MH-PC-B, we do not include the level k variables in $TPC_k(Y)$ into conditioning sets, i.e. we replace line 13 in ALGORITHM 4.1 with the following statement:

$$\textbf{if } \exists \boldsymbol{S} \subset TPC(Y) \backslash TPC_k(Y),\ s.t.\ X \perp\!\!\!\perp Y \mid \boldsymbol{S}$$

As MH-PC-B conducts the tests conditioning only on the lower level variables, it can have higher efficiency than MH-PC-F, but at the same time it may produce some false positives. However, since in the backward stage (lines 17-21 of ALGORITHM 4.1), we do another check of the candidate causes remained in $TPC_k(Y)$, it is expected that the false discoveries are removed. As we will see from the next section, the experiments show that MH-PC-B is more efficient than MH-PC-F, while producing the same results as MH-PC-F with the data sets used.

4.2.4 False discoveries of multi-level HITON-PC

Since MH-PC-F (and MH-PC-B) follows the idea of HITON-PC, we firstly analyze the quality of HITON-PC in term of false discoveries. In HITON-PC, possible false decisions mainly come from two sources [13, 116]: the use of *max-k*, the maximum size of conditioning sets ($\boldsymbol{S}$, see 4.1) used for conditional independence tests, and incorrect results of statistical tests. Using a smaller *max-k* reduces the number of conditional independence tests, thus improves efficiency, but results in false positive discoveries. Fortunately, when *max-k* = 3 or 4, the false positive rate is not high, as shown in [14]. When we do not have enough number of samples, the statistical tests may produce incorrect results.

In the following, we will discuss the false discoveries coming from the interactions between variables. As mentioned above, the proposed method only focuses on non-redundant combined variables (Definition 4.2). While this strategy is used for reducing complexity, it may lead to false discoveries. However, we argue that our algorithms can still obtain high-quality causal findings. For non-cause variable A, if another non-cause variable B cannot improve the relationship between A and the target Y, then the combination AB, in most cases, may not improve the relationship between X and Y. This type of combined variables are unlikely to be combined causes. The experiment results in Section 4.3 have confirmed this intuition.

4.3 Experiments

We implemented MH-PC-F and MH-PC-B based on the semi-interleaved HITON-PC implementation in the R package, *bnlearn* [117]. In the experiments, the maximum level of combination (i.e. k_{max} in ALGORITHM 4.1)

Table 4.1: A Brief Description of Data Sets

Name	#Records	#Variables	Distributions	Ground Truth
Syn-7	1000	6	39.1% & 60.9%	2, 1
Syn-10	1000	9	42.2% & 57.8%	3, 2
Syn-12	2000	11	72.1% & 27.9%	4, 3
Syn-16	2000	15	45.2% & 54.8%	4, 3
Syn-20	2000	19	55.6% & 44.4%	4, 4
Syn-50	5000	49	30.8% & 69.2%	5, 5
Syn-60	5000	59	30.8% & 69.2%	5, 5
Syn-80	5000	79	30.5% & 69.5%	5, 5
Syn-100	5000	99	30.0% & 70.0%	5, 5
Syn-120	5000	119	30.4% & 69.6%	5, 5
CMC	1473	22	57.3% & 42.7%	-
German	1000	60	30.0% & 70.0%	-
House-votes-84	435	16	61.4% & 38.6%	-
Hypothyroid	3163	51	4.8% & 95.2%	-
Sick	2800	58	6.1% & 93.9%	-
Census	299285	495	6.2% & 93.8%	-

is restricted to 2, i.e. a combined cause at most consists of two component variables. We set the threshold of p-value to 0.01 to prune redundant combined variables and 0.05 to test causal relationships, for both synthetic and real world data sets.

4.3.1 Data sets

10 synthetic and 6 real world data sets were used in the experiments, and a summary of the data sets is shown in Table 4.1. The variables in all data sets are binary, i.e. each variable has two possible values, 1 or 0. The class variable in each data set is specified as the target variable. The numbers of variables shown in the table refer to the numbers of single predictor

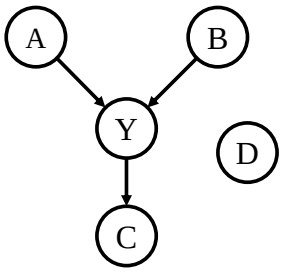
variables. The distribution of each data set indicates the percentages of the two different values of class variables. For synthetic data sets, the ground truth (i.e. the number of true causes) is shown in the table, where the first value is the number of single causes each consisting of one predictor variable and the second value is the number of combined causes each consisting of two predictor variables.

The first five synthetic data sets (with small number of variables) in Table 4.1 were generated with two main steps: (1) generating a data set based on a BN (Bayesian network) created randomly by the TETRAD software tool (http://www.phil.cmu.edu/tetrad/), and (2) generating the final synthetic data set by "splitting" some causes of the target into two new variables. Specifically, we firstly created a random BN using the TETRAD software, whose structure and conditional probability tables were both generated randomly. In the obtained BN, one of the variables was designated as the target and the others as predictor variables. Records of all variables were generated based on the conditional probability tables, using the built-in Bayes Instantiated Model.

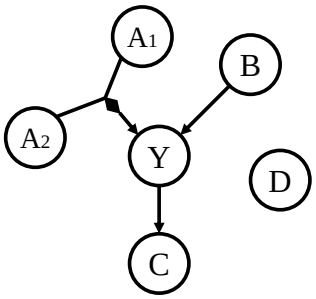
The following example is used to illustrate the process of the data generation.

Example 4.1 *A Bayesian network (shown in Figure 4.2(a)) is created randomly by the TETRAD software tool for the original variable set. Then we selected and split a parent node, e.g. A, of the target Y into two variables, e.g.* A_1 *and* A_2 *(see Figure 4.2(b)), such that (1)* $A_1 \wedge A_2 = A$ *(i.e.* A_1 *and* A_2 *both are equal to 1 if and only if A is 1), and (2)* A_1 *or* A_2 *is not an individual cause of the target.*

Note that, for combined causes in the synthetic data, we do not have a complete ground truth, since it may include some combined causes that we do not observe.



(a) The Bayesian network for the original variable set.



(b) The Bayesian network for the synthetic variable set.

Figure 4.2: A toy example to show the data generation (The solid arrows denote causal relationships and the diamond represents that the combination of two variables causes an outcome.).

For the next five larger synthetic data sets (Syn-50, ..., Syn-120), it is unpractical to generate them based on randomly drawn BNs, since it takes too long time to generate one. We firstly drew a simple BN where some variables were the parents of the target, some were not. Then we adopted logistic regression to generate the data based on the BN. Next, we employed the aforementioned splitting process to obtain the final data sets.

All real world data sets shown in Table 4.1 are obtained from the UCI Machine Learning Repository [118]. The first five real world data sets were employed to assess the effectiveness of the proposed algorithms, while the Census data set was used for evaluating the efficiency. The CMC (Contraceptive Method Choice) data set is an extraction of the National Indonesia Contraceptive Prevalence Survey in 1987. The German data set is a data set for classifying people's credit risks based on a set of attributes. House-

votes-84 contains the United States Congressional Voting Records in 1984. Hypothyroid and Sick are two medical data sets, which are from the Thyroid Disease data set of the repository (discretised using the MLC++ discretisation utility [119]). The Census data set is the Census Income (KDD) data set from the UCI Machine Learning Repository. In our experiments, all continuous attributes have been removed from the original data sets.

4.3.2 Performance evaluation

Three sets of experiments with the synthetic data were done to assess the accuracy of MH-PC-F and MH-PC-B by examining the results against the ground truth.

Firstly we compared MH-PC-F and MH-PC-B with two naïve approaches using HITON-PC and PC-Select [64] respectively (denoted as Naïve-H and Naïve-S in the following). PC-Select is an effective method for discovering the parents and children of a target variable, so we employed it as a benchmark for accuracy comparison.

Because the two naïve methods, especially Naïve-S, cannot handle large data sets, two small synthetic data sets (Syn-7 and Syn-10) were used in this set of experiments. Moreover, it is easier for small data sets to provide a good visualization of the detailed results.

The ground truth of the Syn-7 data set is that $V3$ and $V4$ are two single causes of the target and $V1\&V2$ is a combined cause (see the Ground truth column of Table 4.2, where Yes means the predictor variable is a cause of the target, and No means otherwise). In Table 4.2, MH-PC-F and MH-PC-B find exactly the ground truth in Syn-7. While Naïve-H identifies the ground truth, it includes a number of redundant results, for example, $V3\&V5$ and

Table 4.2: Comparison of the proposed algorithms with the naïve methods

Data set	Predictor variables	Ground truth	Naïve-H	Naïve-S	MH-PC-F	MH-PC-B
Syn-7	$V3$	Yes	√	√	√	√
	$V4$	Yes	√	√	√	√
	$V1\&V2$	Yes	√		√	√
	$V3\&V5$	No	√	√		
	$V4\&V5$	No	√	√		
Syn-10	$V5$	Yes	√	√	√	√
	$V6$	Yes	√	√	√	√
	$V7$	Yes				
	$V1\&V2$	Yes	√	√	√	√
	$V3\&V4$	Yes	√	√	√	√
	$V1\&V3$	No	√			
	$V4\&V5$	No	√			
	$V5\&V6$	No	√	√		
	$V5\&V9$	No		√		

$V4\&V5$ since $V3$ and $V4$ are causes already. Naïve-S misses the combined cause ($V1\&V2$) and it finds some redundant combined causes too. Similar results can be observed with the Syn-10 data set. MH-PC-F and MH-PC-B miss the true single cause, $V7$, and the naïve methods do not find it either.

Then we compared MH-PC-F and MH-PC-B with CR-CS [22] and CR-PA [20] using three synthetic data sets, Syn-12, Syn-16 and Syn-20. CR-CS and CR-PA are both designed to explore causal relationships from association rules, and they are also capable of finding both single and combined causes. The results are shown in Table 4.3, where P, R and F_1 represent the Precision, Recall and F_1-measure respectively. Here we used odds ratio greater than 1.5 as the threshold to indicate a significant result in both CR-CS and CR-PA. We can see that MH-PC-F and MH-PC-B both achieve higher accuracy than CR-CS and CR-PA, based on the known ground truth. Actu-

Table 4.3: Comparison of combined causes discovered by CR-CS, CR-PA, MH-PC-F and MH-PC-B with small synthetic data sets

		Syn-12	Syn-16	Syn-20
CR-CS	P	0.25	0.23	0.36
	R	1.00	1.00	1.00
	F_1	0.40	0.38	0.53
CR-PA	P	0.16	0.17	0.27
	R	1.00	1.00	1.00
	F_1	0.27	0.29	0.42
MH-PC-F	P	0.67	0.50	1.00
	R	0.67	1.00	1.00
	F_1	0.67	0.67	1.00
MH-PC-B	P	0.67	0.50	1.00
	R	0.67	1.00	1.00
	F_1	0.67	0.67	1.00

ally, CR-CS and CR-PA both perform very well in term of Recall, but they also include many false positives, since a main aim of these two methods is for explorations and they tolerate false positives and seek high recall.

In the next set of experiments, the last five larger synthetic data sets in Table 4.1 were used. From Table 4.4, all the four algorithms (i.e. CR-CS, CR-PA, MH-PC-F and MH-PC-B) can recover the ground truth very well from the data sets with relatively large sizes.

Based on the results of three sets of experiments, it is reasonable to conclude that MH-PC-F and MH-PC-B are capable to find single and combined causes. Another finding is that the causes (single and combined) identified by MH-PC-F and MH-PC-B are always the same, and this indicates two algorithms can achieve consistent results. This is also demonstrated by the results of two algorithms with all real world data sets, as described in the following.

Table 4.4: Comparison of combined causes discovered by CR-CS, CR-PA, MH-PC-F and MH-PC-B with larger synthetic data sets

		Syn-50	Syn-60	Syn-80	Syn-100	Syn-120
CR-CS	P	0.71	1.00	0.71	0.83	1.00
	R	1.00	1.00	1.00	1.00	1.00
	F_1	0.83	1.00	0.83	0.91	1.00
CR-PA	P	0.71	1.00	0.83	0.83	1.00
	R	1.00	1.00	1.00	1.00	1.00
	F_1	0.83	1.00	0.91	0.91	1.00
MH-PC-F	P	0.83	0.71	1.00	0.83	0.83
	R	1.00	1.00	1.00	1.00	1.00
	F_1	0.91	0.83	1.00	0.91	0.91
MH-PC-B	P	0.83	0.71	1.00	0.83	0.83
	R	1.00	1.00	1.00	1.00	1.00
	F_1	0.91	0.83	1.00	0.91	0.91

To investigate combined causes in the real world cases, we ran the proposed algorithms on the first six real world data sets in Table 4.1 for performance evaluation, where MH-PC-F and MH-PC-B still return consistent results as shown in Table 4.5. The proposed algorithms find many combined causes, and some of the combined causes discovered are reasonable as judged by common sense, shown in Table 4.6. For example, from the Sick data set it

Table 4.5: Number of (single and combined) causes discovered by MH-PC-F and MH-PC-B in real world data sets

	No. of single causes	No. of combined causes
CMC	2	0
German	1	13
House-votes-84	0	18
Hypothyroid	3	4
Sick	3	8

Table 4.6: Examples of combined causes identified from Sick and German data sets

Sick	T3 < 1.151 & TT4 < 87.5
	sick = true & T3 < 1.151
German	Checking.account = no-account & Savings.account < 100DM
	Property = real.estate & Other.installment.plans = none

is found that a low level of TT4 (Total T4) and T3 may result in thyroid disease (Table 4.6, where T4 and T3 are hormones produced by thyroid), and being sick and having a low level of T3 can lead to thyroid disease too. Some interesting combined causes are also discovered in the German data set. If one person has a private real estate and does not apply for any other installment plan, then this person is very likely to have a low default risk.

4.3.3 Efficiency and scalability

We ran Naïve-H, Naïve-S, MH-PC-F and MH-PC-B with various data sets on the same computer with a 3.4 GHz quad-core CPU and 16 GB of memory.

The running time of the algorithms on subsets of the Census data containing 30, 50, 70, 100 and 150 variables with the same sample size (50K) is shown in Figure 4.3. The two naïve methods are much slower than MH-PC-F and MH-PC-B, and Naïve-S is the most inefficient one. While the two naïve methods do not scale well with the number of variables, the two proposed algorithms both perform good scalability.

When applying the algorithms to the synthetic data sets containing different numbers of variables, both naïve methods do not return results after 5 hours. So no results of naïve methods are shown in Figure 4.4. From the figure, both proposed algorithms again scale well.

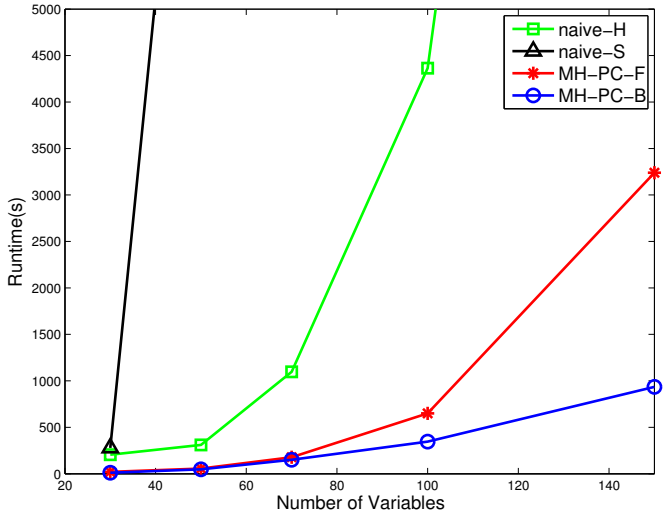


Figure 4.3: Scalability with number of variables - Census data

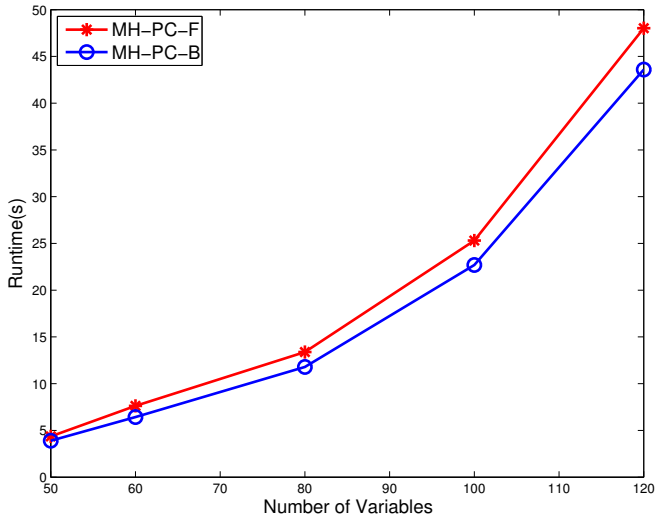


Figure 4.4: Scalability with number of variables - Synthetic data

We then ran the algorithms with 50K, 100K, 150K, 200K, and 250K samples respectively from the Census data set with 100 variables selected randomly, and the execution time of MH-PC-F and MH-PC-B is shown in Figure 4.5. No results are obtained for Naïve-S, and Naïve-H also cannot handle data sets with more than 50K samples. Similarly, MH-PC-B is more efficient and

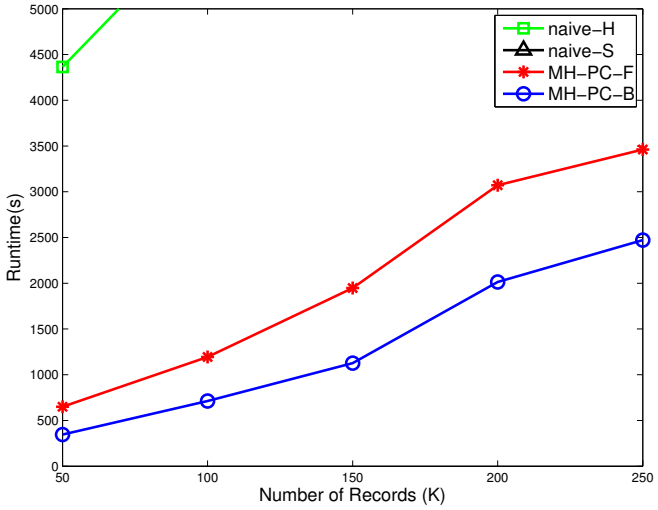


Figure 4.5: Scalability with umber of records - Census data

scalable than MH-PC-F.

To summarize, MH-PC-F and MH-PC-B are much faster than the naïve methods, and both proposed algorithms scale well in terms of the number of variables and number of samples. The experiments have also confirmed the discussions in Section 4.2.4 that MH-PC-B can achieve higher efficiency than MH-PC-F.

4.4 Conclusion

In practice, it is useful to identify a cause consisting of multiple variables, which individually are not causes of the target variable. However, finding such combined causes is challenging as the number of combined variables will increase exponentially with the increase of the number of individual variables. As far as we know, there has been very little work on discovering the combined causes, and the problem has not been studied in causal

Bayesian network research either.

In this chapter, we have proposed two efficient algorithms to mine the combined causes from large data sets. The proposed algorithms are based on a well-designed local causal discovery method, the semi-interleaved HITON-PC algorithm, with the novel extensions for dealing with combined causes. Experiments have shown that the proposed algorithms can find single and combined causes with a low number of false discoveries from synthetic data sets, and discover many reasonable combined causes from real world data. Additionally, the algorithms have been shown to scale up well with respective to the number of variables and the number of samples with both synthetic and real world data.

Chapter 5

Data-driven discovery of cause interactions

In Chapter 4, we have discussed the concept of combined causes and proposed a novel method to detect combined causes from observational data. The method can be considered as a step forward on understanding the complex causal mechanisms behind the phenomena. In this chapter, we further investigate the problem by detecting the interactions between multiple causal factors.

The identification of cause interactions themselves is indeed very useful, as the knowledge of cause interactions has many real world applications [29, 120]. For example, researchers are interested in the interaction between occupational exposure to asbestos dust and smoking status on lung cancer risk [28]. Many diseases result from not only genetic defects and environmental exposures respectively, but also the interactions between them [1]. For example, *dietary fat intake* is more strongly related to *obesity* in carriers of the $Pro12Ala$ *allele* [121]. So knowing such cause interactions is helpful for understanding and preventing diseases.

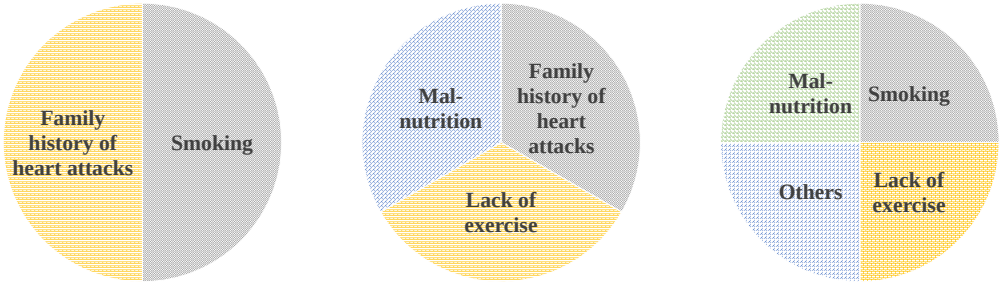


Figure 5.1: Examples of sufficient causes of an immature death.

Rothman et al. developed a theoretical model about causal mechanisms, the sufficient-component cause model [32, 74, 122]. Under this model, a causal mechanism, known as a sufficient cause, consists of a minimal set of interacting component causes such that the occurrence of all the component causes of the mechanism will inevitably lead to the occurrence of the outcome. By a minimal set it means that the causal mechanism will not take effect if any of its components does not occur. Furthermore, for the same outcome there may be multiple alternative causal mechanisms, i.e. it may have different sufficient causes. As illustrated in Figure 5.1, any of the causal mechanisms or sufficient causes (the pie charts in the figure) leads to an immature death and each of the mechanisms comprises a minimal set of components that interact with each other.

Although it is quite challenging to find complete causal mechanisms by using the sufficient-component cause model [74], it is a useful tool for investigating the interactions between the causal factors. According to the theorem in [74] and [32], there is an interaction between the components of a causal mechanism, called cause interactions hereafter, which in turn leads to produce the outcome. VanderWeele et al. [30, 31] extended the theorem to detect sufficient cause interactions under the sufficient-component cause model. However, the work largely stays at theoretical level, and there is no concrete computational methods for discovering the interactions between component causes, especially from large data sets.

Therefore, we utilize the theorem in [30, 32] as a theoretical basis to develop a true data mining method to discover cause interactions from observational data. We present a data mining framework and develop multiple instantiation algorithms of the framework to discover cause interactions from data. They are the first data mining algorithms for discovering cause interactions between causal factors.

5.1 Problem statement and definitions

In this section, we firstly introduce the notation used in this chapter, present the sufficient-component cause model, which is the theoretical basis of cause interactions discussed here, and then define the research problem studied in this chapter.

5.1.1 Notation

We use upper and lower case letters, e.g. X and x, to represent a random variable and its value, respectively. Bold-faced upper and lower case letters, e.g. $\boldsymbol{X}$ and $\boldsymbol{x}$, represent a set of variables and the corresponding values, respectively. We use the symbol "\" to denote the set difference operator, and we use the shorthand, e.g. $\boldsymbol{X}_{\setminus i}$ to represent $\boldsymbol{X}_{\setminus \{X_i\}}$. Particularly, we denote the set of predictor variables and the target variable with $\boldsymbol{V}$ and Y, respectively.

Note that $\boldsymbol{X} = \mathbf{1}$ represents all variables in $\boldsymbol{X}$ equal to 1 and $\boldsymbol{X}_{\setminus i} = \mathbf{1}$ represents that all variables in $\boldsymbol{X}$ except X_i are equal to 1. For simplicity, we also use "$ACE_{\{\boldsymbol{X}\}}$" to denote the average causal effect of predictor variables

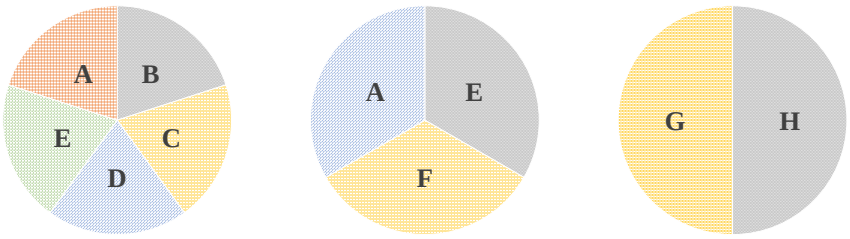


Figure 5.2: The sufficient-component cause model for the target Y, which has three sufficient causes.

$\boldsymbol{X}$ on the target Y, which is estimated by using the potential outcome model.

5.1.2 The sufficient-component cause model

The sufficient-component cause model [32] provides a natural and logical view of causation and the model focuses on the mechanisms that lead to the occurrence of an outcome.

A key concept of the sufficient-component cause model is the sufficient causes or causal mechanisms of a target of interest. As formally defined below, a sufficient cause comprises a minimal set of component causes whose occurrence inevitably results in the occurrence of the target.

Definition 5.1 (Sufficient Cause and Component Cause [32]) *Given a binary target variable Y and the population Ω, a set of binary variables $\boldsymbol{X} = \{X_1, \ldots, X_m\}$ is a sufficient cause of Y, if* (1) $Y_{\boldsymbol{X}=\boldsymbol{1}}(\omega) = 1$ *for all* $\omega \in \Omega$, *and* (2) $\nexists \boldsymbol{X_s} \subsetneq \boldsymbol{X}$ *s.t.* $Y_{\boldsymbol{X_s}=\boldsymbol{1}}(\omega) = 1$ *for all* $\omega \in \Omega$. *Each element* X_i $(1 \leq i \leq m)$ *is called a component cause of sufficient cause* $\boldsymbol{X}$.

Figure 5.2 is a schematic illustration of the sufficient causes of the target variable Y. Y has three causal mechanisms (or sufficient causes), i.e. the

three pie maps in the figure, and each member in a sufficient cause is a component cause. Each of the three sufficient causes can result in the occurrence of Y. For any one sufficient cause, the absence of its components makes it not sufficient any more to produce the effect. In other words, each pie map is sufficient to produce the outcome Y, while each component cause is necessary for the presence of the pie map. For example, the second pie map (or sufficient cause) AEF in Figure 5.2 is sufficient to generate Y, and each member, A, E and F, is necessary to make the sufficient cause AEF appear. Such a co-participation in a sufficient cause is defined as sufficient cause interaction between component causes, which is briefly named as cause interaction and its formal definition is expressed in the following.

Definition 5.2 (Cause Interaction [32]) *Let $\boldsymbol{X}$ be a set of binary predictor variables and Y be the binary target variable. The variable set $\boldsymbol{X_s}$ has a cause interaction on Y, if (1) $\boldsymbol{X_s} \subset \boldsymbol{X}$, and (2) $\boldsymbol{X}$ is a sufficient cause of Y.*

The presence of cause interactions between a set of variables implies that the overall effect of these variables is different from the sum of their individual effects. Cause interactions between multiple variables enable these variables to have stronger overall effects on the target. This proposition of cause interactions is expressed as follows.

Proposition 5.1 *If a set of variables have cause interactions, then the overall effect of these variables exceeds the additivity of their separate effects.*

Meanwhile, the following property of sufficient causes establishes a further connection between cause interactions and sufficient causes [12, 74].

Proposition 5.2 *If a set of variables have cause interactions, then these variables must participate in a same sufficient cause.*

For example, if E and F are detected to have a cause interaction on Y, then E and F are located in a same sufficient cause of Y (as shown by the top right pie chart in Figure 5.2).

5.1.3 Problem definition

Definitions 5.1 and 5.2 (originally given in [32]) initiate a theoretical formalization of cause interactions, which has laid the theoretical foundation for our work. However, the definitions are too strict to be applied to the real world cases [30]. In this chapter, we develop the practical definitions and criteria of cause interactions based on the sufficient-component cause model and the potential outcome framework. Then we take the advantages of data mining techniques to develop efficient algorithms to detect cause interactions from data.

Normally there exist two types of cause interactions [12]: (1) positive cause interactions, also named superadditive interactions in [123], which facilitate to produce the outcome, and (2) negative cause interactions, also named subadditive interactions in [123], which are to prevent the occurrence of the outcome. Both kinds of cause interactions are to strengthen the overall effect of multiple variables on the target. The only difference is that positive cause interactions increase the effects in the positive direction (i.e, producing the outcome), while negative cause interactions increase the effects in the negative direction (i.e. preventing the outcome).

For example, if doctors suggest us to take two or more drugs, it is expected that these drugs will have cause interactions and together help us to recover

from the disease more effectively, i.e. positively increase the effect of drugs on recovery. An example of negative cause interaction is that both high price and high maintenance cost have negative effects on the action of buying a car respectively, while the negative cause interaction between these two variables will strengthen their individual negative effects and thus the presence of both variables results in a lower chance of buying a car.

The following assumption of monotonic effect is commonly made in literatures [31, 124, 125].

Definition 5.3 (Monotonic Effect [30]) *Let* $\boldsymbol{X} = \{X_1, \ldots, X_m\}$ *and* Y *be a set of binary variables and a binary target, respectively. If we have* $E[Y|\boldsymbol{X} = \boldsymbol{x}] \geq E[Y|\boldsymbol{X} = \boldsymbol{x'}]$ *whenever* $\boldsymbol{x} \geq \boldsymbol{x'}$*, then* $\boldsymbol{X}$ *has positive monotonic effects on* Y*. If we have* $E[Y|\boldsymbol{X} = \boldsymbol{x}] \leq E[Y|\boldsymbol{X} = \boldsymbol{x'}]$ *whenever* $\boldsymbol{x} \geq \boldsymbol{x'}$*, then* $\boldsymbol{X}$ *has negative monotonic effects on* Y*.*

With the monotonic effect assumption, we develop a definition of cause interactions under the potential outcome framework. For the simplicity of the presentation, we only introduce the definition of positive cause interactions under the positive monotonic effect assumption and the corresponding criteria to detecting positive cause interactions. They can be easily adapted under the assumption of negative monotonic effects to define and detect negative cause interactions by changing only the direction of the effects of variables. In the experiments, we will show the results of the discovery of both positive and negative cause interactions from data.

Definition 5.4 (Positive Cause Interactions) *Let a variable set* $\boldsymbol{X} = \{X_1, \ldots, X_m\}$ *satisfy the assumption of positive monotonic effects. If* $\exists X_i \in \boldsymbol{X}$ *s.t.* $ACE_{\{X_i, \boldsymbol{X}_{\setminus i}=\boldsymbol{1}\}} - \sum_{j \neq i} ACE_{\{X_i, X_j=0, \boldsymbol{X}_{\setminus \{i,j\}}=\boldsymbol{1}\}} > 0$*, i.e. the causal effect of* X_i *with the presence (participation) of all remaining* $m-1$ *variables*

exceeds the sum of the causal effects of X_i with the presence of $m-2$ (i.e. 1 less) variables, then $\boldsymbol{X}$ has m-way positive cause interactions.

In Definition 5.4, we use $ACE_{\{X_i,\boldsymbol{X}_{\setminus i}=\mathbf{1}\}}$ to denote the average causal effect of X_i with all variables in $\boldsymbol{X}$ except X_i set to 1, and $ACE_{\{X_i,X_j=0,\boldsymbol{X}_{\setminus\{i,j\}}=\mathbf{1}\}}$ represents the average causal effect of X_i with $X_j=0$ and $\boldsymbol{X}_{\setminus\{i,j\}}=\mathbf{1}$.

The research problem in this chapter is to detect cause interactions from data, which is showed as follows.

Problem 5.1 *Given an observational data set $\boldsymbol{D}$ for the predictor variables $\boldsymbol{V}$ and the target Y, find the variable sets $\boldsymbol{X}\subset\boldsymbol{V}$, s.t. $\boldsymbol{X}$ has cause interactions with respect to the target Y by Definition 5.4.*

5.2 Cause interactions

In this section, we illustrate the main idea to detect the 2-way cause interactions and then present the criterion for detecting interactions among m variables (m-way cause interactions).

5.2.1 2-way cause interactions

Taking the study of the causes of lung cancer as an example, let $X_1=1$ or 0 denotes *occupational exposure to asbestos* or not, $X_2=1$ or 0 denotes *smoking* or not and $Y_{x_1x_2}$ represents *suffering the lung cancer* or not when $X_1=x_1$ and $X_2=x_2$, $x_1,x_2\in\{0,1\}$. Then the average causal effect of X_1 when X_2 is present ($X_2=1$) and not present ($X_2=0$) can be

respectively estimated as $ACE_{\{X_1 1\}} = E[Y_{11}] - E[Y_{01}]$ and $ACE_{\{X_1 0\}} = E[Y_{10}] - E[Y_{00}]$, where $ACE_{\{X_1 x_2\}}$ denotes the ACE of X_1 when $X_2 = x_2$. If we have $ACE_{\{X_1 1\}} > ACE_{\{X_1 0\}}$, i.e. the causal effect of X_1 with X_2's presence is larger than the causal effect of X_1 without X_2's presence, then based on Definition 5.4, we can conclude that there is a positive cause interaction between X_1 (*exposure to asbestos*) and X_2 (*smoking*) on Y (*lung cancer*). Note that the condition also can be presented as $ACE_{\{1X_2\}} > ACE_{\{0X_2\}}$.

Criterion 5.1 *Let X_1 and X_2 have positive monotonic effects on the target Y. There is a positive cause interaction between X_1 and X_2, if any of the following conditions is met*

$$ACE_{\{X_1 1\}} - ACE_{\{X_1 0\}} > 0 \tag{5.1}$$

$$ACE_{\{1X_2\}} - ACE_{\{0X_2\}} > 0 \tag{5.2}$$

where $ACE_{\{X_1 x_2\}}$ represents the causal effect of X_1 on Y when $X_2 = x_2$ and $ACE_{\{x_1 X_2\}}$ the causal effect of X_2 on Y when $X_1 = x_1$.

Note that these two conditions (Equations 5.1 and 5.2) are identical, if the ACE in the above conditions are expanded by risk difference.

Criterion 5.1 presents the conditions for detecting 2-way cause interactions, where we suppose there are no confounders, i.e. no other variables influencing the causal effects of X_1 and X_2 on the target Y. However, as mention in Section 3.2, randomized assignment of treatments is not possible in observational studies, so we employ the stratification strategy to address the confounding issue, which stratifies the data into a number of strata. Then we estimate the average causal effects in each stratum and aggregate them over all the strata. So we have the following criterion for detecting the 2-way positive cause interaction.

Criterion 5.2 *Let X_1 and X_2 have positive monotonic effects on the target Y and $\boldsymbol{C}$ be the covariates. There is a positive cause interaction between X_1 and X_2, if any of the following conditions is met*

$$\sum_{\boldsymbol{C}=\boldsymbol{c}} w_{\boldsymbol{c}}(ACE_{\{X_1 1|\boldsymbol{c}\}} - ACE_{\{X_1 0|\boldsymbol{c}\}}) > \eta \tag{5.3}$$

$$\sum_{\boldsymbol{C}=\boldsymbol{c}} w_{\boldsymbol{c}}(ACE_{\{1X_2|\boldsymbol{c}\}} - ACE_{\{0X_2|\boldsymbol{c}\}}) > \eta \tag{5.4}$$

where $w_{\boldsymbol{c}}$ is the weight of stratum $\boldsymbol{C} = \boldsymbol{c}$ and η is a threshold specified by users.

5.2.2 m-way cause interactions

Similar to 2-way cause interactions, we can derive the criterion for detecting m-way cause interactions from data. However, when $m > 2$, the situation becomes more complicated, since there exist more ways for selecting variables X_i and X_j (see Definition 5.4). In the following, we discuss how to do the variable selection for the m-way cause interactions. We firstly look at the case of 3-way (i.e. $m = 3$) cause interactions.

Given three binary variables, X_1, X_2 and X_3, if we focus on X_1, and consider the causal effect of X_1 when both X_2 and X_3 are present, in comparison to the causal effects when either X_2 or X_3 is absent, based on Definition 5.4, we have: if $ACE_{\{X_1 11\}} - (ACE_{\{X_1 01\}} + ACE_{\{X_1 10\}}) > 0$, the three variables have a positive cause interaction. Similarly when we focus on X_2 and X_3 respectively, we obtain the other two conditions for 3-way cause interactions: $ACE_{\{1X_2 1\}} - (ACE_{\{0X_2 1\}} + ACE_{\{1X_2 0\}}) > 0$, and $ACE_{\{11X_3\}} - (ACE_{\{01X_3\}} + ACE_{\{10X_3\}}) > 0$. If any of these three conditions or criteria holds, we say that X_1, X_2, and X_3 have 3-way positive cause interactions with respect to the target Y.

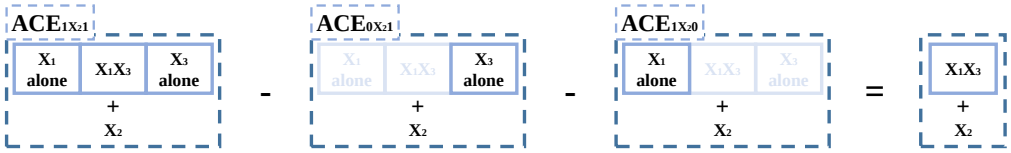


Figure 5.3: An illustration of 3-way cause interactions.

To understand the physical meaning of the conditions for 3-way cause interactions, as illustrated in Figure 5.3 (when focusing on X_2), we can consider that the overall causal effect of X_2 with the participation of X_1 and X_3 ($ACE_{\{1X_21\}}$) on Y in fact attributes to three elements: the causal effect of X_2 with the presence of X_1's contribution, the causal effect of X_2 with the presence of X_3's contribution and the causal effect of X_2 with the presence of the combined contribution of X_1 and X_3. The 3-way cause interaction is indeed the third element, the causal effect of X_2 with the presence of the combined contribution of X_1 and X_3. Therefore, we take away $ACE_{\{1X_20\}}$ and $ACE_{\{0X_21\}}$ from the overall causal effect of X_2 with the participation of X_1 and X_3 ($ACE_{\{1X_21\}}$) to get the causal effect of the 3-way cause interaction.

Now let us develop the criterion for detecting m-way cause interactions. Firstly we need to introduce the concept of a subordinate set [30], which enables us use the ACE concept and notation to present the causal effects of $m-1$ variables on Y in the criterion to be developed (Criterion 5.3).

Definition 5.5 (Subordinate Set) *Given a set of binary variables* $\boldsymbol{X} = \{X_1, \ldots, X_m\}$, *let* $\boldsymbol{u}_m^i = \{(x_1, \ldots, x_m) \in \{0,1\}^m : \boldsymbol{x}_{\setminus i} = \boldsymbol{1}, x_i = 0\}$ *and* $\boldsymbol{u}_m^{i,j_i} = \{(x_1, \ldots, x_m) \in \{0,1\}^m : \boldsymbol{x}_{\setminus\{i,j_i\}} = \boldsymbol{1}, x_i = x_{j_i} = 0, i \neq j_i\}$. *The set* $\mathcal{S}_k = \{\boldsymbol{u}_m^{1,j_1}, \ldots, \boldsymbol{u}_m^{k-1,j_{k-1}}, \boldsymbol{u}_m^{k+1,j_{k+1}}, \ldots, \boldsymbol{u}_m^{m,j_m}\}$ *is a subordinate set of order* m, *if* $\exists l \in \{1, \ldots, k-1, k+1, \ldots, m\}$, *s.t.* $\boldsymbol{u}_m^{l,j_l} \leq \boldsymbol{u}_m^k$.

In Definition 5.5, $\boldsymbol{u}_m^i$ ($\boldsymbol{u}_m^{i,j_i}$) represents an m-length vector with i-th (i-th

and j_i-th) element(s) equal to 0 and the remaining equal to 1. Now we use the following example to explain the subordinate set and the detection of the 3-way cause interaction.

Example 5.1 *For the estimation of 3-way cause interactions ($m = 3$), if we focus on the causal effect of X_2 (i.e. $k = 2$) on Y, then $\mathcal{S}_2 = \{\boldsymbol{u}_3^{1,j_1}, \boldsymbol{u}_3^{3,j_3}\}$, which has three possible combinations based on Definition 5.5: $\{\boldsymbol{u}_3^{1,3}, \boldsymbol{u}_3^{3,2}\}$, $\{\boldsymbol{u}_3^{1,2}, \boldsymbol{u}_3^{3,2}\}$ and $\{\boldsymbol{u}_3^{1,2}, \boldsymbol{u}_3^{3,1}\}$. Any one of three combinations can be selected for detecting 3-way cause interactions, e.g. $\mathcal{S}_2 = \{\boldsymbol{u}_3^{1,2}, \boldsymbol{u}_3^{3,2}\}$ (i.e. $i = 1$, $j_i = 2$ and $i = 3$, $j_i = 2$), and thus a condition of 3-way cause interaction is expressed as $ACE_{\{1X_21\}} - (ACE_{\{0X_21\}} + ACE_{\{1X_20\}}) > 0$.*

As illustrated in the above example, the subordinate set $\mathcal{S}_k$ normally is not unique. The different conditions can be created based on different subordinate sets and the cause interactions will exist, once any one condition is satisfied [30].

The formal criterion for detecting an m-way cause interaction is showed in the following, with the confounding taken into account.

Criterion 5.3 *$\boldsymbol{X} = \{X_1, \ldots, X_m\}$ satisfies the positive monotonic effect assumption, and $\boldsymbol{C}$ be the covariate set. The variable set $\boldsymbol{X}$ exhibits an m-way positive cause interaction, if $\exists X_k \in \boldsymbol{X}$ and some subordinate sets $\mathcal{S}_k$, s.t.*

$$\sum_{\boldsymbol{C}=\boldsymbol{c}} w_{\boldsymbol{c}}(ACE_{\{X_k, \boldsymbol{X}_{\setminus k}=1|\boldsymbol{c}\}} - \sum_{\{i,j_i\}:\boldsymbol{u}_m^{i,j_i} \in \mathcal{S}_k} ACE_{\{X_{j_i}, X_i=0, \boldsymbol{X}_{\setminus\{i,j_i\}}=\boldsymbol{1}|\boldsymbol{c}\}}) > \eta,$$

where η is the threshold and $w_{\boldsymbol{c}}$ is the weight of the stratum $\boldsymbol{C} = \boldsymbol{c}$.

5.2.3 Algorithm

In this section, we present the proposed framework, Data-driven Approach to Cause Interaction Discovery (DACID), and the specific algorithms instantiated from the framework for discovering cause interactions in observational data. As shown in ALGORITHM 5.1, the framework includes three main steps: (1) candidate variable Generation (lines 1-4), to only include the variables associated with the given target and to generate candidate sets for testing cause interactions; (2) data stratification (lines 10-11), to balance observed covariates between control and treatment groups to reduce bias; and (3) cause interaction discovery (lines 12-22), to detect the cause interactions between multiple variables with respect to the target. In each step, different data mining and statistical algorithms can be involved. Meanwhile a pruning schema is employed to improve the efficiency of the algorithms.

5.2.3.1 Candidate variable generation

Based on Proposition 5.2, if a set of variables $\boldsymbol{X}$ has a cause interaction with respect to a target, Y, then every variable X in $\boldsymbol{X}$ is a component cause of one specific sufficient cause of Y. Thus it is reasonable to assume that every variable in $\boldsymbol{X}$ is associated with Y. In our framework, only variables associated with Y will be considered to generate the candidate sets for testing cause interactions. In ALGORITHM 5.1, the function *predictorSelection*() in line 1 is used for the associated variable selection. Various methods can be implemented for this function as introduced in the following.

Correlation is a commonly used criterion to describe the association between two variables. Pearson correlation, Chi-square test, Student's t-test and etc. can be employed to estimate the correlation between variables.

ALGORITHM 5.1 Data-driven Approach to Cause Interaction Discovery (DACID)

Input: A binary data set $\boldsymbol{D}$ for predictor variable set $\boldsymbol{V}$ and the target Y, the significant threshold η for testing cause interactions, and the maximum level of cause interactions k_0.

Output: $\boldsymbol{CI}_Y = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q\}$, where $\boldsymbol{X}_k \subset \boldsymbol{V}$ is a set of variables with cause interactions w.r.t Y.

$\boldsymbol{V'} \leftarrow$ *predictorSelection*$(\boldsymbol{V}, Y)$
$\boldsymbol{CI}_Y \leftarrow \emptyset$
Let $m = 2$
Pairwise generate the 2-way candidate set $\boldsymbol{V}_2$ based on $\boldsymbol{V'}$
while $m \leq m_0$ **do**
  if $m + 1 \leq m_0$ **then**
    $(m+1)$-way candidate set $\boldsymbol{V}_{m+1} \leftarrow \emptyset$
  end if
  for each $\boldsymbol{X}_m$ in $\boldsymbol{V}_m$ **do**
    $\boldsymbol{C} \leftarrow$ *stratifyVariable*$(\boldsymbol{V}, \boldsymbol{X}_m, Y)$
    stratification$(\boldsymbol{X}_m, \boldsymbol{D}, \boldsymbol{C})$
    for each stratum $\boldsymbol{C} = \boldsymbol{c}$ **do**
      $CIValue_{\boldsymbol{c}} \leftarrow$ *causeInteraction*$(\boldsymbol{X}_m, Y, \boldsymbol{c})$
      Get $w_{\boldsymbol{c}}$ from $\boldsymbol{c}$
    end for
    $CIValue \leftarrow \sum w_{\boldsymbol{c}} CIValue_{\boldsymbol{c}}$
    if $CIValue > \eta$ **then**
      $\boldsymbol{CI}_Y \leftarrow \boldsymbol{CI}_Y \cup \boldsymbol{X}_m$
      if $m + 1 \leq m_0$ **then**
        for each $X \in \boldsymbol{V'} \backslash \boldsymbol{X}_m$, $\boldsymbol{V}_{m+1} \leftarrow \boldsymbol{V}_{m+1} \cup \{\boldsymbol{X}_m, X\}$
      end if
    end if
  end for
  $m = m + 1$
end while
Output $\boldsymbol{CI}_Y$

Association rule mining is another way to identify statistical dependence between predictors and a target. Measures, such as support and odds ratio [126], are used to pick up variables associated with the target.

These methods for association analysis primarily vary in term of the types of variables (categorical or continuous variables). Under our framework, one may use any of the association analysis methods suitable to their data types to select the associated variables.

Once the associated variables selected, we pairwise generate the 2-way candidate set $\boldsymbol{V}_2$ (line 4) for testing cause interactions in the next sections.

5.2.3.2 Stratification

Stratification attempts to balance observed covariates by obtaining similar covariate distributions between treatment and control groups to reduce estimation bias. As described in Section 3.2, ideally, within a stratum where the distributions of covariates (predictor variables other than the ones under consideration of cause interactions) of the control group and the treatment group are indistinguishable, we can use risk difference to obtain an unbiased estimation of the causal effects, and then aggregate the causal effects across all the strata. To this end, the key to proper stratification of a data set is the choice of stratifying variables from the set of predictor variables. Using too few stratifying variables may not effectively avoid confounding to reduce bias, while having too many stratifying variables will lead to small sized strata thus to lose the statistical power of the tests.

In the DACID framework, we do not conduct stratification based on all covariates, instead the covariates that are associated with the target are selected for the stratification (represented by stratifying variables $\boldsymbol{C}$). This

is done by function *stratifyVariable*() in ALGORITHM 5.1 (line 10). For example, We do not control one person's facial features (e.g. a hawk nose), which may never impact the lung cancer risk, when estimating the cause interactions of occupational exposure to asbestos and the smoking status, regarding lung cancer.

Once the stratifying variable set is selected, either perfect stratification or approximate stratification on $\boldsymbol{C}$ can be employed. The simplest stratification is the Perfect Stratification (PS). PS requires that all individuals within a stratum have exactly the same values for the stratifying variables $\boldsymbol{C}$, to remove the effects of stratifying variables on the target Y. Thus PS is capable of eliminating bias in the estimations of causal effects and cause interactions. In the proposed DACID framework, PS is employed as an option when performing data stratification.

However, for a high-dimensional data set, PS may have a low statistical power for detecting dependency in data, as too many strata are generated and each stratum has a small size. An alternative option, Stratification on Generalized Propensity Score (SGPS), is provided to increase the statistical power. The basic idea of SGPS is to stratify individuals to different strata, such that individuals in the same stratum have similar propensity scores. The generalized propensity score [108] was developed to deal with the case with multiple treatments, based on the propensity score [33]. Please refer to Chapter 3 for detailed introduction about the generalized propensity score. Specifically we use multinomial logistic regression to obtain the generalized propensity score vector [109].

In the proposed framework, Stratification on Generalized Propensity Score (SGPS) is employed as an alternative option of (PS). Specifically, K-means clustering is adopted to separate individuals with similar generalized propensity score vectors into the same stratum [127].

5.2.3.3 Cause interaction discovery

The stratification puts individuals with similar distribution of stratifying variables $\boldsymbol{C}$ in the same stratum. Within each stratum $\boldsymbol{C} = \boldsymbol{c}$, a contingency table is generated for the estimation of cause interactions. For simplicity, in the following, we discuss how to discover 2-way cause interactions.

Given a data set $\boldsymbol{D}$, X_1 and X_2 are binary predictors of the binary target Y. Each stratum $\boldsymbol{C} = \boldsymbol{c}$ of data $\boldsymbol{D}$ is divided into 4 treatment groups based on different values of X_1 and X_2, as in the contingency table below: *treatment* 1 ($\{X_1 = 1, X_2 = 1\}$), *treatment* 2 ($\{X_1 = 1, X_2 = 0\}$), *treatment* 3 ($\{X_1 = 0, X_2 = 1\}$) and *treatment* 4 ($\{X_1 = 0, X_2 = 0\}$), where n_{ij} ($i \in \{1, 2, 3, 4\}, j \in 1, 2$) denotes the frequencies of the different values of X_1, X_2 and Y.

		Y	
X_1	X_2	1	0
1	1	n_{11}	n_{12}
1	0	n_{21}	n_{22}
0	1	n_{31}	n_{32}
0	0	n_{41}	n_{42}

Since individuals are indistinguishable in terms of the stratifying variables in the same stratum, no matter which treatment group the individuals are in, we can unbiasedly estimate the causal effects between any two treatment groups in the stratum.

We use the following example to explain how to estimate the causal effects and cause interactions. As an illustration, we show how to calculate the causal effect of X_2 in the cases of $X_1 = 1$ and $X_1 = 0$ respectively, and then based on the causal effects to determine if there is a cause interaction between X_1 and X_2.

Table 5.1: Two 2×2 contingency tables

		Y	
X_1	X_2	1	0
1	1	n_{11}	n_{12}
1	0	n_{21}	n_{22}

		Y	
X_1	X_2	1	0
0	1	n_{31}	n_{32}
0	0	n_{41}	n_{42}

Example 5.2 *For clear presentation, we break the above contingency table into two* 2×2 *contingency tables, shown in Table 5.1. Then the causal effect of* X_2 *in the case of* $X_1 = 1$ *can be estimated by comparing the individuals receiving* treatment *1 (*$\{X_1 = 1, X_2 = 1\}$*) and* treatment *2 (*$\{X_1 = 1, X_2 = 0\}$*), see the first contingency table in Table 5.1. Similarly, ACE between* treatment *3 group and* treatment *4 group is estimated from the second contingency table, to get the causal effect of* X_2 *when* $X_1 = 0$*. Probabilities (Prob) are the most common approach to calculating the ACEs [31].*

$$ACE_{\{1X_2|\boldsymbol{c}\}} = n_{11}/(n_{11} + n_{12}) - n_{21}/(n_{21} + n_{22})$$
$$ACE_{\{0X_2|\boldsymbol{c}\}} = n_{31}/(n_{31} + n_{32}) - n_{41}/(n_{41} + n_{42})$$

Odds Ratio (OR) [128] is another widely used measure and may be more suitable to determine the ACEs when the response variable is discrete [129], especially binary.

$$ACE_{\{1X_2|\boldsymbol{c}\}} = n_{11}n_{22}/n_{12}n_{21}$$
$$ACE_{\{0X_2|\boldsymbol{c}\}} = n_{31}n_{42}/n_{32}n_{41}$$

With either Prob or OR, the causal effects of X_2 *with* $X_1 = 1$ *and* $X_1 = 0$ *on the stratum* $\boldsymbol{C} = \boldsymbol{c}$ *are obtained. Then the cause interactions between* X_1 *and* X_2 *on* Y *can be determined based on Criterion 5.2, by aggregating ACEs in all strata. Here the weight* $w_{\boldsymbol{c}}$ *of one stratum* $\boldsymbol{c}$ *is set as the ratio of the sample size of* $\boldsymbol{c}$ *to the size of data* $\boldsymbol{D}$*.*

Table 5.2: The difference between the four algorithms.

Algorithms	Stratification	ACE Estimation
PS-OR	PS	OR
PS-Prob	PS	Prob
SGPS-OR	SGPS	OR
SGPS-Prob	SGPS	Prob

The above is an example of 2-way cause interaction estimation. m-way cause interactions can be obtained with the similar progress.

Probability (Prob) and Odds Ratio (OR) are employed here for the estimation of causal effects. Combined with two different stratification strategies, Perfect Stratification (PS) and Stratification on Generalized Propensity Score (SGPS), the proposed DACID framework is instantiated into four different algorithms. The difference between four different algorithms is shown in Table 5.2.

5.2.3.4 Pruning schema and complexity analysis

As mentioned in Definition 5.2, variables in the same sufficient cause have cause interactions on the target. Therefore, high level cause interactions and low level cause interactions always co-occur in a sufficient cause. For example, if AEF is a sufficient cause of the target Y, then 2-way cause interactions of AE, AF and EF and the 3-way cause interaction of AEF always coincide with each other.

Therefore, if a set of k variables do not have cause interaction, any of its superset of variables do not have cause interaction either. Using this property, we can prune the search space by generating high level cause interaction candidates based on confirmed lower level interactions, as shown in

line 19 of ALGORITHM 5.1. Thus the DACID framework identifies cause interactions in a level by level manner.

Now we analyze the time complexity of the proposed algorithms. In the case without the pruning schema, the number of all possible cause interaction candidate sets is $\sum_{k=2}^{|\boldsymbol{V}|} \binom{|\boldsymbol{V}|}{k}$, and the number of cause interaction tests is $O(2^{|\boldsymbol{V}|})$. For the case with the pruning schema, if the number of all k-way cause interaction candidate sets is $N(k)$ and the number of sets confirmed is $N'(k)$ ($N'(k) \ll N(k)$), then the number of $(k+1)$-way cause interaction candidate sets is $N(k+1) = N'(k)(|\boldsymbol{V}| - k)$. The number of cause interaction candidate sets is $\binom{|\boldsymbol{V}|}{2} + \sum_{k=2}^{|\boldsymbol{V}|-1} N'(k)(|\boldsymbol{V}| - k)$. Thus in the case with pruning schema, the number of cause interaction tests is $O(|\boldsymbol{V}|^2)$. Therefore, the pruning schema significantly improves the efficiency of cause interaction discovery.

This pruning schema may bring about some false negatives, when the cause interaction discovery approach misses the lower level cause interactions, since the higher level (e.g. $(k+1)$-way) candidates are generated from the lower level (e.g. k-way) cause interactions. Fortunately, for a $(k+1)$-way candidate, if any k-way cause interaction found is a subset of this $(k+1)$-way candidate, then this candidate will not be missed. For example, when we missed the cause interaction between X_2 and X_3, the candidate $X_1X_2X_3$ still could be generated if we detected the cause interaction between X_1 and X_2. Thus the cause interaction discovery approach can obtain high-quality cause interaction discovery.

5.3 Experiments

In this section, a collection of synthetic and frequently used real world data sets are used to evaluate the effectiveness and efficiency of four instantiated

Table 5.3: A summary of synthetic data sets

Name	#Records	#Variables	Distribution	#GT
V50-5K	5000	49	20.0% & 80.0%	5
V80-5K	5000	79	20.0% & 80.0%	5
V100-5K	5000	99	20.0% & 80.0%	5
V120-5K	5000	119	20.0% & 80.0%	5
V150-5K	5000	149	20.0% & 80.0%	5
V200-5K	5000	199	20.0% & 80.0%	5
V6-2K	2000	5	20.0% & 80.0%	1
V10-2K	2000	9	20.0% & 80.0%	2
V20-2K	2000	19	20.0% & 80.0%	3
V30-2K	2000	29	20.0% & 80.0%	4

algorithms of the proposed framework: PS-OR, PS-Prob, SGPS-OR, and SGPS-Prob. For synthetic data, we set the threshold for detecting cause interactions to 0.10 for all four algorithms (i.e. $\eta = 0.10$ in Criterion 5.3). For the real world data sets, we use 0.10 as the threshold to pick up cause interactions, and then results with strong cause interactions are extracted for detailed discussions.

5.3.1 Synthetic data

Table 5.3 summarizes the synthetic data sets used for the accuracy evaluation, each of which in fact refers to a group of 100 data sets with the same numbers of variables and samples. The performance results presented in this section is the average of the results of the 100 data sets. All data sets in the table are binary and the distribution of each data set indicates the percentages of the two different values of the target variable. The #GT (Ground Truth) column in the table shows the numbers of sufficient causes contained in the synthetic data sets. In Table 5.3, the first six synthetic data sets contain five sufficient causes, each composing of two component

Table 5.4: Cause interactions discovered by PS-OR, PS-Prob, SGPS-OR and SGPS-Prob from sufficient causes, each of which composes of two component causes.

		PS			SGPS		
		P	R	F_1	P	R	F_1
V50-5K	OR	1.00	1.00	1.00	1.00	1.00	1.00
	Prob	1.00	1.00	1.00	0.99	1.00	1.00
V80-5K	OR	1.00	1.00	1.00	1.00	1.00	1.00
	Prob	1.00	1.00	1.00	0.99	1.00	0.99
V100-5K	OR	1.00	1.00	1.00	1.00	1.00	1.00
	Prob	1.00	1.00	1.00	0.99	1.00	1.00
V120-5K	OR	1.00	1.00	1.00	1.00	1.00	1.00
	Prob	1.00	1.00	1.00	0.99	1.00	1.00
V150-5K	OR	1.00	1.00	1.00	1.00	1.00	1.00
	Prob	1.00	1.00	1.00	0.99	1.00	0.99
V200-5K	OR	1.00	1.00	1.00	1.00	1.00	1.00
	Prob	1.00	1.00	1.00	0.99	1.00	0.99

causes, and the next four synthetic data contains one or more sufficient causes, each composing of three component causes.

In order to obtain these synthetic data, three main steps are performed: (1) samples of the predictor variables $\boldsymbol{V}$ and the target Y are randomly generated, (2) a set of predictor variables $\boldsymbol{X}$ are randomly picked up from $\boldsymbol{V}$ and values of these variables $\boldsymbol{X}$ are modified such that their values (and values of Y) satisfy the definition of sufficient causes, i.e. let $\boldsymbol{X}$ be a sufficient cause of Y, and (3) repeat steps 2 and 3 to obtain another sufficient causes (and their samples) of Y. For example, a variable set $\boldsymbol{X} = \{X_1, X_2\}$ is picked up, then we modify the samples of X_1 and X_2 such that for the whole data set we have $Y = 1$ once $X_1 = 1$ and $X_2 = 1$. In this way, $\boldsymbol{X}$ is now a sufficient cause of Y, based on the definition of sufficient causes.

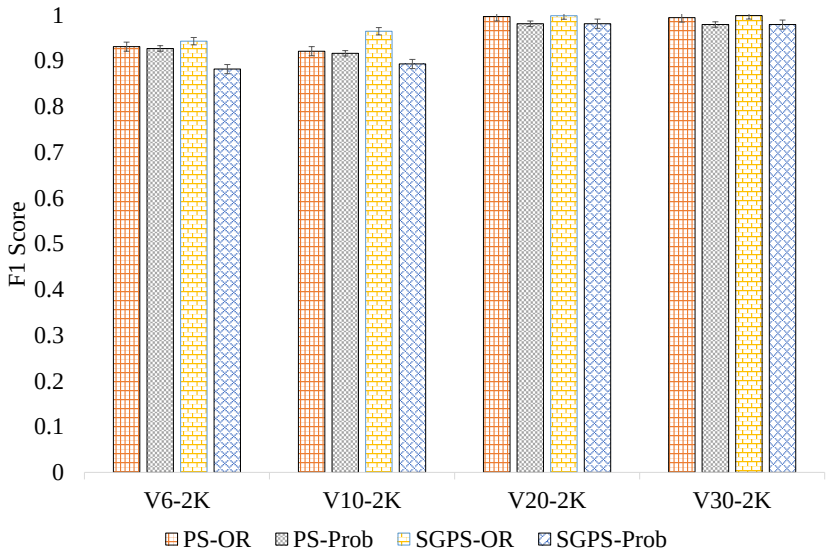
Firstly, we do the experiments on the first six groups of synthetic data sets

(in Table 5.3). Table 5.4 shows the performance of the proposed algorithms, where P, R and F_1 represent the Precision, Recall and F_1-measure respectively. We can that see all of the four algorithms perform very well on these data sets. The reason is that if two variables are sufficient to cause the occurrence of the target, then these two variables have a strong cause interaction on the target, and thus this type of cause interactions are easier to be discovered from the data.

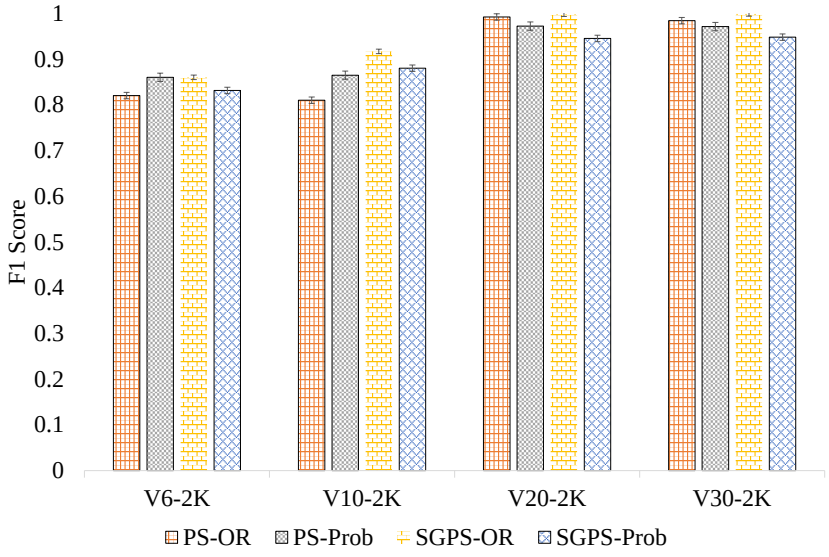
We also employ the last four groups of synthetic data sets in Table 5.3 to evaluate the proposed algorithms' abilities of discovering cause interactions between a sufficient cause consisting of three component causes. As mentioned previously (in Section 5.1), within such a sufficient cause, there exist 2-way and 3-way cause interactions. For example, $X_1X_2X_3$ is a known sufficient cause, then X_1X_2, X_1X_3 and X_2X_3 will show 2-way cause interactions respectively and $X_1X_2X_3$ has a 3-way cause interaction. The results of 2-way and 3-way cause interactions will be discussed in the following, respectively.

The ground truth of 2-way cause interactions consists of subsets of sufficient causes, e.g. X_1X_2, X_1X_3 and X_2X_3 in the above example. Figure 5.4a shows the F_1 score of the discoveries of the 2-way cause interactions by the four algorithms. Results show that the proposed algorithms achieve high accuracy on detecting the 2-way cause interactions.

In 3-way cause interaction discovery, the ground truth is the known sufficient causes built in synthetic data sets. The F_1 scores of the discoveries of the 3-way cause interactions are shown in Figure 5.4b. In our experiments, the threshold for detecting 3-way cause interactions is set as the maximum *CIValue* (line 16 in ALGORITHM 5.1) of the 2-way cause interactions between subsets of current candidates. To explain the rational of this setting, within a sufficient cause, a superset always has stronger a cause interaction than the ones its subsets have, since a superset has a higher chance to



(a) 2-way cause interactions



(b) 3-way cause interactions

Figure 5.4: Cause interactions found from sufficient causes, consisting of three component causes.

Table 5.5: Description of the Car Evaluation data set.

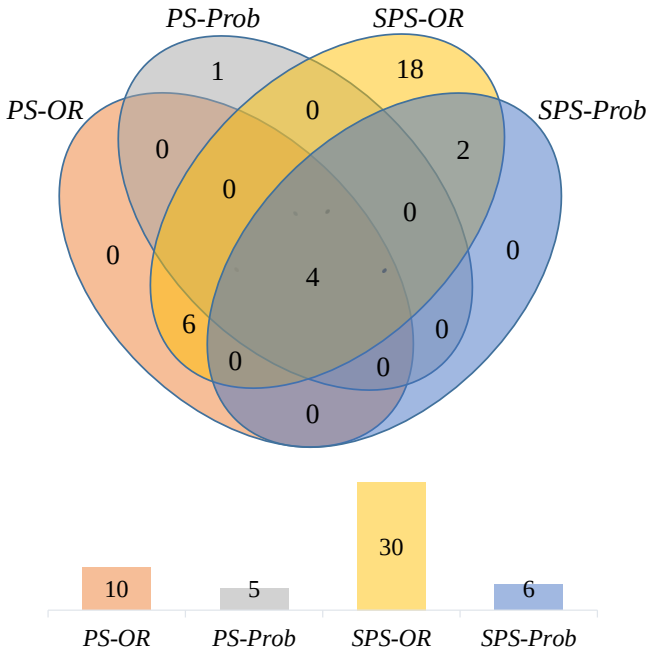
Variables	Values	Comment
price	*vhigh*, *high*, *med* , *low*	buying price
maintC	*vhigh*, *high*, *med* , *low*	cost of the maintenance
doors	2, 3, 4, ≥ 5	number of doors
passengers	2, 4, ≥ 5	capacity in terms of passengers
lugBoot	*small*, *med*, *big*	size of luggage boot
$safety$	*low*, *med*, *high*	estimated safety of the car
buying	*yes*, *no*	outcome, buying the car or not

produce the outcome. Figure 5.4b show that the four algorithms can consistently work well to find cause interactions, however, the overall accuracy is lower than the 2-way cause interaction discovery shown in Figure 5.4a. The reason is that a pruning schema introduced in Section 5.2.3.4 is employed to improve the efficiency, which results in that a miss of a lower level cause interaction may impact the cause interaction discovery in a higher level.

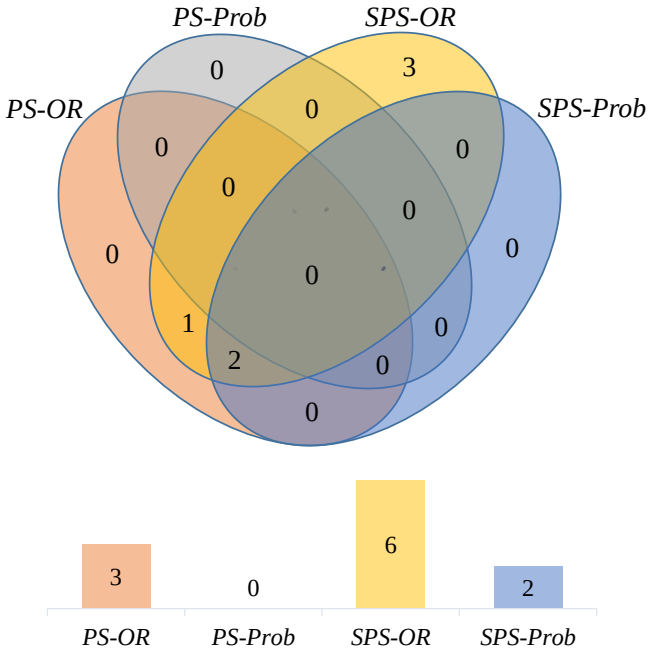
5.3.2 Benchmark data

To investigate the capability of the proposed algorithms in detecting cause interactions in real world cases, we firstly run experiments on a benchmark data set, the Car Evaluation data [118], to detect the cause interactions. The Car Evaluation data set describes the decision making of buying a car, based on six input attributes: price (*price*), maintenance cost (*maint*), the number of doors (*doors*), the number of seats (*seats*), the luggage boot size (*boot_size*) and estimated safety of the car ($safety$) (see detailed description in Table 5.5).

As mentioned above, on all four instantiated algorithms, we use 0.10 as the threshold for the detection of cause interactions from the Car Evalua-



(a) Positive cause interactions



(b) Negative cause interactions

Figure 5.5: Venn diagram of the number of cause interactions discovered by the four algorithms from the Car Evaluation data set.

Table 5.6: Some cause interactions identified by the proposed algorithms, from the Car Evaluation data set.

positive cause interactions	PS-OR	PS-Prob	SGPS-OR	SGPS-Prob
$seats = 4$ & $safety = high$	√	√	√	√
$seats > 4$ & $safety = high$	√	√	√	√
$seats > 4$ & $safety = med$	√	√		√
$boot_size = big$ & $safety = med$	√	√	√	√
negative cause interactions	PS-OR	PS-Prob	SGPS-OR	SGPS-Prob
$price = high$ & $maint = very_high$	√		√	√
$price = very_high$ & $maint = high$	√		√	√

tion data set. Under this setting, the four algorithms find 13, 5, 36 and 8 cause interactions, respectively, which include both positive cause interactions (i.e. increase the chance to buy a car) and negative cause interactions (i.e. decrease the chance to buy a car). Figure 5.5 shows the comparison of the four algorithms in terms of the number of discovered cause interactions. One observation is that SGPS-OR detects the largest number of cause interactions. Moreover, the results discovered by SGPS-OR can cover most of the ones found by other algorithms.

Specifically, TOP 5 positive and TOP 3 negative cause interactions are extracted for a further discussion, where cause interactions discovered by at least two algorithms are shown in Table 5.6. We can see that the four algorithms obtain highly consistent results. As mentioned in Proposition 5.2, the presence of both variables having cause interactions is able to increase the chance of the occurrence of the outcome, which is confirmed by the discovered results. For example, good safety (including medium and high safety) is very likely to have a strong cause interaction with good comfort level (including more seats and big luggage boot size), and the proposed algorithms consistently find such type of cause interactions. An interesting finding is that $seats > 4$ has cause interactions with both $high\ safety$ and $medium\ safety$, while $seats = 4$ only interacts with $high\ safety$, but not $medium\ safety$. The reason would be that when the comfort level reduces,

one will have a higher requirement on the safety factor. We also find that high price and high maintenance cost interact with each other to prevent the buying action, see the negative cause interactions in Table 5.6.

5.3.3 Clinical data

The METABRIC data set [130] contains clinical traits and outcomes for 1981 primary cancer tumors, collected from participants of the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) trial. We follow [131] to do the clinical variable selection and transformations to convert continuous variables to categorical ones. Then we dichotomise them to get the binary data set. The transformed data set contains 35 variables and 1358 samples. The outcome variable in our experiments is 5 year survival, i.e. whether patients survive more than 5 years after being diagnosed as having breast cancer. In addition, the R package "impute" [132] is employed to impute missing data, when the maximum percentage of missing data in any row or column is less than 60%. If more than 60% missed, then this row or column is dropped.

With the default threshold setting ($\eta = 0.10$), 42, 22, 63 and 19 cause interactions are detected by four instantiated algorithms. We extract TOP 5 positive and TOP 5 negative cause interactions for the evaluation, and list the ones detected by at least two algorithms in Table 5.7. The results indicate that $chemo = no$ (the patient did not have chemotherapy) interacts with some symptoms, e.g. $stage = 1$ (early stage), $size$ = 0-19 (small size of tumor), and $positive_lymph = 0$ (no positive lymph nodes), to have a high chance to increase the 5 year survival rate. Chemotherapy usually is not recommended for patients with non-invasive breast cancer, which may be a reason that $chemo = no$ interact with various features to get positive results in our experiments.

Table 5.7: Some cause interactions identified by four algorithms, from the METABRIC data set. (*diag_age*: age at diagnosis of the disease, *size*: size of tumor in cm, *grade*: grade of disease, *stage*: composite of size and # positive nodes, *pos_lymph*: # positive lymph nodes, *removed_lymph*: # lymph nodes removed, *ER.Expr*: estrogen receptor expression, *chemo*: whether patient had chemotherapy, *radiation*: whether patient had radiation therapy, *hormone*: whether patient had hormone therapy.)

positive cause interactions (5 year survival)	PS-OR	PS-Prob	SGPS-OR	SGPS-Prob
$positive_lymph = 0$ & $chemo = no$	√	√		√
$stage = 1$ & $chemo = no$	√	√		√
$diag_age$ = 55-69 & $chemo = no$	√		√	√
$size$ = 0-19 & $chemo = no$	√	√		
$removed_lymph$ = 4-9 & $radiation = yes$			√	√
$ER.Expr = -$ & $radiation = yes$			√	√
negative cause interactions (5 year survival)	PS-OR	PS-Prob	SGPS-OR	SGPS-Prob
$ER.Expr = +$ & $chemo = yes$	√	√	√	√
$positive_lymph$ = 2-3 & $grade = 2$	√			√
$grade = 2$ & $chemo = yes$	√	√		
$diag_age$ = 70-84 & $positive_lymph$ = 2-3		√	√	
$stage = 3$ & $Er.Expr = +$		√		√
$stage = 3$ & $hormone = yes$			√	√
$positive_lymph = 1$ & $removed_lymph$ = 1-3			√	√

5.3.4 Genomic data

BRCA is a breast cancer data set, including the expression profiles of messenger RNAs (mRNAs) and microRNAs (miRNAs, a class of gene regulators). We use the Limma package [133] to find the significantly differentially expressed miRNAs and mRNAs between the tumor and normal samples (p-value < 0.05, adjusted by the Benjamini-Hochberg (BH) method). TOP 100 differentially expressed miRNAs are chosen as the predictor variables. As an example to illustrate the performance of proposed algorithms on genomic data, we then select 3 mRNAs (i.e. BCL-2, RAS and VEGF) from 5 differentially expressed mRNAs in the "pathways in cancer" as the targets. We run the experiments to discover cause interactions between the 100 miRNAs in the regulation of the 3 different mRNAs (we only discover 2-way cause interactions in this set of experiments). TargetScan (v7.0) [134] provides predictions on direct binding information using the sequence data and is

used as the input to filter miRNAs incapable of binding the target mRNAs. Thus the output of experiments is the cause interactions of miRNAs discovered by the proposed algorithms and also included in TargetScan v7.0. For simplicity, we denote the output as "miRNA-mRNA-miRNA", which represents (1) each miRNA is capable of binding the target mRNA, and (2) these two miRNAs have a cause interaction with respect to co-regulating the mRNA.

Although there is no complete ground truth of the relationships between miRNAs and mRNAs, the enrichment analysis can be employed to see how many identified cause interactions satisfy: (1) both of the interacting miRNAs and the target mRNA are in the same pathway; and (2) the miRNAs having cause interactions have been confirmed experimentally to regulate the same target mRNA. Specifically, in the experiments, we focus on one significant KEGG (Kyoto Encyclopedia of Genes and Genomes) [135] pathway, "pathways in cancer". The MiRSEA package (https://cran.r-project.org/web/packages/MiRSEA/index.html) is used to detect the set of miRNAs that co-regulate in the pathway "pathways in cancer". The experimentally confirmed miRNA-mRNA regulatory relationships are downloaded from [136].

A cumulative hypergeometric distribution model is employed to assess the statistical significance of cause interactions discovered. Let N be the number of possible patterns (miRNA-mRNA-miRNA cause interaction triplets) in our data set, K be the number of patterns in the data satisfying the above two conditions, n be the number of patterns discovered and k be the number of validated patterns. The p-value of the validation results is obtained using the cumulative hypergeometric test formula:

$$p(X \geq k) = \sum_{i=k}^{n} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} \tag{5.5}$$

Table 5.8: Statistical significance of validated miRNA-mRNA-miRNA cause interaction triplets.

	BCL-2	RAS	VEGF
PS-OR	1.51E-16	7.81E-01	1.89E-03
PS-Prob	1.31E-12	8.39E-03	8.04E-03
SGPS-OR	1.10E-16	1.51E-02	1.09E-08
SGPS-Prob	1.46E-05	1.55E-01	5.07E-02

We set 0.10 as the threshold to discover miRNA-mRNA-miRNA cause interaction triplets from data. TOP 50 detected cause interactions on each mRNA are extracted to analyze the performance of proposed algorithms. Tables 5.8 shows the statistical significance (p-value) of validated miRNA cause interactions on the three different mRNAs. The low p-values in Table 5.8 suggest that the discovered cause interactions are statistically significant

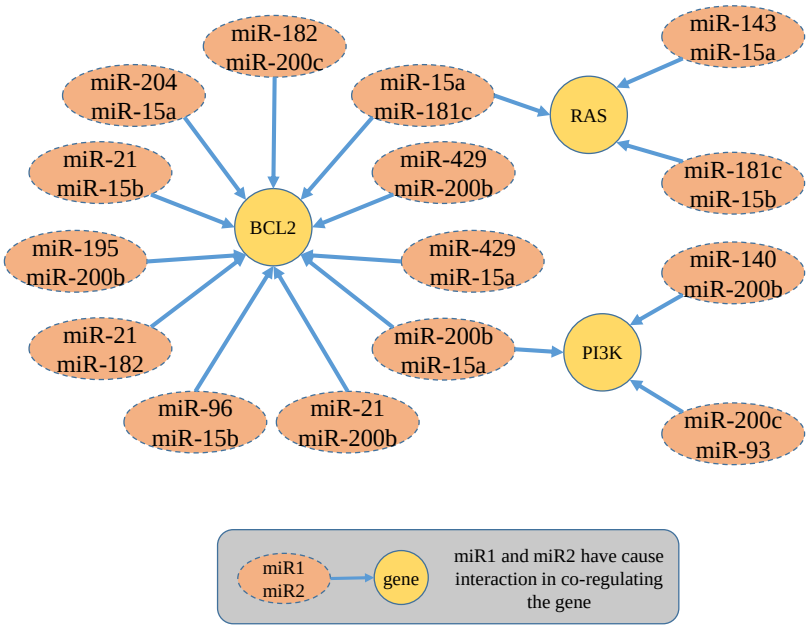


Figure 5.6: The cause interactions discovered by at least three proposed algorithms, which are associated with the KEGG pathway, pathways in cancer, and have the experimentally confirmed regulatory relationships.

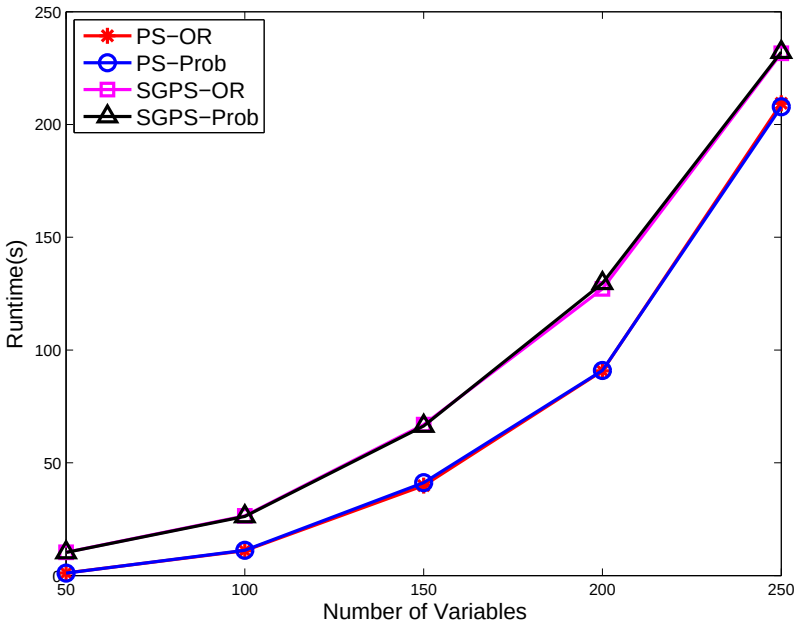


Figure 5.7: Scalability with the number of variables.

and not obtained by chance. The results show that the four proposed algorithms have a good performance in detecting cause interactions between miRNAs in co-regulating a target mRNA.

Figure 5.6 shows an example of the miRNA cause interactions discovered by at least three proposed methods, with all miRNAs and mRNAs appearing in the "pathways in cancer" and the miRNA-mRNA regulatory relationships are experimentally confirmed.

5.3.5 Scalability

We run PS-OR, PS-Prob, SGPS-OR, and SGPS-Prob with ten data sets to test the scalability in terms of the numbers of variables and samples. The data sets are randomly generated by using logistic regression, where the predictor variables and the target variable are binary. All the scalability

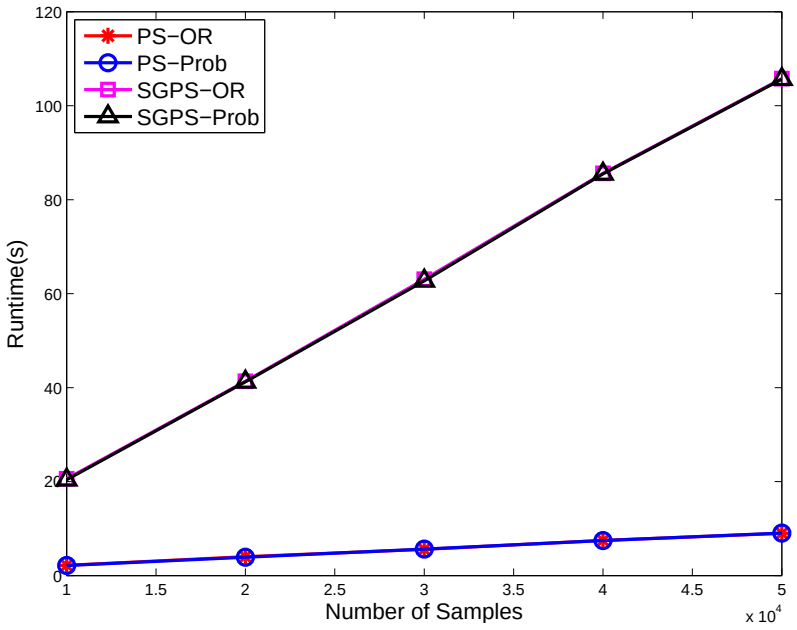


Figure 5.8: Scalability with the number of samples.

evaluation experiments are run on the same computer with a 3.4 GHz Quad-core CPU and 16 GB of memory.

Figure 5.7 shows the running time of the algorithms using data sets with the same sample size (5K) but different number of variables (50, 100, 150, 200 and 250). The four proposed algorithms all scale well with the number of variables. The algorithms using PS (PS-OR and PS-Prob) are slightly more efficient than the ones with SGPS (SGPS-OR and SGPS-Prob), since PS is faster than SGPS. Another observation is that PS-OR and PS-Prob (SGPS-OR and SGPS-Prob) have the same efficiency curve. And the similar situation also appears in Figure 5.8. It is because that the different measures of the ACE estimation (OR or Prob) do not impact the efficiency of algorithms, when the stratification strategies are the same.

We then apply the algorithms to the data sets with 50 variables but different sample sizes (10K, 20K, 30K, 40K and 50K). The execution time is shown in Figure 5.8. The running time of SGPS-OR and SGPS-Prob in-

creases sharply as the data sets get larger, while the other two algorithms are scalable in terms of the sample size. The main reason is that propensity scores are obtained using the logistic regressions, whose time complexity is polynomial to the number of samples. Thus SGPS-OR and SGPS-Prob may not handle data sets with too large sample size, while PS-OR and PS-Prob scale well with both number of variables and number of samples.

5.4 Conclusion

Cause interaction discovery is an important topic in the field of causal discovery, as understanding the interactions between causal factors helps us gain valuable insights into the underlying causal mechanisms. It also provides us an alternative and effective way for identifying important causal factors when it is impossible to discover all the components of a causal mechanism.

The research on cause interactions has had a long history, but finding the interactions directly from data is still a new and challenging topic. There is a severe lack of data-driven approaches to discovering cause interactions, particularly from large data sets.

In this chapter, the concept of cause interactions is re-formalized by considering the detection of cause interactions as a data mining task, such that they can be applied to discovering cause interactions from observational data. A general framework, DACID, has been developed to address the problem of cause interaction discovery around a given target variable, and DACID is instantiated in various ways. The resulting algorithms are sound under the monotonic effect assumption.

In the set of experiments on synthetic data sets, all instantiated algorithms

achieve a high accuracy for cause interaction discovery. Experiments on the benchmark data sets have shown that the proposed framework can find many cause interactions justifiable based on common knowledge. Meanwhile, the proposed algorithms have a good performance in detecting cause interactions between miRNAs in co-regulating a target mRNA. The algorithms also achieve a high computational efficiency, especially for algorithms using perfect stratification.

Chapter 6

A general framework for context specific causal relationship discovery

In the previous two chapters, we have developed the methods to mine combined causes and to identify the interactions between multiple causal factors. These methods are designed to capture the causal knowledge derived from the combination of a set of variables, and each component of the combination has weak (or even no) causal signal on the outcome. In this chapter, we investigate another type of causal relationships involving multiple factors, context specific causal relationships.

A variable has a strong causal effect on an outcome only when the other variable(s) having a specific value. The former variable is called a cause, while the latter is a modifier or context. Such causal relationships are called as context specific causal relationships in this chapter. For example, a treatment is only effective for a group of patients with certain genetic profile [19].

As discussed in Chapter 1, context specific causal relationships are hidden and difficult to be discovered. One challenge is that the overall causal effect may be averaged out to be marginal in the whole population. For example, for a treatment, some patients respond positively and some respond negatively, and hence the overall effect among all patients is marginal. The other main challenge is to distinguish potential causes from contexts, since they corresponds to totally different procedures in a randomized experiment (a context is only a modifier or condition to define a sub-population, while a cause is the action or treatment taken on the sub-population).

A straightforward solution is to find both contexts and the causal relationships under the contexts in one go. However, it is difficult to find context specific causal relationships directly from data without knowing or assuming a cause. Our goal is to design an automated approach to discovering context specific causal relationships from data without any assumptions about the causal relationships. Our proposed framework, Context Specific Causal Rule discovery (CSCR) adapts data mining and machine learning methods to tackle the efficiency issue. We directly make use of the highly efficient and mature data mining to select tentative causal relationships. Then within the much reduced search space, we employ the potential outcome model [33, 34] to assess the relationships to identify causes and their contexts.

6.1 Problem statement and definitions

Given a data set, the research problem of finding context specific causal relationships can be defined as follows.

Problem 6.1 *Given a data set $\boldsymbol{D}$, find all the predictor variables $X_p \in \boldsymbol{V}$ and the corresponding context variables $\boldsymbol{X_c}$ ($\boldsymbol{X_c} \subset \boldsymbol{V}$), such that $X_p \rightarrow Y$ is a causal rule when $\boldsymbol{X_c} = \boldsymbol{x_c}$.*

The potential outcome model [33, 34] is used to estimate the causal effect of a predictor variable X_p on the outcome Y in this chapter. As introduced in Section 3.2, in this model, each individual in a population has two potential outcomes with respect to a treatment: when taking the treatment ($X_p = 1$), the potential outcome is Y_1; and when not taking the treatment ($X = 0$), the potential outcome is Y_0. Then the Average Causal Effect (ACE) can be evaluated by aggregating the causal effects of individuals.

$$ACE(X_p \to Y) = E[Y_1] - E[Y_0] \tag{6.1}$$

where $E[.]$ denotes the expectation. The causal effect metric ACE will be employed here to determine if a candidate rule $X_p \to Y$ is a causal rule or not. The formal definition is given as follows.

Definition 6.1 (Causal rule) *Given a data set $\boldsymbol{D}$, a rule $X_p \to Y$ is a causal rule in the context $\boldsymbol{X_c} = \boldsymbol{x_c}$, if*

$$ACE(X_p \to Y) \geq \eta \tag{6.2}$$

where η is the minimal causal effect threshold specified by users.

In order to identify context specific causal rules, the context specific causal effect (which is represented as the conditional average treatment effect in [18]) is defined as the causal effect under a specific context:

$$ACE(X_p \to Y|\boldsymbol{X_c} = \boldsymbol{x_c}) = E[Y_1 - Y_0|\boldsymbol{X_c} = \boldsymbol{x_c}] \tag{6.3}$$

Thus, the definition of context specific causal rules is expressed as follows.

Definition 6.2 (Context specific causal rule) *Given a data set $\boldsymbol{D}$, a rule $X_p \to Y|\boldsymbol{X_c} = \boldsymbol{x_c}$ is a causal rule in the context $\boldsymbol{X_c} = \boldsymbol{x_c}$, if*

$$ACE(X_p \to Y|\boldsymbol{X_c} = \boldsymbol{x_c}) \geq \eta \tag{6.4}$$

where η is the minimal causal effect threshold specified by users.

Then we can refine the research problem as follows.

Problem 6.2 *Given a data set* $\boldsymbol{D}$ *with a set of predictor variables* $\boldsymbol{V}$ *and the target variable* Y*, find the predictor variables* X_p *(*$X_p \in \boldsymbol{V}$*) and corresponding contexts* $\boldsymbol{X_c} = \boldsymbol{x_c}$ *(*$\boldsymbol{X_c} \subset \boldsymbol{V}$*), such that* $ACE(X_p \to Y|\boldsymbol{X_c} = \boldsymbol{x_c})$ *is a context specific causal rule based on Definition 6.2.*

6.2 Context specific causal relationship discovery

In this section, we firstly present an overview of the proposed framework for context specific causal relationship discovery, Context Specific Causal Rule discovery (CSCR). Then we introduce the detailed process of CSCR and present several algorithmic instantiations from the framework to detect context specific causal rules.

6.2.1 Overview of CSCR

CSCR includes two main steps. The first step of CSCR is to select tentative causes and context variables with respect to a given target, by employing data mining approaches. Most existing work for identifying causal relationships using ACE is given a tentative causal relationship from domain experts, and ACE is used as a way similar to a hypothesis test. In our task, we explore context specific causal relationships without prior causal relationships. In other words, every variable is a potential cause and every value of a variable combination is a possible context. The search space is huge. Thus we will utilize some data mining means to reduce the search

space and make the algorithm practical for large data sets. Then a confidence test is conducted to remove tentative causes and context variables which may have a weak causal sign to the target.

In the second step, a well-known causal framework, the potential outcome model, is utilized to assess causal relationships by separating causes and contexts. Within the process of causal tests, we use a pruning strategy to reduce the search space and improve the efficiency.

6.2.2 Selecting rules with strong causal signals

CSCR firstly picks up a proper search base of tentative causes and context variables by invoking some efficient data mining approaches, e.g. decision tree [137] and association rule mining [3]. The use of an efficient data mining approach is motivated by the following two reasons. Firstly, we take the advantage of data mining technique for high computational efficiency, which is capable of handling both large sized and high dimensional data. Hence basing CSCR on data mining technique will be practical for various applications. Secondly, the outcome of data mining research encodes association relationships between predictor variables and the outcome, which are likely the indicators of causal relationships.

The results of data mining approaches are transformed to the form of rules, e.g. $(\boldsymbol{X} = \boldsymbol{x}) \rightarrow (Y = y)$ or $\boldsymbol{x} \rightarrow y$, such that the results can be easily utilized in the following causal tests. For instance, given a decision tree, each path in the decision tree is considered as a decision rule $(\boldsymbol{X} = \boldsymbol{x}) \rightarrow (Y = y)$, where $\boldsymbol{X}$ and Y are the predictor variables and the target on a path of the decision tree, and $\boldsymbol{x}$ and y are the corresponding values respectively.

Next, for each rule $\boldsymbol{x} \rightarrow y$, two tests are performed to refine the rules

and pick up the ones with strong causal signals. Firstly a support test is conducted on each rule such that it has enough samples for the ACE estimation, where support is an indication of how frequently the rule appears in the data. A rule $\boldsymbol{x} \rightarrow y$ is frequent if its support exceeds the minimum support threshold α specified by users, i.e. $supp(\boldsymbol{x} \rightarrow y) > \alpha$. Then we test the confidence to remove the rules if it has low confidence, since causal signal in a low confidence rule is weak. Here the confidence of $\boldsymbol{x} \rightarrow y$ is defined as the proportion of individuals containing $\boldsymbol{x}$ which also contains y, which is expressed in the following:

$$conf(\boldsymbol{x} \rightarrow y) = \frac{supp(\boldsymbol{x}y)}{supp(\boldsymbol{x})}. \tag{6.5}$$

Only a rule $\boldsymbol{x} \rightarrow y$ with high support and confidence (compared with user defined thresholds) will be inserted into the candidate set for causal rule discovery.

6.2.3 Causal rule tests

Given refined rules, the potential outcome framework [33, 34] is employed to detect the causal signals, including global and context specific causal rule discovery, shown in ALGORITHM 6.1. Lines 3-8 shows the process of global causal rule discovery and context specific causal discovery is conducted in lines 9-20.

In global causal rule discovery, a function $causalTest()$ is invoked to perform the causal tests, which is shown in ALGORITHM 6.2. The potential outcome framework [33, 34] is utilized to estimate the Average Causal Effect (ACE). As mentioned in Chapter 3, the major issue for ACE estimation is to unbiasedly estimate the counterfactual outcomes, e.g. what the effect would be if a person had not taken a treatment (actually the person

ALGORITHM 6.1 Causal rule discovery

Input: A data set $\boldsymbol{D}$, refined rules R and the minimal causal effect threshold η.

Output: $\boldsymbol{C_Y}$, the set of causal rules to the target Y.

for each rule $R \in \boldsymbol{R}$ **do**
  extract the predictor variables $\boldsymbol{X}$ and Y and their corresponding values $\boldsymbol{x}$ and y from R
  // Global causal test
  for each variable $X_p \in \boldsymbol{X}$ **do**
    $ACEValue = causalTest(\boldsymbol{D}, X_p, \emptyset, \emptyset, Y)$
    if $ACEValue > \eta$ **then**
      $\boldsymbol{C_Y} = \boldsymbol{C_Y} \cup \{X_p \rightarrow Y\}$
    end if
  end for
  // Context specific causal test
  for each variable $X'_p \in \boldsymbol{X}$ **do**
    $\boldsymbol{X'} = \boldsymbol{X} \backslash \{X'_p\}$
    for each variable set $\boldsymbol{X_c} \subset \boldsymbol{X'}$ **do**
      if $RedundantTest(X'_p, \boldsymbol{X_c}, \boldsymbol{C_Y})$ **then**
        continue
      end if
      $ACEValue = causalTest(\boldsymbol{D}, X'_p, \boldsymbol{X_c}, \boldsymbol{x_c}, Y)$
      if $ACEValue > \eta$ **then**
        $\boldsymbol{C_Y} = \boldsymbol{C_Y} \cup \{X'_p \rightarrow Y | \boldsymbol{X_c} = \boldsymbol{x_c}\}$
      end if
    end for
  end for
end for
Output $\boldsymbol{C_Y}$

ALGORITHM 6.2 $causalTest(\boldsymbol{D}, X_p, \boldsymbol{X_c}, \boldsymbol{x_c}, Y)$

1: **if** $\boldsymbol{X_c} \neq \emptyset$ **then**
2: $\quad \boldsymbol{D} = subdata(\boldsymbol{D}, \boldsymbol{X_c}, \boldsymbol{x_c})$
3: **end if**
4: $\boldsymbol{C} = stratifyAttr(\boldsymbol{D}, \boldsymbol{V}, X_p, Y)$ // Select covariates
5: $stratification(\boldsymbol{D}, X_p, \boldsymbol{C})$ // Stratify data
6: $ACEValue = ACE(X_p, Y, \boldsymbol{C})$
7: return $ACEValue$

did take the treatment). Stratification [111] is a commonly used approach to mimicking randomized treatment assignment in non-controlled studies. The first step for the stratification is to determine covariates $\boldsymbol{C}$, see line 4 in ALGORITHM 6.2, such that the treatment assignment satisfies the unconfoundedness assumption for the variable selection introduced in Chapter 3.

For the convenience of discussion, a causal graphical model shown in Figure 6.1 is used to clarify the confounding issue. In this figure, the nodes represent single variables and the edges denote the causal links between nodes. For example, the connection $X \rightarrow Y$ means X is a cause of the target Y. Apart from treatment X and the target Y, variables are categorized into four different types: (i) Indirect causes (e.g. U), which indirectly cause a change of Y; (ii) Confounders (e.g. C), which are the common causes of X and Y; (iii) Direct causes (e.g. Z), apart from X, which are directly cause

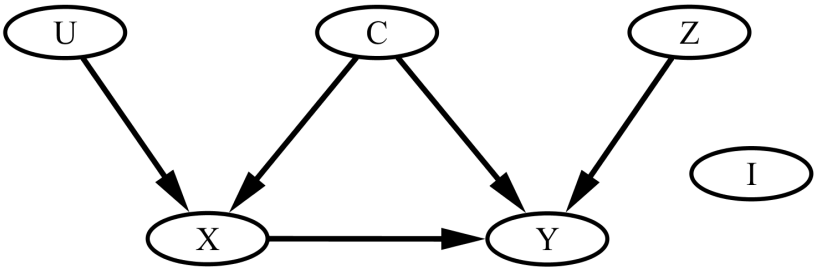


Figure 6.1: A causal graphical model

Y; and (iv) Irrelevant variables (e.g. I), which are totally independent with both X and Y.

From this causal structure, we can see confounders C are the ones that may influence the causal effect estimation of the treatment X on Y. Thus to achieve the assumption of unconfoundedness, variables known to make causal effects on both treatment assignment and the outcome are required to be included in stratification. Unfortunately, the causal graph is normally unknown. All variables that are associated with both the treatment variable X and the target Y are included into covariates $\boldsymbol{C}$ for stratification. Covariates $\boldsymbol{C}$ are the superset of confounders C shown in Figure 6.1, as a variable can never be a cause of another if they are independent.

The second step is to choose a distance measure for stratification, shown in line 5 in ALGORITHM 6.2. Propensity score [33] is used here to measure distance between individuals, which summarize covariates $\boldsymbol{C}$ into the probability of the individual receiving the treatment conditioning on $\boldsymbol{C}$. If treatment assignment is unconfounded given covariates $\boldsymbol{C}$, then treatment assignment is also unconfounded given the propensity score $e(\boldsymbol{C})$. The goal of stratification is that the individuals with similar propensity scores are divided into the same stratum, that is in each stratum the individuals are indistinguishable apart from being treated or not. Finally, ACE of treatment X_p on the target T can be estimated by invoking Formula 6.1.

Then we move to context specific causal rule discovery. The discovery of context specific causal rules from a rule $(\boldsymbol{X} = \boldsymbol{x}) \rightarrow (Y = y)$ includes two nested loops (lines 9 to 20 in ALGORITHM 6.1). In the outer loop, we traverse the predictor variables $\boldsymbol{X}$ in R as the candidate treatment variable X'_p, while the inner loop enumerates the subsets of $\boldsymbol{X} \backslash \{X'_p\}$ finding the contexts. In other words, we enumerate all possible pairs of treatment and context variables, $\langle X'_p, \boldsymbol{X_c} \rangle$ from $\boldsymbol{X}$ in R.

Then causal tests are performed to estimate the causal effect of X'_p on Y, under the context $\boldsymbol{X_c}$. With each subset $\boldsymbol{X_c}$ of $\boldsymbol{X}\backslash\{X'_p\}$, the subset of data is extracted from the original data with $\boldsymbol{X_c} = \boldsymbol{x_c}$ (see lines 1-3 in ALGORITHM 6.2), where $\boldsymbol{x_c}$ is the value of $\boldsymbol{X_c}$ as indicated in the rule R. Next, we select the covariate variables and perform stratification on them. Formula 6.3 is employed to estimate the causal effect of tentative context specific causal rules.

A bottle-neck for context specific causal rule discovery is the enumeration of different contexts in the variable set of the antecedent of a rule. Thus a pruning strategy is developed to address the efficiency problem, before the procedure of causal tests. Function *RedundantTest*() (lines 12 to 14 in ALGORITHM 6.1) is invoked to test if the rule $X_p \rightarrow Y|\boldsymbol{X_c} = \boldsymbol{x_c}$ is redundant. Only if the rule is not redundant, then causal test (line 15) is performed to estimate the causal effect of X'_p on Y under the context $\boldsymbol{X_c} = \boldsymbol{x_c}$.

As we know, if a causal relationship holds in a population, then it should hold in each of the sub-populations, which is redundant and less interest to us. In other words, if a rule is identified as a context specific causal rule, its more specific rules (i.e. with more conditions than the general one) are implied by the general causal rule. For example, if Children's Panadol is effective for relieving child under 12 from fever and pain (i.e. $Panadol \rightarrow recovery|age < 12$), then we can conclude that it is also effective for boys under 12 (i.e. $Panadol \rightarrow recovery|\{age < 12, sex = male\}$). We call such more specific rule as a redundant rule, which is formally defined in the following.

Definition 6.3 (Redundant rule) *A rule* $X_p \rightarrow Y|\{\boldsymbol{X_c} = \boldsymbol{x_c}, \boldsymbol{X_a} = \boldsymbol{x_a}\}$ *is redundant, if* $X_p \rightarrow Y|\boldsymbol{X_c} = \boldsymbol{x_c}$ *is a context specific causal rule, where* $\boldsymbol{X_a}$ *is an additional condition defining a specific sub-population.*

We are not interested in redundant rules as the causal relationships (if any), since the causal relationships are already implied by their more general context specific causal rules. Thus we exclude redundant rules in the algorithm to reduce the search space. We implement the CSCR framework level-by-level, and in each level check if the tentative causal rule is redundant or not. In other words, once we find a context specific causal rule (including a global causal rule where the context variable set $X_c = \emptyset$), we stop searching for its more specific context specific rules.

6.2.4 Instantiations of CSCR

The outcome of data mining approaches, including single decision tree, multiple decision trees and association rules, will be utilized to reduce the search space of context specific causal rule discovery. In this section, we will detail three instantiation algorithms of CSCR.

6.2.4.1 CSCR with a single decision tree

Decision tree encodes context specific relationships between predictor variables and the outcome, which are likely the indicators of context specific causal relationships.

We use the following example to illustrate the process of CSCR with a single decision tree.

Example 6.1 *Figure 6.2 shows a decision tree learned from the synthetic data set. The leftmost branch in the decision tree can represent a decision rule* $(V8 = 0, V4 = 0) \rightarrow (outcome = no)$*. Then the predictor variables in*

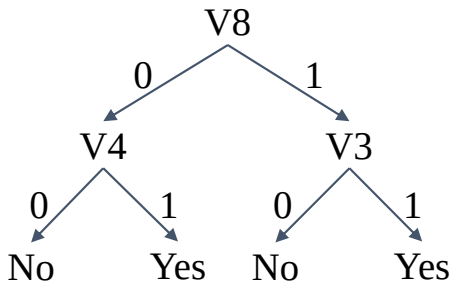
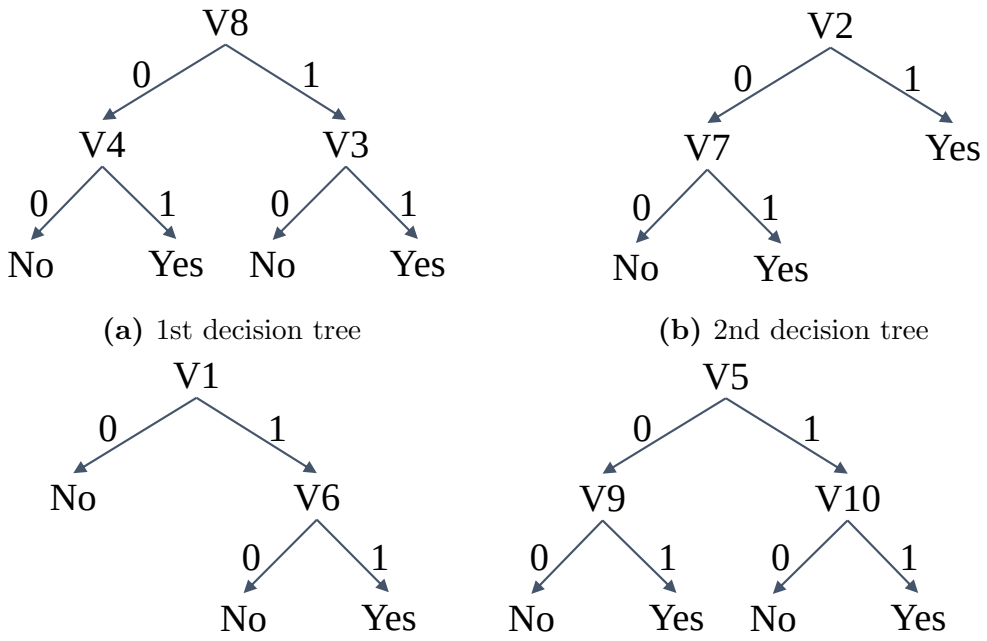


Figure 6.2: A decision tree of the toy data set.

the rule, i.e. $V8$ and $V4$, are considered as the search base of both potential causes and contexts. Meanwhile, the values of predictor variables in this path may indicate context specific causal relationship. For example, we may test the causal relationship between $V4$ and the outcome under the context $V8 = 0$. However, we do not test the causal relationship between $V4$ and the outcome under the context $V8 = 1$, as this rule only implies that $V4$ correlates to outcome when $V8 = 0$. Thus we can significantly reduce the search space of tentative context specific causal relationships. Then causal tests are performed to confirm context specific causal relationships. CSCR with a single decision tree is notated as CSCR-S.

The performance of CSCR could be sensitive to the results of decision tree construction. Although we can control the size of a decision tree by setting different parameters, a decision tree normally makes use of a small subset of attributes in decision. Therefore, a key limitation of using decision tree for our purpose is that it may not cover all possible causal factors and the context variables, and thus we may miss some potential causal relationships. In the following two sections, we will introduce other two instantiations of CSCR, which are expected to achieve better performance of causal discovery.



(a) 1st decision tree

(b) 2nd decision tree

(c) 3rd decision tree

(d) 4th decision tree

Figure 6.3: Diversified multiple trees of the toy data set.

6.2.4.2 CSCR with multiple decision trees

In this section, we present a instantiation of CSCR with multiple decision trees. Multiple trees easily include more variables and thus potentially cover more possible causal factors and the context variables into the search space, which in turn improves the chance of context specific cause detection.

Li et al. [138] developed a ensemble classifier, Diversified Multiple Trees (DMT). The basic idea of DMT is to build a set of disjunct trees. Specifically, DMT sequentially builds m decision trees, where variables used in a tree are not to be used in the construction of the next tree. Thus the output decision trees are disjunct, i.e. without using overlapping variables. In the process, C4.5 [137] is utilized to build each decision tree.

Figure 6.3 shows DMT learned from the toy data set mentioned in Section

6.2.4.1, where $m = 4$. The first decision tree of DMT (Figure 6.3a) is the same as single decision tree shown in Figure 6.2, as both DMT and single decision tree employ C4.5. When building the second decision tree (Figure 6.3b), variables included in the first tree are knocked off. Similar process is performed to build the third and forth decision trees. We can see that these four decision trees are disjunct.

Then each path of each decision tree in DMT is extracted as a decision rule. Similarly, the predictor variables in each decision rule are treated as the search base of both potential causes and contexts. Finally, causal tests are performed to detect context specific causal relationships. To avoid confusion, we use notation CSCR-m to denote CSCR with m decision trees.

6.2.4.3 CSCR with association rule mining

Association rule mining [3] is another popular tool to detect interesting associations among variables. Compared with decision tree, it is capable of discovering more comprehensive association relationships. This section is deliberately presented to show that the proposed algorithm is able to be applied to association rules to detect context specific causal relationships.

In this chapter, measures of interestingness, support, confidence and lift, are used to identify association rules, which are formally defined as follows.

Definition 6.4 (Association rule) *A rule $\boldsymbol{x} \rightarrow y$ is an association rule, if (1) $supp(\boldsymbol{x} \rightarrow y) > \alpha$, (2) $conf(\boldsymbol{x} \rightarrow y) > \beta$, and (3) $lift(\boldsymbol{x} \rightarrow y) > \gamma$, and $LHS(\boldsymbol{x} \rightarrow y) = \boldsymbol{x}$ is the LHS (left-hand-side) of the rule and $RHS(\boldsymbol{x} \rightarrow y) = y$ is the RHS (right-hand-side).*

where the lift of a rule $\boldsymbol{x} \rightarrow y$ is defined as

$$lift(\boldsymbol{x} \rightarrow y) = \frac{supp(\boldsymbol{x}y)}{supp(\boldsymbol{x})supp(y)} \tag{6.6}$$

In the definition, the LHS is the predictor variables, which are the search bass of tentative causes and contexts, and the RHS is the outcome of a rule. Association rules detected from data are treated as the search space of causal rules, where each association rule is tested if it is causal or not. We use CSCR-Ass to notate the approach, CSCR with association rule mining.

6.3 Experiments

In this section, we firstly introduce the process of the synthetic data set generation. Then we implement experiments for the instantiations of CSCR on the synthetic data sets for evaluation. Moreover, we also applied the proposed methods to two real world data sets. Finally, we compare our methods with an existing causal tree method, CRT in term of context specific causal relationship discovery.

6.3.1 Synthetic data generation

In order to evaluate the proposed framework, we generate several synthetic data sets containing context specific causal relationships.

The synthetic data sets are generated with four main steps: (i) create randomly two Causal Bayesian Networks with the same number of variables by using the TETRAD software tool (http://www.phil.cmu.edu/tetrad/),

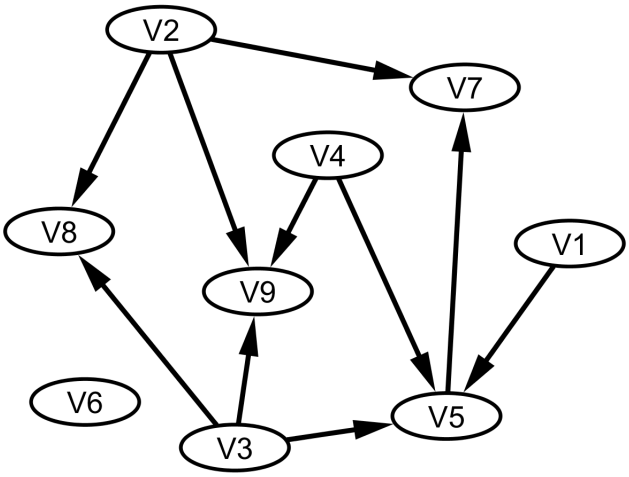
where the Direct Acyclic Graphs (DAGs) and Conditional Probability Tables (CPTs) are both created randomly by the software; (*ii*) generate two data sets for two causal Bayesian networks based on their respective conditional probability tables via the built-in Bayes Instantiated Model; (*iii*) add one more column to these two data sets, as the context variable X_c, such that $X_c \equiv 0$ in the first data set and $X_c \equiv 1$ in the another one; and (*iv*) merge these two new data sets by columns to obtain the final data set.

We use the following example to show the process of synthetic data set generation.

Example 6.2 *We generate two Direct Acyclic Graphs and their respective Conditional Probability Tables using the TETRAD software tool, shown in Figures 6.4 and 6.5. Data part I is generated by using DAG (Figure 6.4a) and CPT (Figure 6.4b), while data part II is created from DAG (Figure 6.5a) and CPT (Figure 6.5b).*

One more variable X_c is added into these two data sets as the context variable, respectively. In data part I, the newly added variable X_c is consistently equal to zero, while all values of X_c in data part II are ones. Thus their variable sets both become $\{X_c, V1, V2, ..., V9\}$, where $V9$ is set as the target in the following experiments. Then we merge these two data sets by columns, i.e. the samples with the same variable names are put in the same columns, to obtain a new data set Syn-10.

In the data set Syn-10, the structure Figure 6.4a only exists in the context $X_c = 0$ and Figure 6.5a only appears within $X_c = 1$. In other words, if we extract a subset of data with $X_c = 0$ from Syn-10, then a causal Bayesian network like Figure 6.4a can be learned, and the CBN in Figure 6.5a can be learned from the subset of data with $X_c = 1$. When we set the variable $V9$ as the target, the data set Syn-10 contains the following context specific causal relationships: in the context $X_c = 0$, $V2$, $V3$ and $V4$ are the causes



(a) DAG of data part I

P(V1)	
V1=0	0.71
V1=1	0.29

P(V2)	
V2=0	0.59
V2=1	0.41

P(V3)	
V3=0	0.84
V3=1	0.16

P(V4)	
V4=0	0.41
V4=1	0.59

P(V7 | V2, V5)	(V2=0, V5=0)	(V2=0, V5=1)	(V2=1, V5=0)	(V2=1, V5=1)
V7=0	0.94	0.08	0.39	0.93
V7=1	0.06	0.92	0.61	0.07

P(V6)	
V6=0	0.60
V6=1	0.40

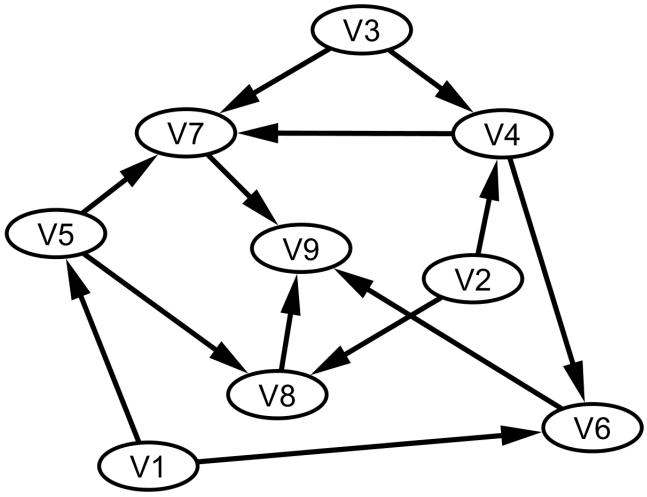
P(V8 | V2, V3)	(V2=0, V3=0)	(V2=0, V3=1)	(V2=1, V3=0)	(V2=1, V3=1)
V8=0	1.00	0.24	0.17	0.18
V8=1	0.00	0.76	0.83	0.82

P(V5 | V1, V3, V4)	(V1=0, V3=0, V4=0)	(V1=0, V3=0, V4=1)	(V1=0, V3=1, V4=0)	(V1=0, V3=1, V4=1)	(V1=1, V3=0, V4=0)	(V1=1, V3=0, V4=1)	(V1=1, V3=1, V4=0)	(V1=1, V3=1, V4=1)
V5=0	0.51	0.54	0.69	0.94	0.10	0.32	0.56	0.79
V5=1	0.49	0.46	0.31	0.06	0.90	0.68	0.44	0.21

P(V9 | V2, V3, V4)	(V2=0, V3=0, V4=0)	(V2=0, V3=0, V4=1)	(V2=0, V3=1, V4=0)	(V2=0, V3=1, V4=1)	(V2=1, V3=0, V4=0)	(V2=1, V3=0, V4=1)	(V2=1, V3=1, V4=0)	(V2=1, V3=1, V4=1)
V5=0	0.00	0.04	0.24	0.91	0.01	0.24	0.00	0.41
V5=1	1.00	0.96	0.76	0.09	0.99	0.76	1.00	0.59

(b) CPT of data part I

Figure 6.4: The causal Bayesian network of synthetic data in the context $X_c = 0$ (data part I).



(a) DAG of data part II

P(V1)	
V1=0	0.16
V1=1	0.84

P(V2)	
V2=0	0.02
V2=1	0.98

P(V3)	
V3=0	0.67
V3=1	0.33

P(V5 | X1)	V1=0	V1=1
V5=0	0.49	0.71
V5=1	0.51	0.29

P(V4 | V2, V3)	(V2=0, V3=0)	(V2=0, V3=1)	(V2=1, V3=0)	(V2=1, V3=1)
V4=0	0.03	0.44	0.11	0.03
V4=1	0.97	0.56	0.89	0.97

P(V6 | V1, V4)	(V1=0, V4=0)	(V1=0, V4=1)	(V1=1, V4=0)	(V1=1, V4=1)
V4=0	0.58	0.44	0.25	0.04
V4=1	0.42	0.56	0.75	0.96

P(V8 | V2, V5)	(V2=0, V5=0)	(V2=0, V5=1)	(V2=1, V5=0)	(V2=1, V5=1)
V8=0	0.04	0.05	0.89	0.62
V8=1	0.96	0.95	0.11	0.38

P(V7 | V3, V4, V5)	(V3=0, V4=0, V5=0)	(V3=0, V4=0, V5=1)	(V3=0, V4=1, V5=0)	(V3=0, V4=1, V5=1)	(V3=1, V4=0, V5=0)	(V3=1, V4=0, V5=1)	(V3=1, V4=1, V5=0)	(V3=1, V4=1, V5=1)
V7=0	0.20	0.15	0.04	0.11	0.01	0.62	0.00	0.25
V7=1	0.80	0.85	0.96	0.89	0.99	0.38	1.00	0.75

P(V9 | V6, V7, V8)	(V6=0, V7=0, V8=0)	(V6=0, V7=0, V8=1)	(V6=0, V7=1, V8=0)	(V6=0, V7=1, V8=1)	(V6=1, V7=0, V8=0)	(V6=1, V7=0, V8=1)	(V6=1, V7=1, V8=0)	(V6=1, V7=1, V8=1)
V9=0	0.10	0.45	0.01	1.00	0.14	0.09	0.41	0.35
V9=1	0.90	0.55	0.99	0.00	0.86	0.91	0.59	0.65

(b) CPT of data part II

Figure 6.5: The causal Bayesian network of synthetic data in the context $X_c = 1$ (data part II).

Table 6.1: Causal rules discovered by CSCR-S on Syn-10.

$V9=1$		$V9=0$	
Cause	Context	Cause	Context
$V2$	$X_c=0$	$V7$	$X_c=1$
$V3$	$X_c=0$	$V8$	$X_c=1$
$V4$	$X_c=0$	$V8$	$X_c=0$

of $V9$*; and in* $X_c=1$*,* $V6$*,* $V7$ *and* $V8$ *are the causes.*

6.3.2 Performance evaluation

Now we use Syn-10 as an example to illustrate the implementation of the experiments. For simplicity, here we only run CSCR-S on Syn-10. Firstly a C4.5 decision tree is built for the generated data set by using Weka [139]. We set the minimum number of instances per leaf to 5‰ of the sample size, and other settings are default. Then we constraint the minimal confidence threshold to 0.70 for high confidence decision rule selection, and if the ACE of a rule is larger than 0.20, then this rule is a context specific causal rule. Here we only discuss the results within the context $X_c=0$ and $X_c=1$, since all we know about the ground truth are the causes under these contexts.

The results discovered by CSCR-S from Syn-10 are listed in Table 6.1. CSCR-S discovers five causal rules consistent with the ground truth, and also includes one false positive $V8 \rightarrow V9|X_c=0$ and one false negative $V6 \rightarrow V9|X_c=1$.

Following the described steps in Section 6.3.1, we generate another four synthetic data with 20, 30, 40 and 50 variables (Syn-20, Syn-30, Syn-40 and Syn-50) respectively. Each data set has 10K samples. we use precision (P), recall (R) and F_1-measure (F_1) as evaluation metrics to quantify the performance of CSCR in term of accuracy. They are defined based on the

similarity between the output results and the ground truth.

$$P = \frac{tp}{tp + fp} \tag{6.7}$$

$$R = \frac{tp}{tp + fn} \tag{6.8}$$

$$F_1 = \frac{2PR}{P + R} \tag{6.9}$$

where tp, fp and fn denote the number of true positives, false positives and false negatives respectively.

For the parameter settings, in CSCR-S and CRCS-m, we still set 5‰ of the sample size as the minimum number of instances per leaf, when building decision trees. In CSCR-Ass, the minimum support, confidence and lift thresholds are set as 0.05, 0.70 and 1.25 respectively, to mine association rules. Meanwhile, for the sake of the high computational efficiency, we only pick up a portion of association rules for the detailed analysis: (1) the length of LHS of an association rule is limited between 3 and 9; (2) all rules are sorted by lift and TOP 100 rules are selected for causal tests; and (3) if a rule has a bigger length but the same or a lower lift, compared with another rule, then this rule will be removed. For all three methods, if the ACE of a rule is larger than 0.20, then this rule is a context specific causal rule.

The results discovered by instantiations of CSCR are shown in Table 6.2. We can see that all instantiations achieve high accuracy and the average of F_1 *score* is mostly larger than 0.85. Specifically CSCR-3 is slightly better performed than other three instantiations. Although CSCR-Ass has the lowest F_1 *score*, it has exactly same *Recall* as other methods. Actually CSCR-Ass generates the most number of tentative context specific causal rules, which also brings some false positives.

Further we examine the performance of instantiation algorithms on a sparse data set, Syn-100. Table 6.3 shows the results of CSCR-S, CSCR-3, CSCR-

Table 6.2: The accuracy of CSCR with synthetic data sets, where P, R and F_1 represent *precision*, *recall* and F_1 *score*.

		CSCR-S	CSCR-3	CSCR-5	CSCR-Ass
Syn-10	P	0.83	0.83	0.83	0.83
	R	0.83	0.83	0.83	0.83
	F_1	0.83	0.83	0.83	0.83
Syn-20	P	1.00	1.00	1.00	0.75
	R	1.00	1.00	1.00	1.00
	F_1	1.00	1.00	1.00	0.86
Syn-30	P	1.00	1.00	1.00	1.00
	R	0.83	0.83	0.83	0.83
	F_1	0.91	0.91	0.91	0.91
Syn-40	P	0.83	1.00	0.75	0.75
	R	1.00	1.00	1.00	1.00
	F_1	0.91	1.00	0.86	0.86
Syn-50	P	0.83	0.83	0.83	0.83
	R	0.83	0.83	0.83	0.83
	F_1	0.83	0.83	0.83	0.83

5 and CSCR-Ass. The accuracy is significantly improved when multiple tree approach is employed, and more decision trees (larger the values of m) bring bigger improvement. This is because that multiple tree approach is able to include more causal factors and context variables in the output trees than the single decision tree, which in turn improves the chance of detecting context specific causes. Similar tend is also shown when compare CSCR-Ass and CSCR-S, where *Recall* is improved significantly.

Table 6.3: Performance of CSCR with Syn-100.

	CSCR-S	CSCR-3	CSCR-5	CSCR-Ass
P	0.80	0.82	0.90	0.82
R	0.80	0.90	0.90	0.90
F_1	0.80	0.86	0.90	0.86

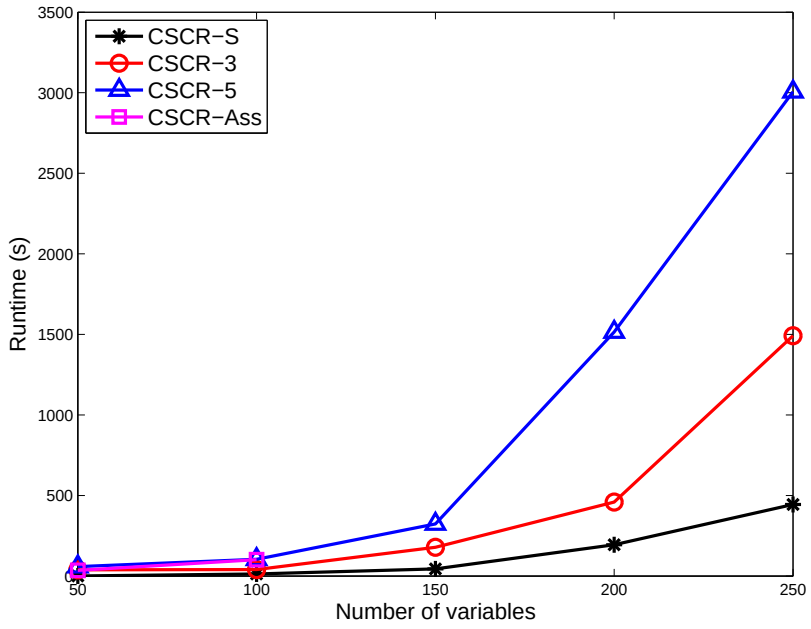


Figure 6.6: Scalability evaluation with CSCR-S, CSCR-3, CSCR-5 and CSCR-Ass, when # context variables = 1, i.e. $|\boldsymbol{X_c}| = 1$

6.3.3 Scalability

To evaluate the efficiency of multiple instantiations of CSCR, we run experiments on five synthetic data sets with 50, 100, 150, 200 and 250 variables, respectively. The experiments are performed on a computer with a 3.4 GHz Quad-core CPU and 16 GB of memory.

We compare CSCR-S, CSCR-3, CSCR-5 and CSCR-Ass in the two cases with $|\boldsymbol{X_c}| = 1$ and $|\boldsymbol{X_c}| = 2$ respectively, where $|\boldsymbol{X_c}|$ means the cardinality of the context variable set. The results in Figure 6.6 show that all four methods scale well in term of the number of variables. CSCR-S is the most efficient one, while the run time of CSCR-m (here $m = 3, 5$) is roughly m times of CSCR-S. Several criteria are set to reduce the number of association rules, such that CSCR-Ass also scales well on Syn-50 and Syn-100 data sets. However, CSCR-Ass did not return results on other three data sets in both

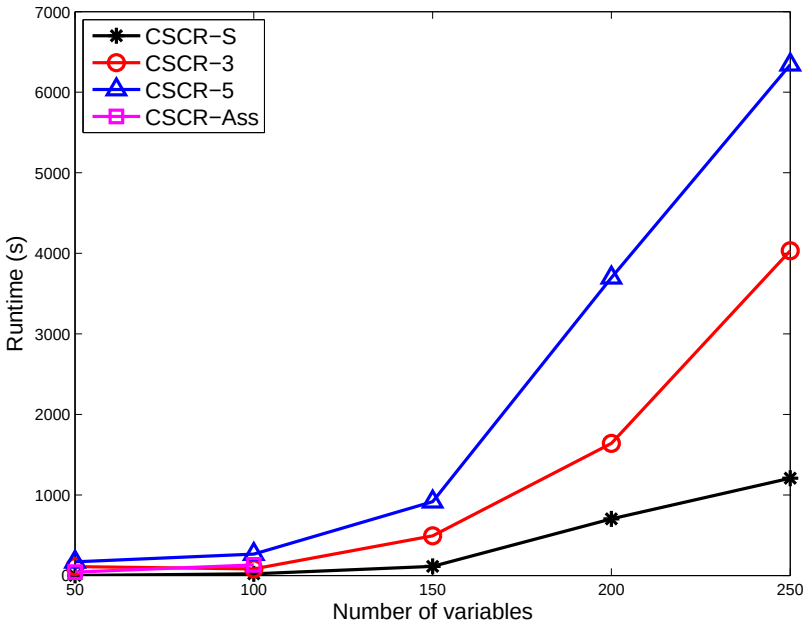


Figure 6.7: Scalability evaluation with CSCR-S, CSCR-3, CSCR-5 and CSCR-Ass, when # context variables = 2, i.e. $|\boldsymbol{X_c}| = 2$

cases of $|\boldsymbol{X_c}| = 1$ and $|\boldsymbol{X_c}| = 2$. This is because that the computer fails to allocate contiguous blocks of memory for association rule generation and thus the implement is stopped. When the size of context variables is set as 2 (see Figure 6.7), they are also efficient and scalable. Specifically, in this setting, methods require slightly more than double of the time used when $|\boldsymbol{X_c}| = 1$.

6.3.4 Applications to real world data sets

We run experiments with two real world data sets, Car Evaluation and Adult data sets, where the threshold of ACE is also set to 0.20.

We firstly apply the proposed framework to the Car Evaluation data set [118] containing decisions of buying a car. A description of the data set is given in TABLE 5.5. A decision of the buying action is made by the

Table 6.4: Comparison of causal rules discovered by CSCR-S, CSCR-3 and CSCR-Ass from the Car Evaluation data set.

Cause	Context	Outcome	CSCR-S	CSCR-3	CSCR-Ass
$safety(high)$	-	$buying$	√	√	√
$passengers(\geq 4)$	-	$buying$	√	√	√
$price(low)$	$safety = med$	$buying$	√	√	√
$price(low)$	$passengers = 4$	$buying$	√	√	√
$price(low)$	$maintC = high$	$buying$	√	√	√
$safety(med)$	$passengers > 4$	$buying$	√	√	√
$safety(med)$	$lugBoot \neq small$	$buying$	√	√	√
$safety(med)$	$price = low$	$buying$	√	√	√
$lugBoot(small)$	$safety = med$	$not\ buying$	√	√	√
$lugBoot(small)$	$passengers > 4$	$not\ buying$	√	√	√

features of a car.

Table 6.4 shows causal rules discovered by three methods with the Car Evaluation data set. We can see that CSCR-S, CSCR-3 and CSCR-Ass find exactly same results. The reason is that in CSCR-S all possible causes and contexts are already included in the decision tree, and multiple decision trees and association rules do not find more causal factors.

In Table 6.4, two context free causal rules are detected, $safety(high) \rightarrow buying$ and $passengers(\geq 4) \rightarrow buying$. The first rule is a reasonable causal rule and the second rule reflects the fact that most cars bought are for four passengers and more. Some interesting context specific causes are detected. For instance, while $safety(high)$ is a context free cause of $buying$, $safety(med)$ can also be a cause of $buying$, but only under some conditions such as $price$ is low or $lugBoot$ is not small (i.e. $price = low$ or $lugBoot \neq small$). Other context specific causal rules are also reasonable. From the rule $lugBoot(small) \rightarrow not\ buying|passengers \geq 5$, we can see that given a car with a large passenger capacity, having a small car boot

Table 6.5: Comparison of causal rules discovered by CSCR-S, CSCR-3 and CSCR-Ass from the Adult data set.

Cause	Context	Outcome	CSCR-S	CSCR-3	CSCR-Ass
$educ(>12)$	-	$\geq 50K$	√	√	√
$age(\geq 30)$	-	$\geq 50K$	√	√	√
$prof(yes)$	-	$\geq 50K$	√	√	√
$sex(male)$	$educ > 12$	$\geq 50K$	√	√	√
$sex(male)$	$prof = yes$	$\geq 50K$	√	√	√
$sex(male)$	$hours \geq 30$	$\geq 50K$	√	√	√
$sex(male)$	$age \geq 30$	$\geq 50K$	√	√	√
$sex(male)$	$educ \geq 9$	$\geq 50K$		√	
$sex(male)$	$nation = USA$	$\geq 50K$			√
$hours(<30)$	$prof = yes$	$< 50K$	√	√	√
$hours(<30)$	$educ > 12$	$< 50K$	√	√	√
$educ(\leq 9)$	$age \geq 30$	$< 50K$		√	
$educ(\leq 9)$	$hours \geq 30$	$< 50K$		√	
$educ(\leq 9)$	$sex = male$	$< 50K$		√	
$educ(\leq 9)$	$prof = yes$	$< 50K$		√	

would lead to *not buying* the car. This is quite reasonable as for a car with a big passenger capacity, a small car boot is unlikely to have enough room for the luggage of all passengers, resulting in "not buying" the car.

The Adult data set [35] is an extraction of 1994 USA census database. The causal rules discovered by CSCR-S, CSCR-3 and CSCR-Ass are shown in Table 6.5, where the target is whether *annual income* is less than $50K$ or not.

Intuitively, education, professional occupation and age are the main factors impacting annual incomes. All three methods capture these three context free causal rules: $educ(>12) \rightarrow salary \geq 50K$, $prof(yes) \rightarrow salary \geq 50K$, and $age(<30) \rightarrow salary < 50K$. In terms of context

specific causal rules, CSCR-S and CSCR-Ass find quite similar results, while CSCR-3 captures more rules than other ones. CSCR-3 finds that $educ(\leq 9) \rightarrow salary < 50K$ is a causal rule under certain conditions, e.g. $age \geq 30$ and $hours \geq 30$, which is ignored by other two methods.

The commonly detected causal rules by three methods indicate that neither *sex* nor *working hours* is the cause of *income*, but under some certain conditions they are. Considering the causal rule $sex(male) \rightarrow salary \geq 50K|hours > 30$, $sex(male)$ alone does not necessarily lead to high salary, but under the condition of hard working (e.g. working more than 30 hours per week) a male deserves a good income. Another interesting finding is that less working hours per week ($hours < 30$) will cause a low annual income, although the person has a good education ($educ > 12$). That is, one having a good eduction will have a great chance to get high salary (i.e. the global causal rule $educ(> 12) \rightarrow salary \geq 50K$), however, even if one has a good education, low working hours can still lead to a low salary. This is an exception of the global causal rule $educ(> 12) \rightarrow salary \geq 50K$, and this exception is understandable. This result indicates that context specific causal rules can provide insights that are hidden by a global causal rule.

6.3.5 Comparison with the existing method

In this section, we compare the performance between our proposed framework and Causal Regression Tree (CRT) [18], an existing causal tree approach. Here we only compare CSCR-S with CRT on two real world data sets.

CRT is designed to estimate the heterogeneity of causal effects across subsets of the population. Specifically, it applies regression tree with a modified MSE (Mean Squared Error) criterion to partition the population into multi-

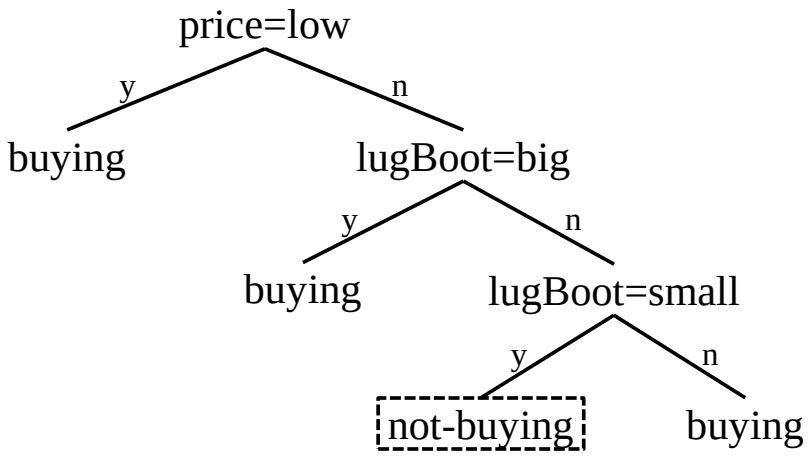


Figure 6.8: A CRT of the Car Evaluation data set, where $safety(med)$ is set as the treatment. The rule in the path from the root to the leaf with dashed box has a low causal effect.

ple subgroups. Different from CRCS-S, in a CRT, the treatment variable is a hypothesized cause of the target variable and the goal of this approach is to estimate the differences in causal effects across different sub-populations.

CRCS-S discovers both context free and context specific causal rules simultaneously, and does not need a specification of a cause in order to find context specific causal rules as CRT does. In the following, we will show that causal rules discovered by CRCS-S are shorter and more interpretable than those discovered by CRT.

We run experiments with the Car Evaluation data set to compare the results discovered by CRCS-S and CRT. Figure 6.8 shows a CRT built from the data set, where $safety(med)$ is assumed as the treatment variable. For example, the result on the root node of the tree can be interpreted as that $safety(med)$ has heterogeneous causal effects on $buying$, across the subgroups $price = low$ and $price \neq low$.

TABLE 6.6 compares context specific causal rules discovered by CRT and

Table 6.6: Comparison of causal rules discovered by CRT and CRCS-S from the Car Evaluation data set.

Cause	Context	Outcome	CRT	CRCS-S
$safety(med)$	$price = low$	$buying$	$\surd$	$\surd$
$safety(med)$	$passengers \geq 5$	$buying$		$\surd$
$safety(med)$	$price \neq low$ $lugBoot = big$	$buying$	$\surd$	$\surd^*$
$safety(med)$	$price \neq low$ $lugBoot = med$	$buying$	$\surd$	$\surd^*$
$safety(med)$	$passengers = 4$ $price = med$	$buying$		$\surd$
$safety(med)$	$passengers = 4$ $maint \neq high$	$buying$		$\surd$

* Which rule is implied by a more general causal rule discovered by CRCS-S.

CRCS-S from the Car Evaluation data set. We can see that all the results found by CRT are covered by either the results of CRCS-S (e.g. the first row in TABLE 6.6), or a more general causal rule detected by CRCS-S (e.g. CRCS-S finds $safety(med) \rightarrow buying|lugBoot \neq small$, which is a general causal rule of the third and forth rows in the table). CRT also misses three reasonable causal rules found by CRCS-S.

Another comparison is performed with the Adult data set, where $sex(male)$ is the treatment. CRT builds a large tree (with 178 leaves) on the data set, and we show the first three levels of the tree in Figure 6.9. We believe that a context more than 3 conditions is not useful since it is too specific and difficult to be interpreted. As the tree is pruned to the three levels, not all leaf nodes represent causal rules because of a low cause effect. We notice that $workclass = private$ is selected as the first branching variable, and this is not a good choice since it has a very low causal effect (-0.07) on the outcome $salary \geq 50K$. All leafs under it produce no context spe-

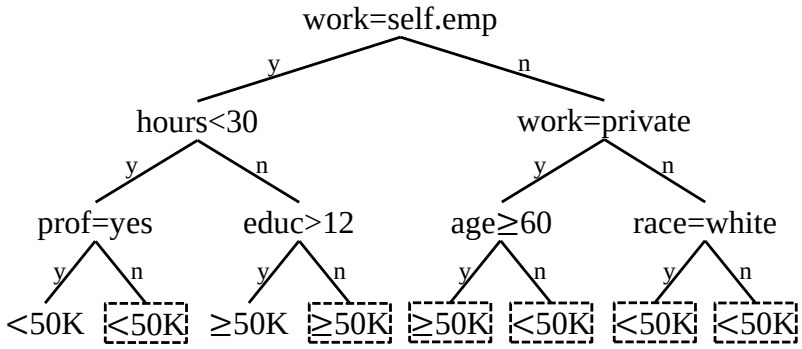


Figure 6.9: A CRT of the Adult data set, where $sex(male)$ is the treatment. The rule in the path from the root to the leaf with dashed box has a low causal effect. ($work$: working class, $self.emp$: self employment.)

cific causal rules. Then we also compare the context specific causal rules detected by CRT and CRCS-S in TABLE 6.7. CRT fails to find short context specific causal rules which are easily understandable. In contrast, the causal rules found by CRT are more complex and more difficult to be

Table 6.7: Comparison of causal rules discovered by CRT and CRCS-S from the Adult data set.

Cause	Context	Outcome	CRT	CRCS-S
$sex(male)$	$educ > 12$	$\geq 50K$		$\surd$
$sex(male)$	$prof = yes$	$\geq 50K$		$\surd$
$sex(male)$	$hours \geq 30$	$\geq 50K$		$\surd$
$sex(male)$	$age \geq 30$	$\geq 50K$		$\surd$
$sex(male)$	$work = self.emp$ $hours \geq 30$ $educ > 12$	$\geq 50K$	$\surd$	$\surd$ [*]
$sex(male)$	$work = self.emp$ $hours < 30$ $prof = yes$	$< 50K$	$\surd$	

[*] Which rule is implied by a more general causal rule discovered by CRCS-S.

interpreted than those discovered by CRCS-S. For example, $sex(male) \rightarrow salary \geq 50K|\{work = self.emp, hours \geq 30, educ > 12\}$ detected by CRT is much more complex than either $sex(male) \rightarrow salary \geq 50K|hours \geq 30$ or $sex(male) \rightarrow salary \geq 50K|educ > 12$ (both results are found by CRCS-S).

6.4 Conclusion

In this chapter, a novel framework, Context Specific Causal Rule discovery (CSCR) has been proposed to explore context specific causal relationships from observational data. Finding causes and contexts simultaneously is important but challenging. Data mining techniques are utilized to make the complex problem manageable. We have designed CSCR based on a well-known causal framework, the potential outcome model, to assess context specific causal relationships. CSCR is instantiated in three ways, based on decision tree, multiple decision trees and association rules.

The experiment results show that CSCR can achieve a high accuracy with the synthetic data sets and find insights from real world data sets. CSCR also outperforms an existing causal tree method, CRT in terms of the exploration of short and meaningful context specific causal relationships and easy operation without specifying a candidate cause. CSCR provides a scalable and automated way to address the increasing need of context specific causal relationships for personalized decision making.

Chapter 7

Causal decision trees

In previous chapters, we have discussed several different types of causal relationships involving multiple factors, and developed efficient methods for discovering and examining the causal relationships from observational data. These methods aim to detect complex causal relationships with multiple factors involved, such that people can gain better insights into the data and understand the causal relationships between random variables. In this chapter, we will discuss how to effectively apply causal knowledge captured from data for correct decision making and timely intervening in processes at risk. To this end, we propose a causality based classification method, Causal Decision Tree (CDT), where each node of the tree has causal interpretations with respect to the outcome variable.

Traditional classification methods are capable of efficiently finding relationships between random variables from data and have the potential to make good predictions. Decision trees [36] are a good example of the classification methods, and they have been widely used for decision making in many areas, including social and medical data analysis. A traditional classification method offering high classification accuracy, however, 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.

In contrast, a CDT is designed to detect real causal relationships with an outcome, and also expect reliable classification accuracy (although it may not achieve optimal classification accuracy). The major difference between CDT and traditional classification methods is that the former assesses the relationship between a (potential) cause variable and an outcome variable by removing the effects of their covariates on the outcome but the latter only considers the relationship between the potential cause variable and the outcome in isolation.

This chapter is based on the work published in [35].

7.1 Problem statement and definitions

Let $\boldsymbol{X} = \{X_1, X_2, \ldots, X_m\}$ 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 $\boldsymbol{D}$ contains n records taking various assignments of values for $\boldsymbol{X}$ and Y, each of which represents the record of an observation. Let us assume that $\boldsymbol{X}$ includes all the attributes for characterising an individual, and the data set is large and thus there is no bias in the sampling process.

7.1.1 Why a decision tree may not encode causality?

Decision trees [36] are a popular classification model, with two types of nodes: branching and leaf nodes. A branching node represents a predic-

tor 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 [36]. 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 $1 - \max(Pr(Y = 1|X), Pr(Y = 0|X))$ [36], where $Pr(Y = 1|X)$ and $Pr(Y = 0|X)$ are the fractions of positive and negative samples respectively given X. In practice, other measures such as entropy and Gini index are often used [36], 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 $(Pr(Y = 1|X) - Pr(Y = 0|X))$, we have the following conceptual definition of a discriminative or branching attribute.

Definition 7.1 (Discriminative attribute) *Given a data set $\boldsymbol{D}'$, a discriminative attribute is the attribute X_i such that $|Pr(Y = 1|X_i = 1) - Pr(Y = 0|X_i = 1)|$ is maximised.*

Note that $\boldsymbol{D}'$ is a subset (of $\boldsymbol{D}$) 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 7.2 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 $Pr(Y = 1|X_i = 1)$–$Pr(Y = 0|X_i = 1)$) is different from that of a causal factor (having significant causal effect $Pr(Y = 1|X_i = 1)$–$Pr(Y = 1|X_i = 0)$).

Secondly, the estimation of $Pr(Y = 1|X_i = 1)$–$Pr(Y = 0|X_i = 1)$ for choosing a discriminative attribute is based on the data set $\boldsymbol{D'}$, while the estimation of the causal effect of X_i on Y is based on the stratified data set $\boldsymbol{D}_{\boldsymbol{C}=\boldsymbol{c}_k}$ 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 ($Pr(Y = 1|X_i = 1)$–$Pr(Y = 1|X_i = 0)$) reflects the causal effect of X_i on Y.

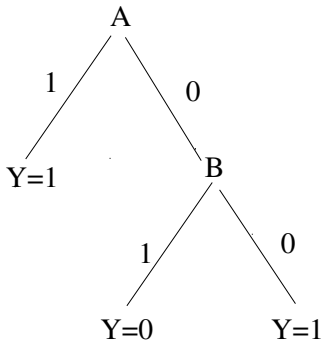
Essentially the main limitation of a decision tree is that it does not consider other attributes in determining a branching attribute. The choice of a branching attribute does not rely on the causal effect of the attribute on the outcome attribute.

Following Observation 3.1, the estimation of causal effect should be based on stratified data where the difference of individuals in a stratum is eliminated.

Now we analyze that the perfect classification decision tree does not code causal relationships, by using an example shown in Figure 7.1.

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)

Figure 7.1: 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

Example 7.1 *In Figure 7.1, the path* $(A=1) \rightarrow (Y=1)$*, with the difference in probabilities,* $(Pr(Y=1|A=1) - Pr(Y=0|A=1)) = 1$*, 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 7.1 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 ($B = 0$, $C = 1$), $E_{\boldsymbol{D}}[Y^1|A = 1] - E_{\boldsymbol{D}}[Y^1|A = 0] = 1 - 1 = 0$. 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 summarized 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.

7.1.2 Problem statement

Our aim is to build a causal decision tree 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 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 7.2 (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 7.3 (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* $(\boldsymbol{D}|\boldsymbol{P} = \boldsymbol{p})$*.*

Note that both Definitions 6.2 and 7.3 are developed to define the case that causal relationships only hold under specific contexts. The only difference between them is that different criteria have been employed to examine the causal relationships, respectively.

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 $\boldsymbol{P}$ 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 7.3, the causal relationship between the root 'age<30' and the outcome '>50K' is context free, while the causal relationship between 'education-num>12' and the outcome is in the context of 'age<30', which is a refinement of the causal relationship 'age<30' (the parent of 'education-num>12') to the outcome and is a more specific relationship.

Definition 7.4 (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.*

Then the research problem can be formally expressed as follows.

Problem 7.1 *Given a data set $\boldsymbol{D}$ with a set of predictor variables $\boldsymbol{X}$ and the target variable Y, build a causality based classifier based on Definition 7.4, by capturing global or context specific causal relationships between $\boldsymbol{X}$ and Y in $\boldsymbol{D}$.*

7.2 Causal decision trees

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 chapter 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). Before that, we firstly present a measure which is used in both two CDT algorithms for the causal effect estimation. At the end of this section, we will also analyze the time complexity of these two CDT algorithms.

7.2.1 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 $\boldsymbol{X}\backslash\{Q\}$ 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 [128] is suitable for measuring the difference of two outcomes. Let the following table summarize the statistics of the k-th stratum, $\boldsymbol{c}_k$.

$\boldsymbol{c}_k$	$Y=1$	$Y=0$	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 [140] is a means to achieve this. Over all the r strata of a data set, the difference can be summarized as:

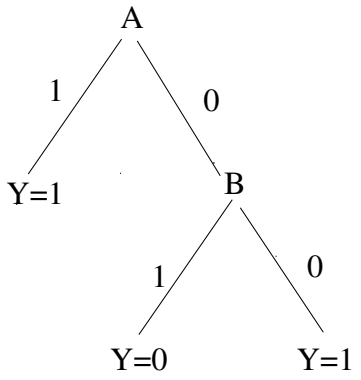
$$\mathrm{PAMH}(Q,Y)=\frac{(|\sum_{k=1}^{r}\frac{n_{11k}n_{22k}-n_{21k}n_{12k}}{n_{..k}}|-\frac{1}{2})^2}{\sum_{k=1}^{r}\frac{n_{1.k}n_{2.k}n_{.1k}n_{.2k}}{n_{..k}^2(n_{..k}-1)}} \tag{7.1}$$

This is the test statistic of the Mantel-Haenszel test [140, 141]. The test is used to find direct causal relationship between two variables [140]. The test statistic has a Chi-square distribution (degree of freedom=1). Given a significance level η, if $\mathrm{PAMH}(Q,Y)\geq\chi^2_\eta$, 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 7.2 in the following.

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

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)

Figure 7.2: An example showing that a CDT represents causal relationships (a) An exemplar data set. (b) a CDT of the data

Example 7.2 *From the data set shown in Figure 7.2, 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 (7.1)), $\mathrm{PAMH}(A, Y)$*. 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 $\text{PAMH}(A, Y) = 17.5$. *For* $\eta = 0.05$ *or* $\chi^2_\eta = 3.84$, *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 $\text{PAMH}(B,Y) = 49 > 3.84$ *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.*

7.2.2 CDT by perfect stratification

As shown in ALGORITHM 7.1, CDT-PS takes 3 inputs: the data set $\boldsymbol{D}$ for a set of predictor attributes $\boldsymbol{X}$ 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 7.1 CDT with Perfect Stratification (**CDT-PS**) and CDT with Stratification on Propensity Scores (**CDT-SPS**)

Input: $\boldsymbol{D}$, a data set for the set of predictor attributes $\boldsymbol{X} = \{X_1, X_2, \ldots, X_m\}$ and the outcome attribute Y; $h_{\max}$, the maximum height the tree; η, significance level for the Mantel-Haenszel (partial association) test and correlation test.

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

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

ALGORITHM 7.1 firstly initiates the CDT (i.e. $\boldsymbol{T}$), 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 (ALGORITHM 7.2) 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 $\boldsymbol{Z} \subset \boldsymbol{X}$ to expand the current subtree (whose root is N) and $\boldsymbol{Z}$ contains only the attributes that have not been used in the tree; the context specific data set $\boldsymbol{D}'$, 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 $\boldsymbol{Z}$ 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 $(\boldsymbol{D}'|1)$ and $(\boldsymbol{D}'|0)$ respectively.

ALGORITHM 7.2 TreeConstruct(N, $\boldsymbol{Z}$, $\boldsymbol{D'}$, h, e)

1: **if** $\boldsymbol{Z} == \emptyset$ 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 $(\boldsymbol{D'}|N = e)$
3: return
4: **end if**
5: find a set of attributes in $\boldsymbol{Z}$ that are correlated with Y in $\boldsymbol{D'}$
6: **for** each correlated attribute X_i **do**
7: calculate propensity scores for each sample of $\boldsymbol{D'}$ given the stratifying attributes, $\boldsymbol{Z}\backslash\{X_i\}$ // for CDT-SPS only
8: Stratify samples via QuickSort and specific grouping criterion // for CDT-PS grouping according to the values of $\boldsymbol{Z}\backslash\{X_i\}$; for CDT-SPS, grouping according to propensity scores and the given number of subgroups, r
9: compute PAMH(X_i, Y) in stratified $\boldsymbol{D'}$
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 $(\boldsymbol{D'}|N = e)$
14: return
15: **end if**
16: **if** $e == null$ **then**
17: let node W be the root of $\boldsymbol{T}$
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 $\boldsymbol{Z}$
22: **for** each $w \in \{0, 1\}$ **do**
23: call TreeConstruct(W, $\boldsymbol{Z}$, $(\boldsymbol{D'}|W = w)$, h, w)
24: **end for**

ALGORITHM 7.3 TreePruning($\boldsymbol{T}$)

1: **for** each leaf in $\boldsymbol{T}$ **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**

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

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

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 $(\boldsymbol{D}'|1)$ and $(\boldsymbol{D}'|0)$ 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 $\boldsymbol{T}$; 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 $\boldsymbol{Z}$ 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 $(\boldsymbol{D}'|W = w)$ where $w \in \{0, 1\}$.

ALGORITHM 7.3 prunes leaves that do not have distinct labels. The func-

tion 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.

7.2.3 CDT by stratification on propensity score

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 [33, 107, 111], 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. Since within a stratum the probability of each individual receiving the treatment is similar, we can still follow Observation 3.1 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 (7.1) 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 [33, 142]. 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 [107, 143]. The use of 5 to 10 subgroups is a current convention [111].

As can be seen from ALGORITHMs 7.1, 7.2 and 7.3, CDT-SPS essentially follows the same steps as CDT-PS except that:

1. CDT-SPS requires an extra input, r (see line 8 in ALGORITHM 7.2), 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 ($\boldsymbol{Z} \setminus \{X_i\}$) into the same strata using QuickSort (see Line 8 of ALGORITHM 7.2), CDT-SPS firstly calculates the propensity scores of each sample (Line 7 of ALGORITHM 7.2) 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 $\boldsymbol{Z} \setminus \{X_i\}$.

7.2.4 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 $\boldsymbol{D}'$ 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 $O(n^\eta)$ where $2 < \eta \leq 3$ depending on different optimisers [144]. The overall time complexity for CDT-SPS is $O(mn \log n + mn^\eta) = O(mn^\eta)$.

CDT-SPS does not scale well as CDT-PS does. The detailed comparisons regarding scalability are given in Section 7.3.4. However, the efficiency of CDT-SPS can be improved based on the research for fast estimation of propensity scores [105].

7.3 Experiments

To evaluate the CDT algorithms, CDT-PS and CDT-SPS, firstly in Section 7.3.1, 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 the normal decision trees. The normal decision trees are built using the C4.5 algorithm [36] implemented in Weka [139] 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 7.3.2, 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 [13], a commonly used Bayesian network learning algorithm.

Another set of experiments with 10 real world data sets are carried out in Section 7.3.3 to compare the prediction performance (i.e. classification accuracy) of the CDTs and the C4.5 trees.

Section 7.3.4 evaluates the scalability of CDT-PS and CDT-SPS and compares 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 7.3.3 where the heights of the CDTs are not limited.

Table 7.1: Summary of 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

7.3.1 CDTs find meaningful causal relationships

Now 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.

7.3.1.1 Adult data set - census income

The Adult data set (Table 7.1) was retrieved from the UCI Machine Learning Repository [118] and it is an extraction of 1994 USA census database.

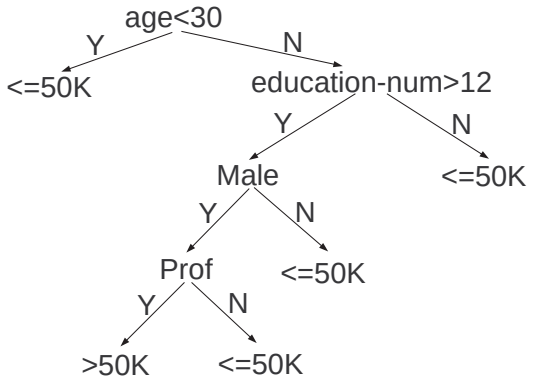


Figure 7.3: CDTs of the Adult data set

It is a well known classification data set used in predicting whether a person earns over 50K or not in a year. 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 7.3) with the Adult data set, and the normal decision tree built using the data is shown in Figure 7.4.

From Figure 7.4, 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. We should note that the major objective of CDTs is for better (causal) interpretation, although we still expect that CDTs have good classification accuracy for prediction. Instead . Hence smaller CDTs are preferred.

The next observation is crucial to show the difference between 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

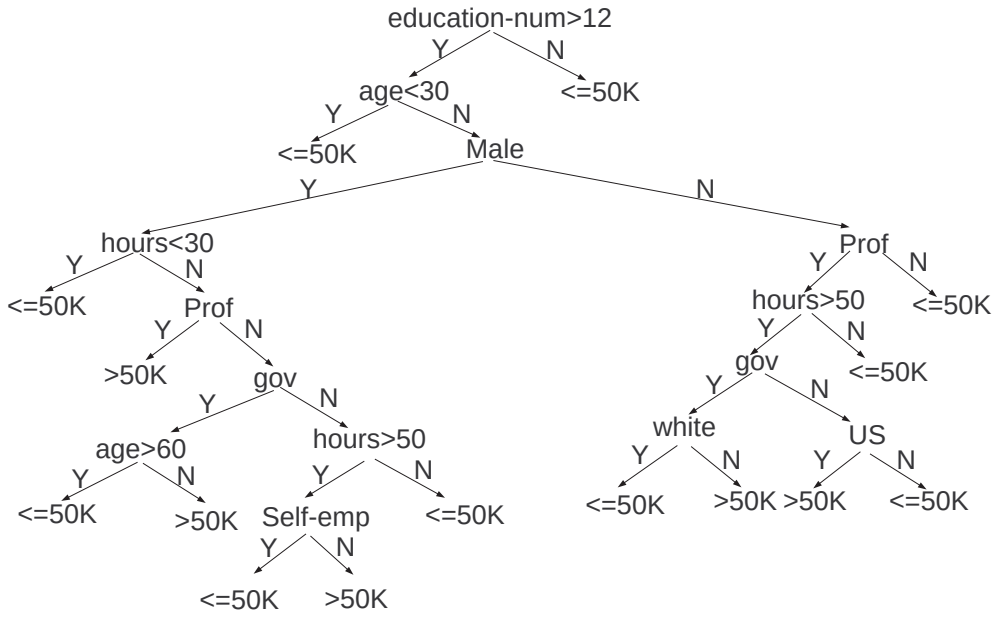


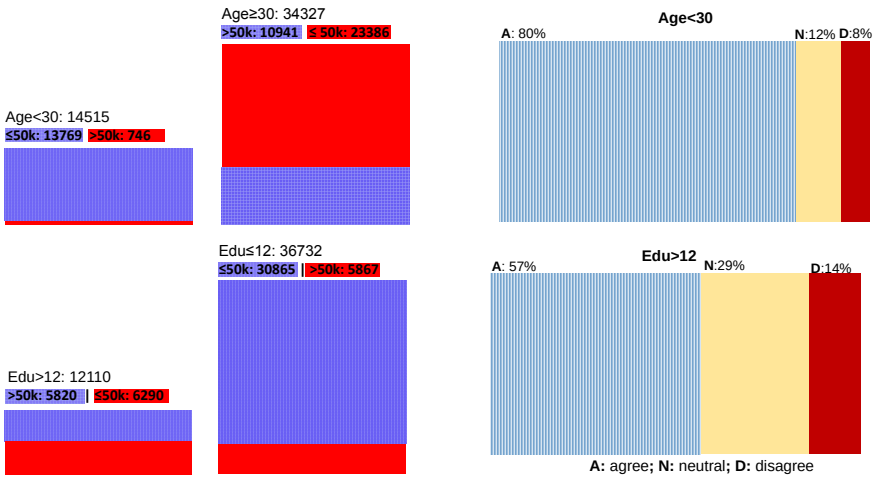
Figure 7.4: A C4.5 decision tree of Adult data set

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$\leq$12	5867	30865
counts	$> 50K$	$\leq 50K$
Age$\geq$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 at-



(a) Sample distribution (b) Strata distribution

Figure 7.5: An illustration of the different choices of a splitting attribute between a normal decision tree and a causal decision tree

tribute 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 [145, 146], which a classification criterion fails to capture.

A visual illustration of the above discussions is shown in Figure 7.5. To minimise classification errors (indicated by the red areas in Figure 7.5 (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 7.5 (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.

7.3.1.2 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 [147]. The description of the data is shown in Table 7.2. 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 7.6 (left), the CDT has captured the operational mecha-

Table 7.2: Summary of Ultra Short Stay Unit 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

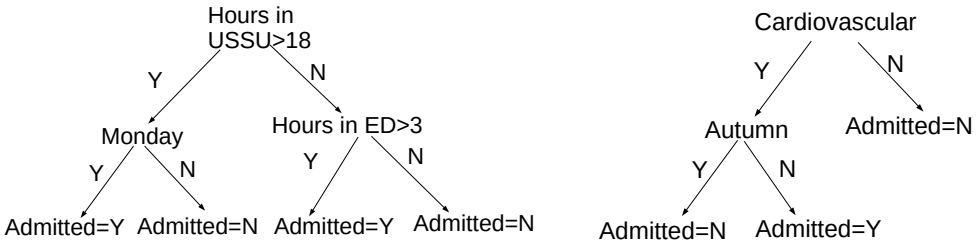


Figure 7.6: CDT (left) and C4.5 tree (right) of the USSU data. The labels of a CDT leaf indicate the proportionally larger number of instances instead of the majority of instances. The distribution of the classes is so skewed and the majority of instances is "Admitted=N" since its domination

nisms 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 7.6, 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 words, 'Cardiovascular disease → 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 → 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 7.6 (right) have 23, 62, and 3428 instances respectively. In contrast, the leaves of the CDT in Figure 7.6 (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.

7.3.1.3 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 randomized 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 7.7.

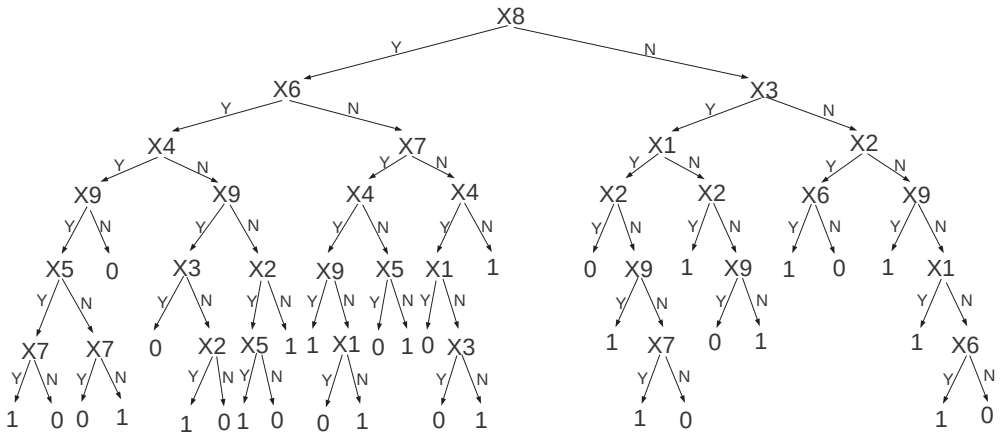


Figure 7.7: A C4.5 decision tree of a randomly generated data set

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.

7.3.2 CDT identifies causal relationships

7.3.2.1 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 [13] 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 speci-

Table 7.3: 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

fied 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.

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 7.3 (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.

7.3.2.2 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 [64] (also known as PC-Simple) 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 7.3 (Part B) summarizes 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 chapter can find context specific causal relationships in the complete data sets.

7.3.3 CDTs for classification

A major goal of CDTs is to discover and represent 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

Table 7.4: 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**

can be classifiers with good interpretability and have reliable prediction performance.

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 7.4), 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++ [119]. 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. Moreover, we do

not set height limit to the CDTs to make them comparable to C4.5 trees. Table 7.4 summarizes 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.

As expected, from Table 7.4, 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.

7.3.4 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 [139] and the PC algorithm [13]. We use 12 synthetic data sets generated with the same procedure as described in Section 7.3.2.1. 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 Figures 7.8 and 7.9. 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

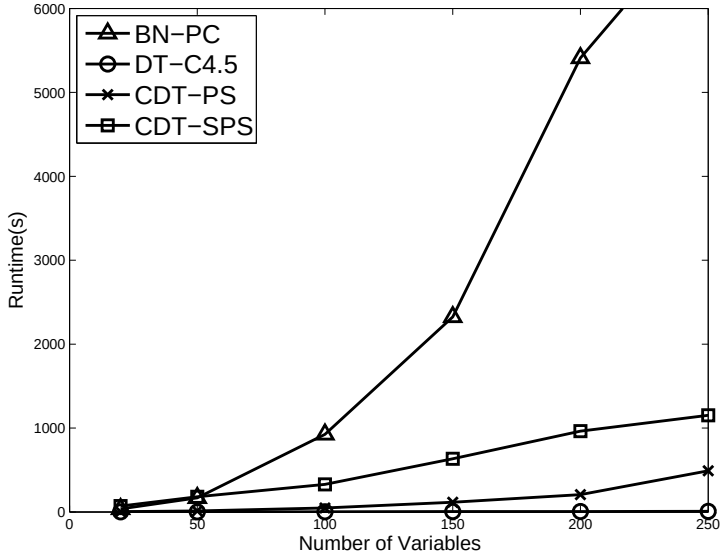


Figure 7.8: The scalability of CDT with the number of variables in comparison to C4.5 and PC

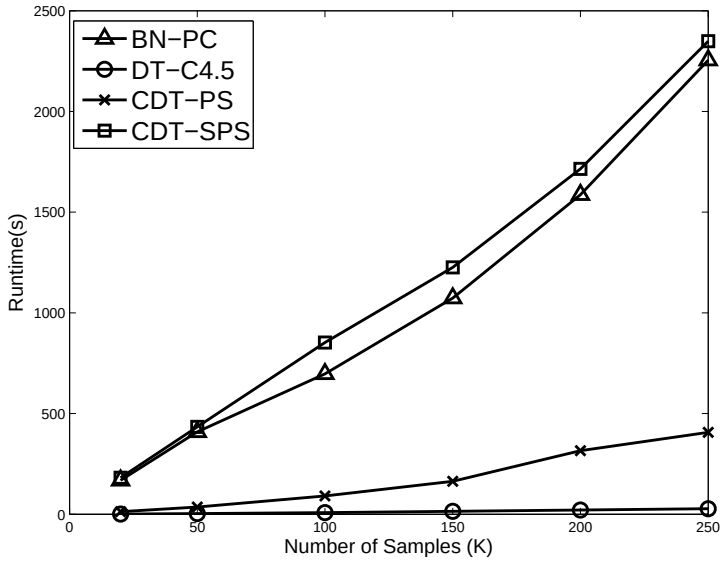


Figure 7.9: The scalability of CDT with the number of samples in comparison to C4.5 and PC

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.

7.4 Conclusion

In this chapter, we have proposed causal decision trees, a novel causality based classification model for prediction and actions by capturing causal relationships in data.

A CDT has a reliable classification accuracy and also 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 most existing causal models.

The algorithms developed for constructing a CDT utilizes 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 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.

Chapter 8

Conclusions

In the past decades, causal discovery from observational data has been studied widely. However, most existing research has focused on detecting causal relationship between a single factor and the outcome, which may not be sufficient for reasoning about the cause of a particular effect. In this thesis, we study complex causal relationships with multiple factors involved, and have developed the methods for discovering combined causes, cause interactions and context specific causes from data. We have also developed a causality based classification model for predictions with causal interpretations.

From data mining perspective, we aim to propose efficient methods for automated discovery of the complex causal relationships in large data sets. Moreover, we do not make any assumptions about the causal relationships between random variables. To address the computational complexity of the causal discovery, we take the advantages of highly efficient data mining techniques when developing methods.

The proposed method for mining combined causes, MH-PC, is able to explore the data to identify combined causes, which comprise multiple com-

ponent variables leading to an outcome together, but not individually. We utilize the idea of a well-known local causal discovery method, HITON-PC, to discover the combined causes. To improve the efficiency of MH-PC, we introduce the concept of redundant combined variables and carefully design the procedure of combined variable generation to reduce the search space of combined causes. A strategy is proposed to avoid redundant conditional independence tests for further efficiency improvement. The results show that MH-PC can achieve high accuracy with synthetic data sets.

A general framework, DACID, has also been developed to detect the cause interactions between multiple causal factors. A conceptual model, the sufficient-component cause model, is employed to investigate the interactions between the causal factors. The potential outcome model is used to test if the interactions between the predictor variables is causal with respect to the target variable, by removing the effects of their covariates. The instantiated algorithms from DACID search all pairs or sets (with more than two variables) of variables for the cause interaction detection. The algorithms have been applied to multiple synthetic and real world data sets, and the results have demonstrated the good performance of the proposed algorithms.

The method developed for context specific causal relationship exploration, CSCR, is capable of efficiently detecting causal relationships which hold only in certain contexts. Different from existing methods, we do not make any assumptions about the causal relationships between predictor variables and an outcome. In other words, CSCR has been designed to detect both contexts and the causal relationships under the contexts simultaneously. It is very challenging for finding context specific causal relationships directly from data, due to the computational complexity. To address this issue, data mining technique is employed to improve the computational efficiency. Experiments have shown that CSCR outperforms the existing causal tree method [18] in detecting short and meaningful context specific causal re-

lationships. More importantly, comparing to the existing methods, CSCR does not require the prior knowledge of a potential cause.

We have also built a causality based classification method, CDT, for precise and causally interpretable predictions. Instead of directly making use of the decision tree algorithm for candidate generation as CSCR, CDT utilizes a causal criterion in the construction of the tree model. The stratification approach enables the method to exclude many spurious relationships that are definitely not causal from the resulting tree. Thus practitioners will have a smaller set of quality causal hypotheses for the further study. The experiments show that CDT can achieve similar classification accuracy as the C4.5 trees. Meanwhile, due to the causal semantics of the predictions made by CDTs, they are more useful for supporting decision making.

Although huge efforts have been made on causal discovery and inference, limitations exist in current methods, including the proposed methods in this thesis. A key point of causality based methods is about how to find confounding variables and then to adjust the bias arisen by them. In practice, researchers normally pick up confounding variables based on domain knowledge. A theoretically complete solution to the adjustment problem is the back-door and front-door criteria [4] developed by Pearl, which assume that a causal structure is known in advance. However, this is impractical in many real world applications since we do not know the structure. Thus it is still an open issue to identify confounding variables from observational data. Besides, as introduced in Chapter 3, some assumptions (e.g. the causal sufficiency and the faithfulness assumption) are required to enable the discovery of causal relationships from observational data, which means that the methods may not work well, when the assumptions are violated.

Nonetheless, the work conducted in this thesis has launched the first attempt at a practical problem of finding causal relationships involving multiple factors from data. We have also proposed efficient and automated methods

for capturing such complex causal relationships. The proposed methods have the potential to provide good insights from observational data for predictions and decision making.

In the following, we identify some future directions to extend the work in this thesis:

1. Develop a unified model to define different types of causal relationships involving multiple factors

   The methods developed in this thesis aim to capture different types of causal relationships involving multiple factors, including combined causes, cause interactions and context specific causes. We have introduced the concepts of these relationships and discussed the difference between them. It is desirable to develop a unified theory to define these types of causal relationships. With this theory, a possible future work is to develop an efficient method to capture all these relationships from observational data.

2. Develop a method to discover causal relationships involving multiple factors without causal sufficiency

   As mentioned in Chapter 3, a common assumption of causal discovery is to assume that there are no hidden common causes to bias the causal analysis (i.e. the causal sufficiency assumption). The methods in this thesis also follow this assumption that all common causes are observed and included in data. However, in many cases (data sets) causal sufficiency may be violated. The potential presence of hidden common causes varies the dependency between random variables and the orientation. Thus it is required to develop new concepts and algorithms to deal with observational data without assuming the causal sufficiency.

   On the other hand, the collection of observational data from different sources and/or of different types has increased dramatically in recent

decades. An important fact is that hidden common causes may exist in one observational data but not in another one. Thus a possible further work is to combine the different results learned from different source/type of observational data into an integrated study. It is expected to minimize the impact of hidden common causes.

3. Estimate causal effects from categorical data

   In the thesis, the proposed methods are for learning (local) causal structures between variable, while, in many application areas (including medicine, epidemiology, and social sciences), it is required to quantify the causal effects of predictor variables on an outcome, not just to infer the causal relationships. Most existing methods assume a known structure and focus on estimating causal effects based on the structure. It is desirable to quantify the causal effects from data directly. However, causal effect estimation is very challenging and two main problems must be addressed: (1) establishing a causal structure/hypothesis between the variables of interest that holds in data; and (2) estimating the causal effects of a cause variable on its effect variables based on the causal hypothesis.

   The IDA (Intervention effect when the DAG is absent) method [148, 149] is the pioneer work on estimating causal effect directly from data. However, IDA only deals with data following Gaussian distribution. Thus a possible future work is to estimate causal effects directly from categorical data.

4. Apply the proposed methods to real world problems

   In this thesis, we have proposed several methods for discovering causal relationships involving multiple factors. The experiments on multiple real world data sets, including clinical data and genetic data, have shown that the proposed methods have the potential to gain good insights from observational data. Thus an intuitive further work is to apply the proposed methods to solve real world problems, such as

identifying the co-regulation mechanisms of multiple types (instead of just one type) of gene regulators.

References

[1] David J. Hunter. **Gene-environment interactions in human diseases**. *Nature Reviews. Genetics*, **6**(4):287–298, April 2005. v, 4, 6, 7, 62

[2] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011. 1, 39

[3] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. **Mining association rules between sets of items in large databases**. In *ACM SIGMOD record*, **22**, pages 207–216. ACM, 1993. 1, 101, 110

[4] Judea Pearl. *Causality: models, reasoning and inference*. Cambridge Univ Press, Cambridge, 2000. 2, 3, 4, 13, 164

[5] Marcia L. Meldrum. **A brief history of the randomized controlled trial: from oranges and lemons to the gold standard**. *Hematology/Oncology Clinics of North America*, **14**(4):745–760, 2000. 3, 12

[6] Sir Ronald Aylmer Fisher, Ronald Aylmer Fisher, Statistiker Genetiker, Ronald Aylmer Fisher, Statistician Genetician, Ronald Aylmer Fisher, and Statisticien Gnticien. *The design of experiments*, **12**. Oliver and Boyd Edinburgh, 1960. 3, 12

[7] Donald B Rubin. **The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials**. *Statistics in Medicine*, **26**(1):20–36, 2007. 3, 12

[8] Paul R Rosenbaum. *Design of observational studies*. Springer series in statistics. Springer, New York, 2010. 3, 12

[9] Edward J Boyko. **Observational researchopportunities and limitations**. *Journal of Diabetes and its Complications*, **27**(6):642–648, 2013. 3, 12

[10] John Concato, Nirav Shah, and Ralph I Horwitz. **Randomized, controlled trials, observational studies, and the hierarchy of research designs**. *New England Journal of Medicine*, **342**(25):1887–1892, 2000. 3

[11] Carl V Phillips and Karen J Goodman. **The missed lessons of Sir Austin Bradford Hill**. *Epidemiologic Perspectives & Innovations*, **1**(1):3, 2004. 3

[12] Kenneth J Rothman and Sander Greenland. **Causation and causal inference in epidemiology**. *Journal Information*, **95**(S1):144–150, 2005. 3, 66, 67

[13] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. The MIT Press, Cambridge, 2000. 4, 13, 15, 16, 30, 31, 32, 42, 50, 145, 154, 159

[14] Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. **Local causal and markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation**. *The Journal of Machine Learning Research*, **11**:171–234, 2010. 4, 6, 19, 35, 43, 50

[15] Laura R. Novick and Patricia W. Cheng. **Assessing interactive causal influence.** *Psychological Review*, **111**(2):455, 2004. 4, 22

[16] Gnter P. Wagner, Mihaela Pavlicev, and James M. Cheverud. **The road to modularity**. *Nature Reviews Genetics*, **8**(12):921–931, 2007. 5

[17] Patrik Dhaeseleer, Shoudan Liang, and Roland Somogyi. **Genetic network inference: from co-expression clustering to reverse engineering**. *Bioinformatics*, **16**(8):707–726, 2000. 5

[18] Susan Athey and Guido Imbens. **Recursive partitioning for heterogeneous causal effects**. *Proceedings of the National Academy of Sciences*, **113**(27):7353–7360, July 2016. 5, 6, 27, 99, 122, 163

[19] Michael J. Duffy and John Crown. **A personalized approach to cancer treatment: how biomarkers can help**. *Clinical Chemistry*, **54**(11):1770–1779, November 2008. 5, 97

[20] Zhou Jin, Jiuyong Li, Lin Liu, Thuc Duy Le, Bingyu Sun, and Rujing Wang. **Discovery of causal rules using partial association**. In *IEEE 12th International Conference on Data Mining*, pages 309–318. IEEE, 2012. 6, 21, 55

[21] Jiuyong Li, Thuc Duy Le, Lin Liu, Jixue Liu, Zhou Jin, and Bingyu Sun. **Mining causal association rules**. In *First IEEE ICDM Workshop on Causal Discovery*, 2013. 6, 21

[22] Jiuyong Li, Thuc Duy Le, Lin Liu, Jixue Liu, Zhou Jin, Bingyu Sun, and Saisai Ma. **From observational studies to causal rule mining**. *ACM Transactions on Intelligent Systems and Technology (TIST)*, **7**(2):14, 2015. 6, 21, 55

[23] Pranjul Yadav, Lisiane Prunelli, Alexander Hoff, Michael Steinbach, Bonnie Westra, Vipin Kumar, and Gyorgy Simon. **Causal inference in observational data**. *arXiv preprint arXiv:1611.04660*, 2016. 6, 22

[24] Marco Bonetti and Richard D. Gelber. **Patterns of treatment effects in subsets of patients in clinical trials**. *Biostatistics (Oxford, England)*, **5**(3):465–481, July 2004. 6, 27

[25] Kosuke Imai, Marc Ratkovic, and others. **Estimating treatment effect heterogeneity in randomized program evaluation**. *The Annals of Applied Statistics*, **7**(1):443–470, 2013. 6, 28

[26] Saisai Ma, Jiuyong Li, Lin Liu, and Thuc Duy Le. **Mining combined causes in large data sets**. *Knowledge-Based Systems*, **92**:104–111, January 2016. 6, 43

[27] R. A. Weideman, I. H. Bernstein, and W. P. McKinney. **Pharmacist recognition of potential drug interactions**. *American Journal of Health-System Pharmacy*, **56**(15):1524–1529, August 1999. 7

[28] F. D. K. Liddell. **The interaction of asbestos and smoking in lung cancer**. *The Annals of Occupational Hygiene*, **45**(5):341–356, 2001. 7, 62

[29] K. S. Vimaleswaran, C. Power, and E. Hyppnen. **Interaction between vitamin D receptor gene polymorphisms and 25-hydroxyvitamin D concentrations on metabolic and cardiovascular disease outcomes**. *Diabetes & Metabolism*, **40**(5):386–389, November 2014. 7, 62

[30] Tyler J. VanderWeele and James M. Robins. **A theory of sufficient cause interactions**. *COBRA Preprint Series*, page 13, 2006. 7, 23, 63, 64, 67, 68, 72, 73

[31] Tyler J. Vanderweele and James M. Robins. **Empirical and counterfactual conditions for sufficient cause interactions**. *Biometrika*, **95**(1):49–61, 2008. 7, 23, 63, 68, 79

[32] Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash. *Modern epidemiology*. Lippincott Williams & Wilkins, 2008. 7, 63, 64, 65, 66, 67

[33] Paul R Rosenbaum and Donald B Rubin. **The central role of the propensity score in observational studies for causal effects**. *Biometrika*, **70**(1):41–55, 1983. 8, 12, 35, 39, 77, 98, 99, 102, 105, 142, 143

[34] Jasjeet S. Sekhon. **The Neyman-Rubin model of causal inference and estimation via matching methods**. *The Oxford Handbook of Political Methodology*, pages 271–299, 2008. 8, 12, 35, 98, 99, 102

[35] J. Li, S. Ma, T. Le, L. Liu, and J. Liu. **Causal decision trees**. *IEEE Transactions on Knowledge and Data Engineering*, **PP**(99):1–14, 2016. 8, 121, 128

[36] Pang-Ning Tan and others. *Introduction to data mining*. Pearson Education India, 2006. 8, 127, 128, 129, 145

[37] Fredrik Johansson, Uri Shalit, and David Sontag. **Learning representations for counterfactual inference**. In *International Conference on Machine Learning*, pages 3020–3029, 2016. 12, 35, 41

[38] Donald B. Rubin. **Estimating causal effects of treatments in randomized and nonrandomized studies.** *Journal of educational Psychology*, **66**(5):688, 1974. 12, 35

[39] Judea Pearl. *Causality*. Cambridge university press, 2009. 12

[40] Gregory F Cooper and Edward Herskovits. **A Bayesian method for the induction of probabilistic networks from data**. *Machine Learning*, **9**(4):309–347, 1992. 14

[41] Remco R Bouckaert. **Probabilistic network construction using the minimum description length principle**. In *Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 41–48. Springer, 1993. 14

[42] Wai Lam and Fahiem Bacchus. **Using causal information and local measures to learn Bayesian networks**. pages 243–250. Morgan Kaufmann Publishers Inc., 1993. 14

[43] Pedro Larranaga, Cindy MH Kuijpers, Roberto H Murga, and Yosu Yurramendi. **Learning Bayesian network structures by searching for the best ordering with genetic algorithms**. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, **26**(4):487–493, 1996. 14

[44] William H Hsu, Haipeng Guo, Benjamin B Perry, and Julie A Stilson. **A permutation genetic algorithm for variable ordering in learning Bayesian networks from data**. In *Proceedings of the Genetic and Evolutionary Computation Conference*, **2**, pages 383–390, 2002. 15

[45] Man Leung Wong, Shing Yan Lee, and Kwong Sak Leung. **A hybrid approach to discover Bayesian networks from databases using evolutionary programming**. In *Proceedings of IEEE International Conference on Data Mining*, pages 498–505. IEEE, 2002. 15

[46] Rosa Blanco, Inaki Inza, and Pedro Larranaga. **Learning Bayesian networks in the space of structures by estimation of distribution algorithms**. *International journal of intelligent systems*, **18**(2):205–220, 2003. 15

[47] Ruben Armaanzas Arnedillo, Iaki Inza Cano, Roberto Santana, Yvan Saeys, Jose Luis Flores, Jose Antonio Lozano, Yves Van de Peer, Rosa Blanco, Vctor Robles Forcada, Maria Concepcion Bielza Lozoya, and others. **A review of estimation of distribution algorithms in bioinformatics**. *BioData Mining*, 2008. 15

[48] Nevin Lianwen Zhang and David Poole. **Exploiting causal independence in Bayesian network inference**. *Journal of Artificial Intelligence Research*, **5**:301–328, 1996. 15

[49] Dan Geiger, Azaria Paz, and Judea Pearl. *Learning causal trees from dependence information*. UCLA, Computer Science Department, 1991. 15

[50] Robert M Fung and Stuart L Crawford. **Constructor: A system for the induction of probabilistic models**. In *Proceedings of the eighth National conference on Artificial intelligence*, **90**, pages 762–769, 1990. 15

[51] Diego Colombo and Marloes H. Maathuis. **Order-independent constraint-based causal structure learning.** *Journal of Machine Learning Research*, **15**(1):3741–3782, 2014. 16

[52] Judea Pearl and Thomas Verma. *A formal theory of inductive causation*. University of California (Los Angeles). Computer Science Department, 1991. 16

[53] Judea Pearl and Thomas S Verma. **A theory of inferred causation**. *Studies in Logic and the Foundations of Mathematics*, **134**:789–811, 1995. 16

[54] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. **Learning high-dimensional directed acyclic graphs with latent and selection variables**. *The Annals of Statistics*, **40**(1):294–321, 2012. 16

[55] Craig Silverstein, Sergey Brin, Rajeev Motwani, and Jeff Ullman. **Scalable techniques for mining causal structures**. *Data Mining and Knowledge Discovery*, **4**(2-3):163–192, 2000. 17

[56] Subramani Mani, Peter L Spirtes, and Gregory F Cooper. **A theoretical study of Y structures for causal discovery**. *arXiv preprint arXiv:1206.6853*, 2012. 17

[57] Gregory F Cooper. **A simple constraint-based algorithm for efficiently mining observational databases for causal relationships**. *Data Mining and Knowledge Discovery*, **1**(2):203–224, 1997. 17

[58] Ioannis Tsamardinos, Constantin F. Aliferis, and Alexander Statnikov. **Time and sample efficient discovery of Markov blankets and direct causal relations**. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 673–678. ACM, 2003. 18

[59] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. **The max-min hill-climbing Bayesian network structure learning algorithm**. *Machine Learning*, **65**(1):31–78, 2006. 18

[60] Constantin F Aliferis, Ioannis Tsamardinos, and Alexander Statnikov. **HITON: a novel Markov Blanket algorithm for optimal variable selection**. **2003**, page 21. American Medical Informatics Association, 2003. 18, 30, 33

[61] Dimitris Margaritis and Sebastian Thrun. **Bayesian network induction via local neighborhoods**. In *NIPS*, **99**, pages 505–511, 1999. 19

[62] Jean-Philippe Pellet and Andr Elisseeff. **A partial correlation-based algorithm for causal structure discovery with continuous variables**. In *Advances in Intelligent Data Analysis VII*, pages 229–239. Springer, 2007. 19

[63] Jean-Philippe Pellet and Andr Elisseeff. **Using Markov blankets for causal structure learning**. *The Journal of Machine Learning Research*, **9**:1295–1342, 2008. 19

[64] Markus Kalisch, Martin Mchler, Diego Colombo, Marloes H Maathuis, and Peter Bhlmann. **Causal inference using graphical models with the R package pcalg**. *Journal of Statistical Software*, **47**(11):1–26, 2012. 19, 54, 156

[65] Elisabeth Kvaavik, G. David Batty, Giske Ursin, Rachel Huxley, and Catharine R. Gale. **Influence of individual and combined health behaviors on total and cause-specific mortality in men and women: the United Kingdom health and lifestyle survey**. *Archives of Internal Medicine*, **170**(8):711–718, April 2010. 20

[66] Anne M Euser, Carmine Zoccali, Kitty J Jager, and Friedo W Dekker. **Cohort studies: prospective versus retrospective**. *Nephron Clinical Practice*, **113**(3):c214–c217, 2009. 20

[67] Yingyi Luo, Xiankai Li, Jue Li, Xiaoming Wang, Yongxia Qiao, Dayi Hu, Philip A. Merriam, and Yunsheng Ma. **Combined effects of smoking and peripheral arterial disease on all-cause and cardiovascular disease mortality in a Chinese male cohort**. *Journal of Vascular Surgery*, **51**(3):673–678, March 2010. 20

[68] Patricia W. Cheng and Laura R. Novick. **A probabilistic contrast model of causal induction.** *Journal of Personality and Social Psychology*, **58**(4):545, 1990. 20

[69] Bob Rehder. **Reasoning with conjunctive causes**. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 1406–1411, Austin, TX: Cognitive Science Society. 21

[70] Peter A. White. **Causal judgement from contingency information: judging interactions between two causal candidates**. *The Quarterly Journal of Experimental Psychology Section A*, **55**(3):819–838, August 2002. 22

[71] Lance W. Hahn, Marylyn D. Ritchie, and Jason H. Moore. **Multifactor dimensionality reduction software for detecting genegene and geneenvironment interactions**. *Bioinformatics*, **19**(3):376–382, February 2003. 23

[72] David J. Miller, Yanxin Zhang, Guoqiang Yu, Yongmei Liu, Li Chen, Carl D. Langefeld, David Herrington, and Yue Wang. **An algorithm for learning maximum entropy probability models of disease risk that efficiently searches and sparingly encodes multilocus genomic interactions**. *Bioinformatics*, **25**(19):2478–2485, 2009. 23

[73] Xia Jiang, Jeremy Jao, and Richard Neapolitan. **Learning predictive interactions using information gain and Bayesian network scoring**. *PLOS ONE*, **10**(12):e0143247, December 2015. 23

[74] Kenneth J. Rothman. **Causes**. *American Journal of Epidemiology*, **104**(6):587–592, 1976. 23, 63, 66

[75] Tyler J. VanderWeele and Mirjam J. Knol. **A tutorial on interaction**. *Epidemiologic Methods*, **3**(1):33–72, 2014. 23

[76] Tyler J. VanderWeele and James M. Robins. **Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect**. *American Journal of Epidemiology*, **166**(9):1096–1104, November 2007. 24

[77] James E. Smith, Samuel Holtzman, and James E. Matheson. **Structuring conditional relationships in influence diagrams**. *Operations Research*, **41**(2):280–297, 1993. 24

[78] Sren Hjsgaard. **Statistical inference in context specific interaction models for contingency tables**. *Scandinavian Journal of Statistics*, **31**(1):143–158, 2004. 24

[79] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 24

[80] Craig Boutilier, Nir Friedman, Moises Goldszmidt, and Daphne Koller. **Context-specific independence in Bayesian networks**. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, pages 115–123, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc. 24

[81] Nevin Lianwen Zhang and David Poole. **On the role of context-specific independence in probabilistic inference**. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, IJCAI '99, pages 1288–1293, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. 25

[82] David Poole and Nevin Lianwen Zhang. **Exploiting contextual independence in probabilistic inference**. *Journal of Artificial Intelligence Research (JAIR)*, **18**:263–313, 2003. 25

[83] Brian Milch, Bhaskara Marthi, David Sontag, Stuart J. Russell, Daniel L. Ong, and Andrey Kolobov. **Approximate inference for infinite contingent bayesian networks**. In *AIStats*, **2005**, page 10th, 2005. 25

[84] Jukka Corander. **Labelled graphical models**. *Scandinavian Journal of Statistics*, **30**(3):493–508, September 2003. 26

[85] Eric Mjolsness. **Labeled graph notations for graphical models**. Technical report, Technical Report 04-03, School of Information and Computer Science, UC Irvine, 2004. 26

[86] Tom Minka and John Winn. **Gates**. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1073–1080. Curran Associates, Inc., 2009. 26

[87] Henrik Nyman, Johan Pensar, Timo Koski, and Jukka Corander. **Stratified graphical models - context-specific independence in graphical models**. *Bayesian Analysis*, **9**(4):883–908, December 2014. 26

[88] Henrik Nyman, Johan Pensar, and Jukka Corander. **Stratified Gaussian graphical models**. *Communications in Statistics - Theory and Methods*, **0**(ja):0–0, June 2016. 26

[89] Patrick Royston and Willi Sauerbrei. **A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials**. *Statistics in Medicine*, **23**(16):2509–2525, August 2004. 27

[90] Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M. Nickerson, and Bogong Li. **Subgroup analysis via recursive partitioning**. *The Journal of Machine Learning Research*, **10**:141–158, 2009. 27

[91] Jared C. Foster, Jeremy MG Taylor, and Stephen J. Ruberg. **Subgroup identification from randomized clinical trial data**. *Statistics in Medicine*, **30**(24):2867–2880, 2011. 27

[92] Leo Breiman. **Random forests**. *Machine Learning*, **45**(1):5–32, 2001. 27

[93] Miroslav Dudk, John Langford, and Lihong Li. **Doubly robust policy evaluation and learning**. *arXiv preprint arXiv:1103.4601*, 2011. 27

[94] Lu Tian, Ash A. Alizadeh, Andrew J. Gentles, and Robert Tibshirani. **A simple method for estimating interactions between a treatment and a large number of covariates**. *Journal of the American Statistical Association*, **109**(508):1517–1532, October 2014. 28

[95] Robert Tibshirani. **Regression shrinkage and selection via the lasso**. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 28

[96] Donald P. Green and Holger L. Kern. **Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees**. *Public Opinion Quarterly*, **76**(3):491–511, January 2012. 28

[97] Wei Chen, Debashis Ghosh, Trivellore E. Raghunathan, Maxim Norkin, Daniel J. Sargent, and Gerold Bepler. **On Bayesian methods of exploring qualitative interactions for targeted treatment**. *Statistics in Medicine*, **31**(28):3693–3707, December 2012. 28

[98] Richard E. Neapolitan and others. *Learning Bayesian networks*, **38**. Pearson Prentice Hall Upper Saddle River, NJ, 2004. 30, 31

[99] Jerzy Neyman and Karolina Iwaszkiewicz. **Statistical problems in agricultural experimentation**. *Supplement to the Journal of the Royal Statistical Society*, **2**(2):107–180, 1935. 35

[100] Stephen L. Morgan and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2014. 35

[101] Paul W. Holland. **Statistics and causal inference**. *Journal of the American Statistical Association*, **81**(396):945–960, 1986. 37

[102] Donald B Rubin. **Inference and missing data**. *Biometrika*, **63**(3):581–592, 1976. 37

[103] F. Stuart Chapin. *Experimental designs in sociological research*. Harper, New York, 1947. 39

[104] Donald B Rubin and Neal Thomas. **Characterizing the effect of matching using linear propensity score methods with normal distributions**. *Biometrika*, **79**(4):797–809, 1992. 40

[105] Brian K. Lee, Justin Lessler, and Elizabeth A. Stuart. **Improving propensity score weighting using machine learning**. *Statistics in Medicine*, **29**(3):337–346, February 2010. 40, 144

[106] Daniel F. McCaffrey, Greg Ridgeway, and Andrew R. Morral. **Propensity score estimation with boosted regression for evaluating causal effects in observational studies.** *Psychological Methods*, **9**(4):403, 2004. 40

[107] Donald B Rubin and Neal Thomas. **Matching using estimated propensity scores: relating theory to practice**. *Biometrics*, pages 249–264, 1996. 40, 142, 143

[108] Guido W. Imbens. **The role of the propensity score in estimating dose-response functions**. *Biometrika*, **87**(3):706–710, 2000. 40, 77

[109] Jeremy A. Rassen, Abhi A. Shelat, Jessica M. Franklin, Robert J. Glynn, Daniel H. Solomon, and Sebastian Schneeweiss. **Matching by propensity score in cohort studies with three treatment groups**. *Epidemiology*, **24**(3):401–409, 2013. 40, 77

[110] Donald B Rubin. **Matching to remove bias in observational studies**. *Biometrics*, pages 159–183, 1973. 40

[111] Elizabeth A Stuart. **Matching methods for causal inference: A review and a look forward**. *Statistical Science: a review journal of the Institute of Mathematical Statistics*, **25**(1):1, 2010. 41, 104, 142, 143

[112] Gary King, Richard Nielsen, Carter Coberley, James E Pope, and Aaron Wells. **Comparative effectiveness of matching methods for causal inference**. *Unpublished manuscript*, **15**, 2011. 41

[113] Jos R. Zubizarreta. **Stable weights that balance covariates for estimation with incomplete outcome data**. *Journal of the American Statistical Association*, **110**(511):910–922, 2015. 41

[114] Jos R. Zubizarreta. **Using mixed integer programming for matching in an observational study of kidney failure after surgery**. *Journal of the American Statistical Association*, **107**(500):1360–1371, 2012. 41

[115] Mar?a de los Angeles Resa and Jos R. Zubizarreta. **Evaluation of subset matching methods and forms of covariate balance**. 2010. 41

[116] Junning Li and Z Jane Wang. **Controlling the false discovery rate of the association/causality structure learned with the PC algorithm**. *The Journal of Machine Learning Research*, **10**:475–514, 2009. 50

[117] Marco Scutari. **Learning Bayesian networks with the bnlearn R package**. *Journal of Statistical Software*, **35**(3), 2010. 50

[118] Kevin Bache and Moshe Lichman. **UCI machine learning repository, 2013**. *URL http://archive.ics.uci.edu/ml*. 53, 86, 119, 146

[119] Ron Kohavi, Dan Sommerfield, and James Dougherty. **Data mining using a machine learning library in C++**. *International Journal on Artificial Intelligence Tools*, **6**(04):537–566, 1997. 54, 158

[120] Sunita KC Basnet, Sarah Diab, Raffaella Schmid, Mingfeng Yu, Yuchao Yang, Todd Alexander Gillam, Theodosia Teo, Peng Li, Tom Peat, Hugo Albrecht, and others. **Identification of a highly conserved allosteric binding site on Mnk1 and Mnk2**. *Molecular Pharmacology*, **88**(5):935–948, 2015. 62

[121] Asli Memisoglu, Frank B. Hu, Susan E. Hankinson, JoAnn E. Manson, Immaculata De Vivo, Walter C. Willett, and David J. Hunter. **Interaction between a peroxisome proliferator-activated receptor gamma gene polymorphism and dietary fat intake in relation to body mass**. *Human Molecular Genetics*, **12**(22):2923–2929, November 2003. 62

[122] Kenneth J. Rothman, Sander Greenland, and Alexander M. Walker. **Concepts of interaction**. *American Journal of Epidemiology*, **112**(4):467–470, 1980. 63

[123] Tyler J. VanderWeele. **On the distinction between interaction and effect modification**. *Epidemiology (Cambridge, Mass.)*, **20**(6):863–871, November 2009. 67

[124] Mirjam J. Knol, Tyler J. VanderWeele, Rolf H. H. Groenwold, Olaf H. Klungel, Maroeska M. Rovers, and Diederick E. Grobbee. **Estimating measures of interaction on an additive scale for preventive exposures**. *European Journal of Epidemiology*, **26**(6):433–438, June 2011. 68

[125] Brandon L. Pierce and Habibul Ahsan. **Case-only genome-wide interaction study of disease risk, prognosis and treatment**. *Genetic Epidemiology*, **34**(1):7–15, January 2010. 68

[126] Jiuyong Li. **On optimal rule discovery**. *IEEE Transactions on Knowledge and Data Engineering*, **18**(4):460–471, April 2006. 76

[127] Chunhao Tu, Shuo Jiao, and Woon Yuen Koh. **Comparison of clustering algorithms on generalized propensity score in observational studies: a simulation study**. *Journal of Statistical Computation and Simulation*, **83**(12):2206–2218, 2013. 77

[128] Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013. 79, 136

[129] James M. Robins. **Marginal structural models versus structural nested models as tools for causal inference**. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 95–133. Springer, 2000. 79

[130] Nicholas D. Soulakis, Matthew B. Carson, Young Ji Lee, Daniel H. Schneider, Connor T. Skeehan, and Denise M. Scholtens. **Visualizing collaborative electronic health record usage for hospitalized patients with heart failure**. *Journal of the American Medical Informatics Association*, **22**(2):299–311, 2015. 89

[131] Zexian Zeng, Xia Jiang, and Richard Neapolitan. **Discovering causal interactions using Bayesian network scoring and information gain**. *BMC Bioinformatics*, **17**(1):1, 2016. 89

[132] Trevor Hastie, Robert Tibshirani, Balasubramanian Narasimhan, and Gilbert Chu. **Impute: imputation for microarray data**. *R package version*, **1**(0), 2012. 89

[133] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. **limma powers differential expression analyses for RNA-sequencing and microarray studies**. *Nucleic Acids Research*, page gkv007, 2015. 90

[134] Benjamin P. Lewis, I.-hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, and Christopher B. Burge. **Prediction of mammalian microRNA targets**. *Cell*, **115**(7):787–798, 2003. 90

[135] Minoru Kanehisa and Susumu Goto. **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Research*, **28**(1):27–30, 2000. 91

[136] Thuc Duy Le, Junpeng Zhang, Lin Liu, and Jiuyong Li. **Ensemble methods for miRNA target prediction from expression data**. *PLoS ONE*, **10**(6):e0131627, June 2015. 91

[137] J. Ross Quinlan. *C4.5: programs for machine learning*. Elsevier, 2014. 101, 109

[138] Jiuyong Li, Lin Liu, Jixue Liu, and Ryan Green. **Building diversified multiple trees for classification in high dimensional noise data**. *arXiv:1612.05888 [cs, stat]*, December 2016. arXiv: 1612.05888. 109

[139] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. **The WEKA data mining software: an update**. *ACM SIGKDD explorations newsletter*, **11**(1):10–18, 2009. 115, 145, 159

[140] MW Birch. **The detection of partial association, I: The 2 2 case**. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 313–324, 1964. 136

[141] Nathan Mantel and William Haenszel. **Statistical aspects of the analysis of data from retrospective studies of disease**. *Journal of the National Cancer Institute*, 1959. 136

[142] Peter C. Austin. **An introduction to propensity score methods for reducing the effects of confounding in observational studies**. *Multivariate Behavioral Research*, **46**(3):399–424, 2011. 143

[143] Paul R Rosenbaum. **From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment**. *Journal of the American Statistical Association*, **79**(385):41–48, 1984. 143

[144] Paul Komarek. **Logistic regression for data mining and high-dimensional classification**. *Robotics Institute*, page 222, 2004. 144

[145] Austin Bradford Hill. **The environment and disease: association or causation?** *Proceedings of the Royal Society of Medicine*, **58**(5):295, 1965. 149

[146] Jiuyong Li, Lin Liu, and Thuc Le. *Practical approaches to causal relationship exploration*. Springer, 2015. 149

[147] Jiuyong Li, Ada Wai-chee Fu, and Paul Fahey. **Efficient discovery of risk patterns in medical data**. *Artificial Intelligence in Medicine*, **45**(1):77–89, 2009. 150

[148] Marloes H. Maathuis, Markus Kalisch, and Peter Bhlmann. **Estimating high-dimensional intervention effects from observational data**. *The Annals of Statistics*, **37**(6A):3133–3164, December 2009. 166

[149] Marloes H. Maathuis, Diego Colombo, Markus Kalisch, and Peter Bhlmann. **Predicting causal effects in large-scale systems from observational data**. *Nature Methods*, **7**(4):247–248, April 2010. 166