

University of Nottingham
UK | CHINA | MALAYSIA

Explain the world — towards leveraging causality in fuzzy rule based systems

Thesis submitted to the University of Nottingham
for the degree of **Doctor of Philosophy**

Te Zhang
20243875

Supervised by
Christian Wagner
Jonathan M. Garibaldi

School of Computer Science
University of Nottingham

Submitted November 2024

Abstract

Artificial intelligence (AI) is increasingly applied across sectors, including risk-sensitive areas such as healthcare and security, driving the growing demand for explainable AI (XAI). Among various XAI approaches, fuzzy rule-based systems provide a tempting architecture for XAI, offering linguistic, human-accessible rules, combined with the capacity for handling complex applications amid varying levels of uncertainty, vagueness and imprecision. Nowadays, rules are frequently obtained through data-driven approaches. However, while these systems explain model behaviour by revealing the relationships between the variables captured, they often capture correlations. Ideally, AI systems are expected to go beyond explainable AI (XAI), that is, systems which not only explain their behaviour, but also communicate their 'insights' in respect to the real world. Thus, rules are expected to capture causal relationships between variables.

This thesis argues that fuzzy rule-based systems, where the rules reflect causal relationships between variables, offer unique benefits in terms of performance and explainability, particularly by enhancing the communication of AI insights to people. In other words, ideally, the rules of such systems can explain not only what happens within the model but also what happens in the real world. Based on this, this thesis focuses on how to automatically generate rules which reflect causal relationships between variables from data sets using data-driven approaches. To achieve this goal, the following two problems are investigated: 1) What are the nature and role of causal relationships in the context of artificial intelligence reasoning. 2) how to automatically generate rules leveraging the causal information obtained from a given data set.

The first problem relates to the concept of causality, which is complex and an ongoing topic for discussion across disciplines, with no universally accepted definition. To solve the first problem, this thesis first summarizes different definitions of causality and introduces the definition adopted in this thesis. Then, this thesis introduces the tool used to represent causal relationships—the causal graph—and provides a detailed analysis of the causal information that can be obtained from a causal graph of a given data set. Following that, this thesis discusses different facets of causal relationships can be derived from a causal graph. Finally, this thesis reviews the established data-driven approaches for generating causal graphs from a given data set.

To solve the second problem, a data-driven causal rule generation framework is established in this thesis. The framework is designed to start by generating a causal graph from a given data set using a data-driven approach. Then, the framework uses causal information between variables obtained from the causal graph to remove variables which are not causally related to the target variable. Finally, the framework uses a data-driven rule generation approach to generate rules from the refined data set, thereby achieving causal rule generation.

To enable users to customize, based on their needs, the generation of causal

explanations provided by a fuzzy system, this thesis proposes three variants of the framework. These variants leverage different types of causal information to generate rules which reflect different facets of causal explanations. This thesis provides a detailed analysis of the differences in causal explanations provided by the rules generated by these variants and their applicable scenarios.

Furthermore, one meta-variant of the framework is proposed to complement an explanation provided by the rules obtained by the established framework. The meta-variant is designed to generate counterfactual explanations based on rules obtained by a variant of the framework. Uniquely, a counterfactual explanation obtained by the meta-variant of the framework articulates how the given inputs would need to be changed to generate a different output, crucial for lay-user insight, verification and sensitivity-evaluation of XAI systems, for example in decision support around credit risk, cyber security and medical assistance.

Beyond the theoretical framework, a software tool is developed to promote the dissemination and application of the established framework. The software tool is a Python library which contains essential functions to implement different variants of the established framework for solving classification problems. This thesis summarises and describes the features of the developed tool. In addition, this thesis demonstrates how to use the developed Python library to implement different variants of the established framework.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisors, Prof. Christian Wagner and Prof. Jonathan M. Garibaldi for their invaluable advice, continuous support, and patience during my PhD journey. I am fortunate to work with them. I am thankful to them for not only being academic supervisors but also mentoring personally throughout my PhD journey in Nottingham.

I would like to extend my sincere thanks to friends/the members of LUCID groups. Their insightful comments, suggestions and encouragements significantly enriched this thesis.

My special thanks to my beloved wife, Mengmeng. I am thankful for her continuous support and encouragement, for being there with me through every joyful and difficult moment of my PhD journey, and for being a constant source of support every step of the way.

Finally, I am deeply grateful to my parents Fei Lan and Xingdong Zhang, as well as my family back in China. I would not start the journey pursuing my PhD in the UK without their unwavering support. This thesis would not have happened without their endless love.

Contents

Abstract i

Acknowledgements iii

1 Introduction 1

1.1 Notations and acronyms used in this thesis 2

1.2 Why AI needs causality? 2

1.3 Why AI needs Rules 5

1.4 Why causal rules: Motivation and objectives 8

1.5 Thesis structure and key contributions 11

1.6 List of publications arising from this thesis 13

2 Background 15

2.1 Causality in AI 15

2.1.1 Definitions of causality 15

2.1.2 Causal models 19

2.2 Classical data-driven causal discovery approaches 22

2.2.1 Causal assumptions for data-driven causal discovery algorithms 23

2.2.2 Causal structure learning approaches 25

2.2.3 Causal direction inference approaches 27

2.2.4 Confounding variable detection 29

2.3 Explanations in AI 30

2.4 Explainability of fuzzy systems 32

2.4.1 Semantic meaning of rules 32

2.4.2 Complexity of fuzzy systems 33

2.5 Overview of rule generation for fuzzy systems 35
2.5.1 Rule reduction and compression approaches 36
2.5.2 Sparse rule generation approaches 37
2.5.3 Iterative rule compression approaches 37
2.6 Summary 39

3 Towards causal fuzzy logic rules by leveraging Markov blankets 41

3.1 Overview of the MABLAR framework 42
3.2 Motivation of MABLAR-ST 45
3.3 MABLAR-ST – The standard Markov blanket rule generation framework 47
3.3.1 The process of MABLAR-ST 47
3.3.2 Rationale of MABLAR-ST 48
3.3.3 Targeted benefits of MABLAR-ST 49
3.4 MABLAR-ST illustration – Diabetes diagnosis 50
3.4.1 Problem and the data set 51
3.4.2 Rule generation using MABLAR-ST 51
3.4.3 Performance of MABLAR-ST on the Pima Indian diabetes data set 55
3.4.4 Model-based explanations of the Pima diabetes classification fuzzy system 55
3.4.5 Domain-based explanations of the Pima diabetes classification fuzzy system 57
3.5 Experiments 59
3.5.1 Experiment setting 59
3.5.2 Performance and explanations 60
3.6 Discussion 62
3.7 Summary 63

4 Towards causal fuzzy system rules using causal direction 66

4.1 Motivation of MABLAR-CD 67

4.2 MABLAR-CD – Markov blanket rule generation using causal direction . . . 69
4.2.1 The process of MABLAR-CD . . . 69
4.2.2 Rationale of MABLAR-CD . . . 70
4.2.3 Comparison between causal relationships captured by MABLAR-ST and MABLAR-CD . . . 72
4.3 Experiments . . . 73
4.3.1 Experiment setting . . . 73
4.3.2 Evaluation on synthetic data sets . . . 75
4.3.3 Evaluation on real-world data sets . . . 81
4.3.4 Evaluation of domain-based explanations . . . 88
4.4 Discussion of the experiment results . . . 91
4.5 Summary . . . 93

5 Towards locally relevant explanations using causal rule discovery 95

5.1 Motivation of MABLAR-CW . . . 96
5.2 MABLAR-CW – Markov blanket rule generation using causal weight . . . 98
5.2.1 The process of MABLAR-CW . . . 98
5.2.2 The rationale of MABLAR-CW . . . 101
5.2.3 The prediction and local explanation generation process of MABLAR-CW . . . 103
5.3 Experiment Results . . . 106
5.3.1 Experiment settings . . . 106
5.3.2 Performance evaluation . . . 108
5.3.3 Local explanations — a case study . . . 110
5.4 Summary . . . 114

6 Towards counterfactual rule generation for fuzzy systems 116

6.1 Motivation of CF-MABLAR . . . 117
6.2 CF-MABLAR – Counterfactual Markov blanket rule generation 119

6.2.1 The question to be answered by CF-MABLAR 119
6.2.2 The process of CF-MABLAR 120
6.2.3 Properties of CF explanations obtained by CF-MABLAR 124
6.3 CF-MABLAR illustration – Iris classification 125
6.3.1 Problem description and settings 125
6.3.2 Illustrative example of the CF-MABLAR framework . . 127
6.4 Experiments 129
6.4.1 Experiment settings 129
6.4.2 Evaluation metric 131
6.4.3 Results 131
6.5 Summary 133

7 *mablars*: a python library implementing the MABLAR framework 135

7.1 Motivation of *mablars* 136
7.2 The python library of the Markov blanket rule generation framework 137
7.2.1 Overview of *mablars* 137
7.2.2 Features of *mablars* 139
7.3 Demonstration of *mablars* 139
7.3.1 Demonstration using the PID data set 140
7.3.2 Computational complexity in practice 144
7.3.3 Impact of the causal discovery settings in practise 145
7.4 Summary 147

8 Conclusions and future work 148

8.1 Main contributions 149
8.1.1 MABLAR — a data-driven causal rule generation framework 149
8.1.2 MABLAR-ST — Implementing MABLAR using the Markov blanket information 150

8.1.3 MABLAR-CD — Implementing MABLAR using the causal direction information . . . 151
8.1.4 MABLAR-CW — Implementing MABLAR using the causal weight and causal path information . . . 152
8.1.5 CF-MABLAR — Implementing MABLAR for CF rule generation . . . 153
8.1.6 *mablars* — Implementing MABLAR in the wider Community . . . 154
8.2 Limitations and future works . . . 154
8.3 List of publications arising from this thesis . . . 157
Bibliography **159**
Appendices **172**
A A supplementary explanation of the experimental settings in Section 4.3 **172**
Appendices **173**
B The execution time in Section 4.3.3 **173**
B.1 The execution time of causal discovery in Section 4.3.3 . . . 173
B.2 The execution time (seconds) of fuzzy system rule generation approaches in Section 4.3.3 . . . 173
Appendices **175**
C The complete rule bases obtained by WM, MB-WM and MBCD-WM in Section 4.3.4 **175**
C.1 The complete rule base obtained by WM . . . 175
C.2 The complete rules obtained by MB-WM . . . 181
C.3 The complete rules obtained by MBCD-WM . . . 184
Appendices **185**
D The complete rule bases obtained by FURIA, MB-FURIA and MBCD-FURIA in Section 4.3.4 **185**
D.1 The complete rule base obtained by FURIA . . . 185
D.2 The complete rule base obtained by MB-FURIA . . . 186

D.3 The complete rule base obtained by MBCD-FURIA 186

List of Tables

1.1 Acronyms 2

1.2 Notation 2

2.1 The procedures of the PC algorithm 26

3.1 Rules in MB-WM on the PID data set 55

3.2 Number of rules in each fold 56

3.3 Rules in WM on the Pima Indian diabetes data sets 57

3.4 Data sets used in our experiments 59

3.5 Number of inputs of MB-WM and WM 60

4.1 $\boldsymbol{D}_{MB}$ and $\boldsymbol{D}_{MBCD}$ of the system output variable on the synthetic data set 76

4.2 Performance and # rules of different fuzzy systems 78

4.3 Data sets used in this chapter 83

4.4 The F1 scores of different fuzzy systems on each data set 85

4.5 The number of rules of different fuzzy systems for each data set 88

4.6 Variables used in different fuzzy systems 89

5.1 Data sets used in this chapter 107

5.2 Performance of different frameworks 108

5.3 The tested samples 111

5.4 Local explanations provided by the WM algorithm 111

5.5 Local explanations provided by MABLAR-ST 111

5.6 Local explanations provided by MABLAR-CD 112

5.7 Local explanations provided by MABLAR-CW 112

6.1 Part of rules of the obtained fuzzy system 127
6.2 Counterfactual rules of the selected sample 128
6.3 Data sets used in this chapter 130
6.4 Results on different data sets 132

7.1 The average running time (seconds) under different settings in the PID data set 145

B.1 The execution time (seconds) of causal discovery on each data set 174
B.2 The execution time (seconds) of fuzzy system rule generation approaches on each data set. 174

C.1 The rule base of WM for the PID data set 175
C.2 The rule base of MB-WM for the PID data set 181
C.3 The rule base of the PID data set of MB-CDWM 184

D.1 The rule base of FURIA for the PID data set 185
D.2 The rule base of MB-FURIA for the PID data set 186
D.3 The rule base of MBCD-FURIA for the PID data set 186

List of Figures

1.1 A basic structure of a Mamdani fuzzy system 6

1.2 The outline of the thesis 9

2.1 An example of a causal graph 19

2.2 Three basic structures of a causal graph 24

3.1 Comparison between classical approaches and the MABLAR framework 42

3.2 The process of MABLAR-ST 47

3.3 The rationale of MABLAR-ST 48

3.4 Step 1: Identification of a causal graph of the PID data set . . . 52

3.5 Step 2: Identification of the Markov Blanket of the Diabetes variable 52

3.6 Step 3: Construction of $\boldsymbol{D}_{MB}$ of the PID data set 52

3.7 Step 4: Generation of causal rules for the PID data set 53

3.8 The fuzzy partition on the Blood Pressure variable 54

3.9 Classification accuracy of MB-WM and WM 61

3.10 Average number of rules of MB-WM and WM 62

4.1 The process of MABLAR-CD 69

4.2 The rationale of MABLAR-CD 70

4.3 The ground truth causal graph of the simulated data set 77

4.4 The causal graph of the simulated data set obtained by adopted causal discovery algorithms 77

4.5 Results of Experiment 2 80

4.6 Experiment results with Experiment Setting 4 82
4.7 The causal graph of the PID dataset 89

5.1 The process of MABLAR-CW 98
5.2 An illustrative example of MABLAR-CW 99
5.3 The causal path from V_1 to V_4 102
5.4 The prediction process of MABLAR-CW 104
5.5 The weighted causal weighted graph of the Iris data set 109
5.6 The weighted causal weighted graph of the MAM data set . . . 110

6.1 The process of CF-MABLAR 121
6.2 Fuzzy partitions of each feature 126

7.1 The basic flowchart of using *mablars* 138
7.2 An example of a causal graph 143
7.3 Part of the obtained rule base 144
7.4 The causal graph of the PID data set obtained by ICA-LiNGAM 146
7.5 The rule base obtained by MABLAR-ST using the WM algorithm based on the causal graph obtained by ICA-LiNGAM . . 146

Chapter 1

Introduction

The rules of a rule-based system provide explanations for its behaviour by revealing the relationships between the variables captured. However, ideally, AI systems are expected to go beyond explainable AI (XAI), that is, systems which not only explain their behaviour, but also communicate their 'insights' in respect to the real world. This requires rules to capture causal relationships between variables, because humans are cause-effect thinkers [1]. When considering real-world processes, we generally seek to determine the causal relationships underlying those processes so as to understand them, be they disease or the motions of stars and planets.

This thesis explores different data-driven approaches to generate rules for fuzzy systems which reflect causal relationships between variables. This thesis argues that those systems where the rules reflect causal relationships between variables represent an important class of fuzzy rule-based systems with unique benefits. Specifically, such systems benefit from improved performance and robustness; facilitate explainability and even contribute towards one of the core ambitions for AI: the ability to communicate important relationships amongst a system's real-world variables to the human users of AI.

The rest of this chapter are organised as follows: Section 1.1 shows the acronyms the notation frequently used in this thesis. Section 1.2 discusses why AI models need to focus on causal relationships between variables. Section 1.3 discusses

why rules are needed in the AI area, including how they are presented and the meanings they can represent. Section 1.4 provides the motivations and the objectives for the research in this thesis. Section 1.5 shows the contributions and the thesis structure.

1.1 Notations and acronyms used in this thesis

This section summaries the acronyms and notation used in this thesis, which are shown in Table 1.1 and Table 1.2, respectively.

Table 1.1: Acronyms

AI	Artificial intelligence
XAI	Explainable artificial intelligence
ML	Machine learning
MB	Markov blanket
MABLAR	the Markov blanket rule generation framework
CF	Counterfactual

Table 1.2: Notation

$\boldsymbol{D}$	The original or given data set
$\boldsymbol{D}_{MB}$	The subset of $\boldsymbol{D}$ which only contains the target variable and its MB
$\boldsymbol{D}_{CD}$	The subset of $\boldsymbol{D}_{MB}$ which only contains the target variable and the direct causes within its MB
$\|D\|$	The number of variables in the data set $\boldsymbol{D}$
X_i	The ith variable within $\boldsymbol{D}$.
$\boldsymbol{Z}$	A set of variables in a causal graph
P	A path in a causal graph
$\cdot \perp \cdot \| \cdot$	Conditionally independent. For example, $X_1 \perp X_3 \| X_2$ means that X_1 and X_3 are conditionally independent given X_2.
$\boldsymbol{R}$	A rule base of a fuzzy system
R^k	The k-th rule of a fuzzy rule-based system
A_j^k	The fuzzy set of the j-th variable in the k-th rule

1.2 Why AI needs causality?

Causal relationships describe the relationship between two entities in terms of causation and its consequences, which reflects the internal mechanism and

laws of a system [2]. Researchers in many research areas, e.g., bioscience and social science, need to explore causality between entities to facilitate a deeper understanding and cognition of the laws governing natural and societal phenomena, thereby promoting the advancement of areas.

The definition of causality is a complex and ongoing topic for discussion across disciplines [2–4], with no universally accepted definition. Section 2.1.1 provides a detailed discussion of different definitions of causality and Section 3.1 establishes the specific working definition used in this thesis. Here, for semantic purpose, this thesis adopts an informal definition of causality as *a relation where one event or state (a cause) contributes, and is at least partially responsible for the realisation of another event or state (an effect)*—adapted from [5, 6].

In recent years, artificial intelligence (AI) technology—particularly machine learning-based approaches—has made significant advancements and garnered considerable attention. AI-related research even won the Nobel Prize in Physics and Chemistry in 2024. Broadly speaking, AI refers to the simulation of human intelligence processes by machines, and encompasses a wide range of approaches, including symbolic reasoning, expert systems, and rule-based logic. This thesis focuses on data-driven approaches for causal rule generation; therefore, the term 'AI' is used primarily to refer to data-driven machine learning methods.

However, to achieve good performance, most existing AI models mainly focus on correlations rather than causal relationships between variables [7, 8]. Consequently, AI models have the risk of making unexpected (or wrong) decisions in unexpected ways on problems which are easy for humans [7]. For example, Moosavi et al. [9] show a case where slight adjustments to the pixel values of an image can fool a deep learning model for image classification into making an incorrect prediction.

At the same time, AI models are increasingly used across sectors, from industry to government, including in risk sensitive areas such as healthcare and security.

As a consequence, the demand for XAI has been increased and a variety of XAI methods have been proposed, such as local interpretable model-agnostic explanations (LIME) [10] and mimic learning [11, 12]. However, existing XAI methods mainly focus on correlations explored by models, resulting they can only explain what happens in the model, but may not able to explain what happens in the real-world [7].

For example, Ribeiro et al. [10] report a case where a deep learning model is trained to discriminate between wolves and huskies using photos. Through the explanations for this deep learning model, users found that the model was making its predictions based on the presence of snow in the photos, because all wolf photos contained snow, while none of the husky photos did. However, the explanations of this model only helped users understand how the model operated; they did not provide guidance on how to distinguish wolves from huskies in the real world (you cannot simply label a husky as a wolf just because it is standing in the snow!). This is because the explanations reveal the correlation between wolves/huskies and snow in the photos but failed to reveal the true causes that differentiate wolves and huskies.

Based on the above discussion, compared to AI models which focus on the correlations between variables, AI models which focus on causal relationships between variables have the potential to achieve better performance and robustness. In addition, if such models are self-explainable, users may understand not only what happens in the model, but also what happens in the real-world through the explanations provided by such models. However, it is still an open problem to build an explainable causal AI model.

Rule-based models are often considered self-explainable and have been widely used in fields with a demand for XAI in recent years [13–15], which are discussed in the next section.

1.3 Why AI needs Rules

Rule-based systems are considered to be self-explainable because their rules show the behaviour of the system in a human-readable way [16, 17]. The most common structure of rule-based systems follows *modus ponens* 'IF-THEN' rules, which can be represented as follows [18]:

$$\begin{aligned} &\text{IF } \textit{condition } 1 \text{ AND } \textit{condition } 2 \text{ AND } \cdots \text{ AND } \textit{condition } n \\ &\text{THEN } \textit{conclusion} \end{aligned} \tag{1.1}$$

As shown in Equation (1.1), an 'IF-THEN' rule consists of an 'IF' part and a 'THEN' part. The 'IF' part is also called the antecedent while the 'THEN' part is also called the consequent.

The antecedent consists of one condition or multiple conditions connected by logical connectors. A condition consists of two parts: an object and its value. The object and its value are connected by an operator. An operator can be a symbolic value, such as 'is', 'are', 'is not' and 'are not', or a mathematical operator, such as '$>$', '$<$', '$\leq$', '$\geq$' and '$\neq$'. For example, '*The body temperature is 37 degrees Celsius*' is a condition where the object is '*the body temperature*' and the value is '*36.5 degrees Celsius*', and they are connected by a symbolic value '*is*'. Meanwhile, '*The body temperature* $>$ *37 degrees Celsius*' is a condition which has the same object and value as the previous one, but the connector is a mathematical operator '$>$'.

Logical connectors of an antecedent in a rule-based system include 'AND', 'NOT' and 'OR'. The conditions of an antecedent can be connected by one or more types of logical connectors. For example, 'IF *body temperature* $>$ *37 degrees Celsius* AND *cough is True* OR *Headache is True*' is an antecedent which has three conditions connected by the logical connectives 'AND' and 'OR'.

Similar to the conditions in the antecedent, the conclusion in the consequent can also be composed of an object and its value connected by a symbolic value

or a mathematical operator. For example, 'THEN *flu is true*' is a consequent where the object 'flu' and its value 'true' are connected by a symbolic value 'is'. A consequent can also contain multiple conclusions. For example, 'THEN *severity of condition is high* AND *treatment measure is urgency*' is a consequent which contains two conclusions.

In recent years, fuzzy systems have become a popular type of rule-based system used in AI, because they have strong ability to model complex non-linear relationships using human-readable rules [13, 19]. Fig. 1.1 shows a basic structure of a Mamdani fuzzy system as an example of fuzzy systems [14].

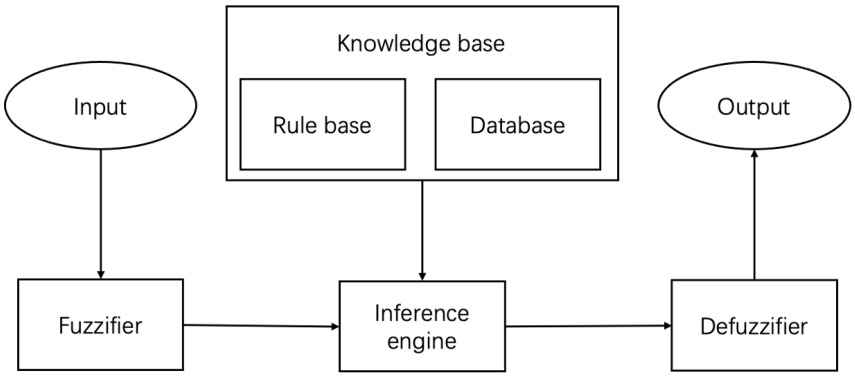


Figure 1.1: *A basic structure of a Mamdani fuzzy system*

In general, as shown in Fig. 1.1, a Mamdani fuzzy system contains the following four parts [14]:

- Fuzzifier: A fuzzifier maps a crisp (real-valued) input into a fuzzy set defined in the universe of discourse (the domain of the fuzzy set) represented by membership functions.

- Knowledge base: A knowledge base of a Mamdani fuzzy system contains the domain knowledge useful for problem solving, and usually consists of a rule base and a data base. A rule base contains a set of 'IF-THEN' rules which represent the knowledge obtained by the system. A database in a Mamdani fuzzy system contains linguistic variables and corresponding fuzzy partitions.

- Fuzzy inference engine: An inference engine performs reasoning by using

the 'IF-THEN' in the knowledge base to produce a fuzzy output based on the fuzzy input obtained by the fuzzifier.

- Defuzzifier: A defuzzifier converts the fuzzy output obtained by the fuzzy inference engine to a crisp value.

The conditions and conclusions in the 'IF-THEN' rules of a Mamdani fuzzy system are composed of the names of variables and their corresponding fuzzy sets. The k-th 'IF-THEN' rule of a Mamdani fuzzy system can be represented as follows: [14]:

$$\begin{aligned} &\text{IF } X_1 \textit{ is } A_1^k \text{ AND } X_2 \textit{ is } A_2^k \text{ AND} \cdots \text{ AND } X_d \textit{ is } A_d^k \\ &\text{THEN } Y \textit{ is } A_y^k \text{ AND} \end{aligned} \tag{1.2}$$

In Equation (1.2), X_i and Y are the input and output variables, respectively. A_i^k and A_y^k are the fuzzy sets of the input and output variables in the k-th rule, respectively.

Rules of fuzzy systems are often generated in two ways: 1) the knowledge-based way and 2) the data-driven way [16]. In the knowledge-based way, rules are usually generated by experts directly and/or generated using historical information of similar systems. However, in some real-word applications, rules of a fuzzy system may not be able to be generated by experts. For example, there may be no experts available in the field where the system is to be used, or the cost of hiring experts may be too high [16].

Nowadays, with the development of data collection techniques and the increase in computing power, generating rules using the data-driven way become more and more popular. For a given data set, fuzzy systems of which rules are generated by data-driven approaches often have better performance (e.g. the classification accuracy) than those obtained by knowledge-based approaches, especially when the background knowledge is incomplete or experts have subjective preferences [16]. However, data-driven approaches mainly focus on improving the performance of rule-based models on the given data set. As a

result, the obtained rule bases usually have a high complexity (e.g., containing a large number of rules), which reduces the explainability of the obtained rule-based systems [16, 20].

In addition, to achieve good performance, most of existing data-driven rule generation approaches exploit statistical correlations between the variables within a data set—rather than necessarily capturing causal relationships. However, as human are cause–effect thinkers [1], when considering real-world processes, we generally seek to determine the causal relationships underlying those processes so as to understand them, be they disease or the motions of stars and planets. Thus, rules which reflect correlations between the variables can mislead users, and cannot in general provide meaningful explanations for real-world processes.

1.4 Why causal rules: Motivation and objectives

As discussed above, fuzzy systems obtained by existing data-driven approaches have the risk of containing rules which reflect (spurious) correlations but not any actual causal relationships between the variables. These rules can increase the complexity of the resulting fuzzy systems and negatively impact performance, particularly in terms of generalisation-capability and explainability of rule-based systems.

The aim of this thesis is to address the aforementioned issue. More specifically, this thesis aims to develop data-driven approaches that can automatically generate causal rules for fuzzy systems from a given data set. The obtained rules should capture and represent causal relationships between variables within the data set. Ideally, the causal relationships captured by the obtained rules will not only hold within the given data set but also in the real world – however it should be noted that verifying this is beyond the scope of this thesis–except

within narrow experimental cases where ground truth information is available. Fig. 1.2 shows the outline of the research of this thesis, including the research aim, research objectives (ROs) and contributions(highlight in grey).

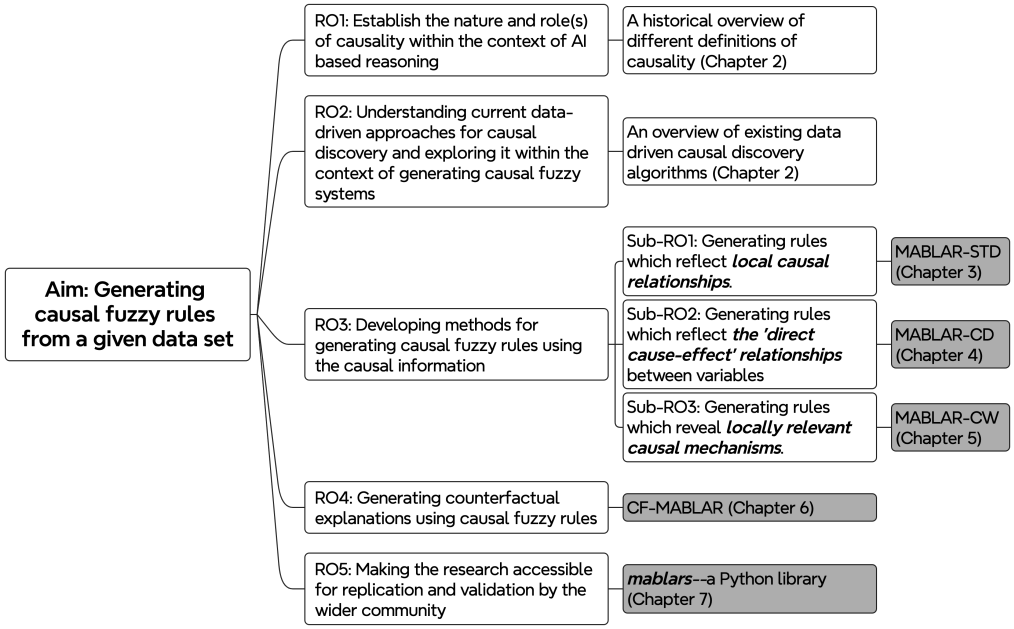


Figure 1.2: *The outline of the thesis*

To achieve the aim of this research, as shown in Fig. 1.2, five research objectives (ROs) are formulated:

1. RO1: Establish the nature and role(s) of causality within the context of AI based reasoning. It is well known that 'correlation does not imply causal relationships' [21, 22]. To distinguish between correlation and causal relationships, researchers from different areas, including the AI area, have proposed multiple definitions of causality (further discussed in Section 2.1). However, there is still no universally accepted definition. The purpose of RO1 is not to provide a perfect definition of causality. Instead, it is to understand different types of causal relationships across different areas. Then, focusing on the context of AI, RO1 determines the model to be used in this thesis for representing causal relationships, i.e., the causal graph model (will be introduced in Section 2.1.2), and establishes the facets of causal relationships which can be derived from

the causal graph model.

2. RO2: Understanding current data-driven approaches for causal discovery and explore it within the context of generating causal fuzzy systems. A data set contains critical information which can be used to generate rules that reflect causal relationships between variables. To exploit causal information within a given data set, many data-driven causal discovery approaches have been proposed. RO2 aims to review and understand current data-driven causal discovery approaches, and exploits how to leverage them for causal rule generation.

3. RO3: Developing methods for generating rules which reflect different facets of causal relationships. As mentioned in RO1, different facets of causal relationships can be derived from the causal model. RO3 aims to establish the roles that rules which reflect different facets of causal relationships can play in the context of fuzzy systems, and develop approaches to generate causal fuzzy rules which reflect different facets of causal relationships. Thus, as shown in Fig. 1.2, RO3 is further divided into three sub-research objectives (sub-ROs) which achieve generating rules that reflect different facets of causal relationships, respectively. The details of each Sub-RO will be further explained in the corresponding chapters.

4. RO4: Generating counterfactual explanations using causal fuzzy rules. Rules reflect causal relationships have the potential to provide counterfactual explanations, which can complement a meaningful explanation [23]. RO4 aims to develop an approach which generates counterfactual explanations based on the causal rules obtained by the approaches proposed in this thesis, and thus providing a complement to the explanations provided by the obtained rules. Ideally, the counterfactual explanations can tell users how to intervene, i.e., how to change the values of the given input variables of a sample to obtain a desired output.

5. RO5: Making the research accessible for replication and validation by the wider community. To disseminate and share the research, RO5 aims to develop a software tool. The tool is designed to help users from different areas in implementing causal fuzzy systems without requiring them to be familiar with causal discovery or fuzzy systems.

1.5 Thesis structure and key contributions

The contribution of this thesis is to articulate the benefits of using rules which reflect causal relationships, from a performance and an XAI point of view, highlighting in particular their potential in communicating AI insight to people. This thesis establishes a data-driven framework for causal rule generation, and proposes three variants and one meta-variant of the proposed framework. Additionally, a Python library is developed to help users to implement the proposed framework.

The rest of this thesis is organised as follows:

- Chapter 2 provides background of this thesis. It begins with a historical overview of different definitions of causality, summarizing popular causal models and offering a detailed explanation of the specific causal model adopted in this thesis. Additionally, it introduces the different facets of causal relationships that can be derived from the adopted model and explores the roles these derived different facets of causal relationships can play in AI reasoning, which covers RO1. The chapter also summarises different types of explanations in AI, discusses factors influencing the explainability of fuzzy systems, and reviews data-driven rule generation algorithms designed to improve their explainability.

- Chapter 3 establishes a data-driven rule generation framework which covers the content of RO3. The proposed framework is called the Markov blanket rule generation (MABLAR) framework. The MABLAR framework is designed to capture the different roles of different facets of causal

relationships within the context of fuzzy rule based systems. It establishes how to source and leverage different types of causal information from a given data set to generate rules which reflect different facets of causal relationships derived from a causal graph of a given data set. Chapter 3 also demonstrates how the MABLAR framework differs from traditional data-driven rule generation approaches.

- After establishing the MABLAR framework, Chapter 3 – Chapter 5 propose three different variants of the MABLAR framework for generating causal rules, which covers the content of Sub-ROs associated with RO3. The proposed variants are Standard Markov blanket rule generation (MABLAR-ST, introduced in Chapter 3), Markov blanket rule generation using causal direction (MABLAR-CD, introduced in Chapter 4) and Markov blanket rule generation using causal weight (MABLAR-CW, introduced in Chapter 5), respectively. These variants of the MABLAR framework leverage different types of causal information exploited from a given data set to generate causal rules. The facets of causal relationships reflected by the rules generated from different variants are distinct. Chapter 3 – Chapter 5 clearly illustrate the differences in causal relationships captured by the rules generated from different variants, and show the scenarios in which each variant is suitable.

- Chapter 6 proposes a meta-variant of the MABLAR framework which is designed to achieve counterfactual rule generation, which covers the content of RO4. This meta-variant is called Counterfactual Markov blanket rule generation (CF-MABLAR). Rules obtained by CF-MABLAR can provide counterfactual explanations which complement the explanation provided by rules obtained by a variant of the MABLAR framework. Furthermore, Chapter 6 explains the benefits of such supplementation from the perspective of XAI.

- Chapter 7 introduces a Python library developed to facilitate users in im-

plementing different variants of the MABLAR framework, which covers the content of RO5. The Python library includes the necessary functions to implement the MABLAR framework, which helps users to implement fuzzy system which focus on exploring causal relationships between variables without requiring them to be familiar with causal discovery or fuzzy systems. The library is actively being developed, and so far, it has supported two variants of MABLAR.

- Chapter 8 summaries the work in this thesis while discussing the limitations of the current work and making suggestions for future research.

1.6 List of publications arising from this thesis

- Te Zhang, Christian Wagner, "Learning Causal Fuzzy Logic Rules by Leveraging Markov Blanket," 2021 IEEE SMC, Melbourne, Australia, 2021, pp. 2794-2799.

  The content of this paper is covered in Chapter 3.

- Te Zhang, Jingda Ying, Christian Wagner and Jonathan M. Garibaldi, "Towards causal fuzzy system rules using causal direction," 2023 FUZZ-IEEE, Songdo Incheon, Korea, 2023, pp. 1-6.

  The content of this paper is covered in Chapter 4.

- Te Zhang, Christian Wagner and Jonathan M. Garibaldi, "Explain the world – Using causality to facilitate better rules for fuzzy systems," IEEE Transactions on Fuzzy Systems (Early access)

  The content of this paper is covered in Chapter 3 and Chapter 4.

- Te Zhang, Christian Wagner, "Generating locally relevant explanations using causal rule discovery," 2024 IEEE WCCI, Yokohama, Japan, 2024

  The content of this paper is covered in Chapter 5.

- Te Zhang, Christian Wagner and Jonathan M. Garibaldi, "Counterfactual rule generation for fuzzy rule-based classification systems," 2022 WCCI, Padua, Italy, 2022, pp. 1-8.

  The content of this paper is covered in Chapter 6.

- Te Zhang, Christian Wagner, "A python library for data-driven causal fuzzy classification rule generation - *mablars*", 2025 SSCI (submitted), Trondheim, Norway.

  The content of this paper is covered in Chapter 7.

Chapter 2

Background

This chapter provides the foundation for understanding the subsequent content of this thesis, including different definitions of causality, data-driven causal discovery algorithms and explainability of fuzzy systems. More specifically, Section 2.1 summarises different definitions of causality, introduces popular causal models used in the AI area, and discusses different facets of causal relationships can be derived from a causal graph. Section 2.2 provides an overview of existing data-driven causal discovery algorithms. Section 2.3 reviews different definitions of explanations in AI. Section 2.4 discusses factors which affect explainability of fuzzy systems. Section 2.5 reviews data-driven approaches, with a particular focus on those designed to improve the explainability of fuzzy systems while maintaining their performance.

2.1 Causality in AI

2.1.1 Definitions of causality

The concept of causality is complex and on ongoing topic for discussion across disciplines [2–4], with no universally accepted definition.

The 'four causes theory' proposed by Aristotle is considered to be the first definition of causality, while the modern discussion of the definition of causality can be traced to Hume's discussion of causality in the 18th century [3, 24, 25].

Hume defines causal relationships between two events that they are always bound to occur together, which is known as the regularity theory of causality [3, 26]. Hume states that a cause should be temporally or spatially contiguous with its effect, occur prior to the effect and have a necessary connection to it. However, Hume's definition cannot distinguish correlation and causal relationships. For example, according to Hume's definition, the day will be the cause of the night. To further distinguish correlation and causal relationships, John Leslie Mackie extended Hume's definition of causality to the scenario where an effect has multiple causes by formalising the concepts of necessity and sufficiency and defining a cause as "an insufficient but non-redundant part of an unnecessary but sufficient condition (INUS condition)" [27].

However, if there are two events which satisfy the INUS condition of an effect, both of them can cause the effect and Mackie's definition cannot distinguish which one is the real cause of the effect [3]. For example, suppose both flu and cold can cause a person's fever. If a person with a fever has only the flu (or only a cold), then according to Mackie's definition, we can say that the flu (or cold) is the cause of the fever. However, if a person has both the flu and a cold, then according to Mackie's definition, we cannot distinguish the contribution of the flu and the cold to the fever separately. Therefore, we cannot determine the cause of the fever based on this definition.

In addition, both Hume's definition and Mackie's definition are based on the regularity theory, which results they focus on *type-level* causality (also called general causality), i.e., causal relationships in general causes [4]. For example, 'flu causes fever' is a statement of type-level causality. In contrast, *token-level* causality focus on the cause of a specific case. For example, '*Jack's* fever is caused by flu' is a statement of token-level causality. Definitions based on regularity theories can be violated in token-level cases. For example, a medicine can cure flu which save people's life, but may also cause the death of a small group of people due to its side effect.

To address the definition of causality in token-level cases, David Lewis defines

causal relationships based on the counterfactual theory. Lewis defines an event A is causally dependent on an event B if 1) had A not happened B would not happen either, and 2) had A happened, B would happen [28].

Lewis's definition provides a way to distinguish causal relationships at the token-level. However, Lewis's definition allows transitivity [3]. In other words, according to Lewis's definition, if event A causes event B, and event B causes event C, then A is a cause of C. This transitivity may produce a counter-intuitive cause of an event. For example, suppose Jack has a fever (event A) which can lead to his death, and Tom gave Jack a medicine (event B) which cures Jack's fever. Then Jack took a plane to New York but the plane crashed (event C) and Jack died (event D). In this case, following Lewis's definition, B is a cause of D, even though its intention is to save Jack's life. This causal relationship arises from the causal chain between B and D. Without B, C would not happen (Jack would have died because of his fever). Thus, B is a cause of C. Similarly, C is a cause of D (without C, D wouldn't happen). Therefore, despite being counter-intuitive, B is a cause of D if we follow Lewis's definition.

The above mentioned definitions of causality treat the causal relationship between two events as a deterministic relationship. In other words, in the above mentioned definitions, if the cause occurs, then the effect will definitely occur [3].However, some researchers believe that in the real world, due to the lack of a complete knowledge base, or even if the knowledge base is complete, causal relationships may not necessarily be deterministic [3]. Thus, many probabilistic theory-based definition of causality have been proposed, such as the Suppes's definition of causality [29] and the screening off causes defined by Reichenbach [30]. Broadly speaking, the definitions based on the probabilistic theory define event A as a cause of event B if $P(B) \neq P(B|A)$ where $P(B)$ means the probability that B happens and $P(B|A)$ represents the probability that B happens given that A has been observed.

Granger [31] defines that if the past value of a variable X can be used to

improve the prediction of the future values of another variable Y, we say that X 'Granger causes' Y, which now is known the Granger causality. The definition of Granger causality explicitly involves time. Thus, it has been widely applied in many areas such as finance [32] and neuroscience [33], where involves lots of time series data. However, one time series being able to predict another only suggests that there might be a causal relationship between them; it does not mean that a causal relationship necessarily exists, as this prediction is based on correlation [34]. Therefore, the Granger causality provides an effective method for analysing predictive relationships between time series data, however, it does mean true causality [3].

Galles and Pearl defined the concept of a *causal model* which describes the world by random variables and uses structural equations to represent causal mechanisms between variables in the world [35]. Based on the concept of a causal model, Halpern and Pearl proposed the 'Halpern-Pearl definition of causality', designed to eliminate imprecision or conceptual fuzziness that may arise when describing causality using natural language semantics by instead providing a precise mathematical description [36].

The above mentioned work has promoted the research of the definition of causality, however, it is still an open problem because many counterexamples can be found against these definitions [3]. Section 2.1 provides an informal definition of causality used in this thesis, and Section 3.1 will establish the specific working definition of causality used in this thesis.

Focusing on the AI area, the next section reviews popular models used to represent causal relationships between variables in the AI area, introduces the model adopted in this thesis, and further outlines the facets of causal relationships can be derived from the adopted model.

2.1.2 Causal models

To formulate causal relationships between variables, causal models are needed. Classical causal models include the potential outcome framework [37], the structural causal model [2] and the causal graph model [38]. The potential outcome framework is designed to compare the effects on the same study subject when receiving an intervention versus not receiving it [37, 39]. The structural causal model is designed to represent the causal relationships between a set of variables using a set of structural equations. Pearl [22] proves that the potential outcome framework and the structural causal model are logically equivalent.

This thesis adopts the causal graph model proposed by Pearl [38] to show the causal relationships between variables of the given data set, as the causal graph model is intuitive and has been widely used. Causal graphs are directed acyclic graphs (DAGs) where each node represents a variable. If two variables have a causal relationship, this is reflected by an edge between them, pointing from the cause to the effect.

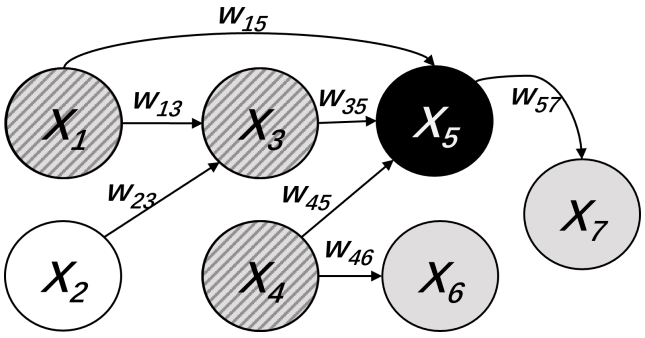


Figure 2.1: *An example of a causal graph*

Fig. 2.1 presents an example of a causal graph. The following causal information can obtained from a causal graph of a given data set:

- *Dependencies between variables*: A causal graph of a data set describes the dependencies between its variables [2]. More specifically, in a causal graph, if there is an edge between two variables, the two variables are dependent. Section 2.2.1 will further explain how causal graphs describe

the dependencies between variables.

- *Causal direction*: Causal direction represents the direction of the edges in a causal graph. Causal direction indicates cause-effect relationships between variables. In other words, if there exists a causal relationship between two variables, the causal direction between them indicates which variable is the cause and which variable is the effect. For example, in Fig. 2.1, the causal direction between 'X_5' (the black node) and 'X_3' (the grey striped node) indicates that the variable 'X_3' is the direct cause of the variable 'X_5'.

- *Causal weights*: The concept of causal weights is complex. For example, it can represent the probability of the causal relationship between variables [40] or the degree of influence of the cause on the effect [41, 42]. Although there are various definitions of causal weights, they share a common element: they describe a notion of the strength of a causal link between two variables.

- *Causal paths*: A causal path of a target variable is a path of which all edges points to the target variable. If the starting node of a causal path has no parents in the causal graph, this path is a complete causal path.

Based on the causal information which can be obtained from a causal graph, at least the following three facets of causal relationships can be derived from a causal graph:

- Local causal relationships: Given a causal graph, using the dependencies between variables information, the Markov blanket (MB) of a variable can be determined, which describes the *local causal relationships* of a target variable [43–45]. The concept of the MB was originally established by Pearl in 1988 [46]—the MB of a variable T is the minimal set of variables given which all other variables are independent of T. Within the context of causal graphs, Aliferis et al. [43] defined the MB of a

variable in a causal graph as follows: the MB of a variable in a causal graph is the set of its parents, children, and spouses. For example, in Fig. 2.1, the MB of X_5 are all grey nodes.

- 'Direct cause-effect' relationships: If the causal influence of a variable T on the target variable is not mediated by other intermediary variables, then the variable is a direct cause of T; otherwise, it is an indirect cause [47]. Thus, using the causal direction information, the direct causes of a target variable can be identified from a causal graph. More specifically, in a causal graph, variables that are directly connected to the target variable T with edges pointing towards T are the direct causes of T. In other words, 'direct cause-effect' relationships are a subset of local causal relationships.

- Causal mechanisms: A causal graph describes different causal mechanisms of a target variable. Here, the term 'causal mechanism' refers to how a set of variables influences the target variable through a specific process. In a causal graph of a given data set, each complete causal path shows the relationships between the target variable and one set of its direct and/or indirect causes. Thus, each complete causal path represents one possible causal mechanism of the target variable as captured by the causal graph.

Different facets of causal relationships derived from a causal graph can benefit AI reasoning from different aspects. For example, the local causal structure defined by the MB can be used to determine whether a variable is redundant for training an AI model in a causal sense, or indeed, whether it is causally linked to the target variable [44].

In addition, as discussed in Section 1.3 and Section 2.3, rules of fuzzy systems can provide explanations based on the relationships their capture. Thus, rules which reflect different facets of causal relationships derived from a causal graph may play different roles in fuzzy rule-based systems, especially in terms of

explainability. Chapter 3 to Chapter 5 will further discuss this in detail.

Nowadays, many data-driven causal discovery approaches have been proposed to generate a causal graph from a given data set, which are reviewed in the next section.

2.2 Classical data-driven causal discovery approaches

Given a data set $\boldsymbol{D} = \{X_1, X_2, \cdots, X_d\}$ with d variables, where X_i represents the ith variable, the aim of causal discovery is to find causal relationships between the variables within a given data set [48]. As the causal relationships between variables of a given data set can be represented by a causal graph, the aim of causal discovery can be also viewed as to find a causal graph of a given data set.

Randomised controlled trials are the gold standard for distinguishing between causal relationships and correlations [2, 49]. However, in most cases, causal relationships between variables can only be obtained through existing observational data rather than active intervention, because of the limitations of realistically possible experimental techniques. For example, when we want to confirm whether there is a causal relationship between smoking and lung cancer, we cannot force a non-smoker to start smoking.

Section 2.1 reviews work which focuses on the definition of causal relationships. These definitions can guide how to learn causal relationships. However, to discover causal relationships from data, specific data-driven approaches are needed [3], which are reviewed in this section.

2.2.1 Causal assumptions for data-driven causal discovery algorithms

As pointed out by Simon [50], causal relationships between variables obtained from data sets depend on the a priori assumptions we set for the data sets. The causal Markov assumption, the causal sufficiency assumption, the causal faithfulness assumption and the assumption about the data generation process are four popular assumptions [51]. Many data-driven causal discovery approaches require one or more of the above assumptions [39]. The four assumptions are introduced as follows:

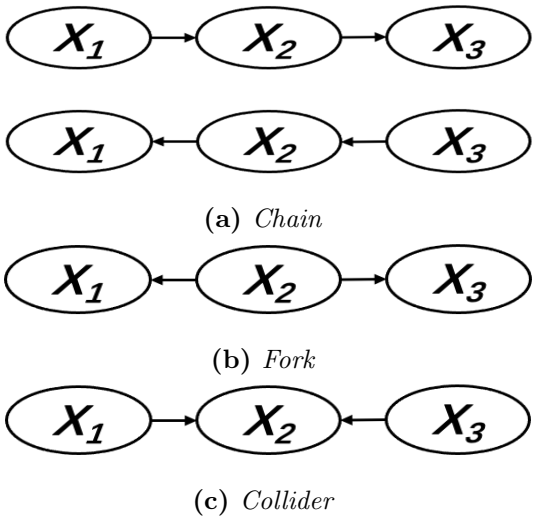
1. The causal Markov assumption: The value of a variable is independent of all other variables conditional on all of its direct causes, and these causal dependencies generate statistical dependencies [51].

2. The causal sufficiency assumption: A given data set $\boldsymbol{D}$ is considered to be causal sufficient if, for any two variables within $\boldsymbol{D}$, their direct causes are also present in $\boldsymbol{D}$ [52]. For example, suppose that Fig. 2.1 is the causal graph which represents the complete causal relationships of a set of variables $\{X_1, X_2, X_3, X_4, X_5, X_6, X_7\}$. Then, the data set $\boldsymbol{D}' = \{X_1, X_2, X_3, X_5, X_6, X_7\}$ does not satisfy the causal sufficiency assumption, because X_4, which is the direct cause of X_5 and X_6, is not included in $\boldsymbol{D}'$.

3. The causal faithfulness assumption: Variables that are causally connected in a particular way in the causal graph are probabilistically dependent [53]. More specifically, given a data set $\boldsymbol{D}$, variables X_i and X_j are independent or conditionally independent. Then in the causal graph G, which consists of all variables within $\boldsymbol{D}$ and their causal dependencies, all paths between X_i and X_j are dependency-separated (d-separated) by the appropriate variables in the data set $\boldsymbol{D}$. The concept of 'd-separation' is further explained below.

4. The assumptions about the data generation process: The assumptions about the data generation process are a new class of assumptions introduced in recent causal discovery methods to enhance the ability to discover causal relationships [54]. These assumptions include whether the data type is discrete or continuous, whether the generation mechanism is linear, nonlinear, or post-linear, and whether the noise distribution is noiseless, Gaussian, or non-Gaussian. For example, the linear non-Gaussian acyclic model (LiNGAM) (which will be further introduced in Section 2.2.3) assumes that the value of a variable is the linear sum of the value of its corresponding cause(s), plus non-Gaussian distributed noise [55].

Here, the concept of 'd-separation' is explained. Fig. 2.2 shows three basic structures of a causal graph: a) chains, b) forks and c) colliders [2]. A path P



(a) *Chain*

(b) *Fork*

(c) *Collider*

Figure 2.2: *Three basic structures of a causal graph*

is blocked by a variable set $\mathbf{Z}$ if one of the two conditions is satisfied:

- 1) There is a chain $X_1 \rightarrow X_2 \rightarrow X_3$ or a fork $X_1 \leftarrow X_2 \rightarrow X_3$ in P and the middle node X_2 is included in $\boldsymbol{Z}$.

- 2) There is a collider $X_1 \rightarrow X_2 \leftarrow X_3$. The collider node X_2 and no descendant of X_2 is included in $\mathbf{Z}$.

If a variable set $\boldsymbol{Z}$ blocks all paths between variable A and B, then A and B are *d-separated* by the set $\boldsymbol{Z}$. This means that $A \perp B|\boldsymbol{Z}$ [2]. For example, in Fig. 2.1, X_5 blocks all paths between X_3 and X_7, which means $X_3 \perp X_7|X_5$. D-separation provides a method in causal graphs to determine the independence or dependence between nodes based on their connection relationships.

Data-driven causal discovery approaches can be classified into the following three categories based on their purpose: 1) Casual structure learning approaches, 2) Causal inference approaches and 3) Confounding variables detection approaches. The rest of this section will review each category.

2.2.2 Causal structure learning approaches

Given the data set D, the goal of causal structure learning is to determine the skeleton of the causal graph of D.

Verma and Pearl [56] proposed the Inductive Causation (IC) algorithm to achieve causal structure learning. Given a data set, the IC algorithm first creates a graph where each node corresponds to a variable within the data set, but all edges are initially empty. Then, the IC algorithm adds undirected edges to the created graph based on the independence and non-independence between variables. Classical methods for testing the dependency between variables include the Chi-Square conditional independence test and Fisher's Z conditional independence test. Finally, the IC algorithm determines the directions of edges as much as possible based on the principles of avoiding cycles and not creating new colliders.

To reduce computation costs of the IC algorithm, Spirtes and Glymour [51] proposed the Peter-Clark (PC) algorithm. Different from the IC algorithm, which starts with a graph whose all edges are initially empty, the PC algorithm starts with a complete undirected graph. Moreover, conditional independence tests are only conducted between adjacent nodes (variables). This reduces the search scope of conditional independence tests required by the algorithm,

which reduces computation costs. Table 2.1 shows the procedures of the PC algorithm.

Table 2.1: The procedures of the PC algorithm

The PC Algorithm
Input: Sample data set ***D***
Output: A directed acyclic graph (DAG) representing the causal structure
Step 1: Identify the variable set ***Z*** from the sample data set ***D***
Step 2: Initialize a complete undirected graph G over the variable set ***Z***
Step 3: Test conditional independence relationships between variable pairs and remove undirected edges from G based on the identified independencies to obtain the skeleton (undirected) graph
Step 4: Identify colliders (See Fig. 2.2) in the graph using the skeleton and the independence information
Step 5: Orient the remaining edges under the constraints of: a: avoiding directed cycles, and b: not introducing new colliders

Both the PC algorithm and the IC algorithm are based on the independence between variables to determine the skeleton of the causal graph, such algorithms are also called constraint-based algorithms [39, 57]. Other constraint-based algorithms include the PC-pop algorithm [58] and PC-stable [59]. The PC-pop algorithm aims to address the issue of computational costs that the original PC algorithm faces with high-dimensional data sets, while the PC-stable algorithm is designed to resolve the problem of variable processing order dependency in low-dimensional settings encountered by the original PC algorithm.

As mentioned in Section 2.2.1, constraint-based algorithms require causal assumptions about data sets. For example, the PC algorithm requires the given data set to satisfy the causal Markov assumption, the causal sufficiency assumption, and the causal faithfulness assumption [48]. However, a given data set may not meet these assumptions, which can lead to the algorithm failing to achieve a global optimum. In addition, constraint-based algorithms cannot determine the causal direction between variables in a Markov Equivalence class[1][61].

[1] A Markov equivalence class is a set of causal graphs that encode the same set of conditional independencies [60]. For example, the two chains in Fig. 2.2 a) are within a Markov Equivalence class, because these two chains encode the same conditional independencies,

To address the above mentioned challenges, Chickering [62] proposed the Greedy Equivalence Search algorithm (GES). GES replaces conditional independence tests with a scoring criterion which measures the fitness of a causal graph to the given data set. GES begins with a random graph and iteratively adjusts it by adding or removing edges. If a modified graph has a higher score, it replaces the original graph. This process is repeated until the obtained graph has the highest score.

The GES algorithm transforms the causal structure learning problem into an optimization problem by scoring the fitness between the graph and the given data set. Such algorithms are also called score-based algorithms [39, 57]. Other score-based algorithms include MMHC [63], MMPC [64], etc. Compared with constraint-based algorithms, score-based algorithms are often more accurate when the given data set contains small number of samples [65]. However, score-based algorithms have the risk of converging to a local optimum [39].

2.2.3 Causal direction inference approaches

Suppose there exists a causal relationship between two variables X_1 and X_2. The goal of causal direction inference is to determine the direction of the edge between X_1 and X_2 in a causal graph, i.e. to distinguish between $X_1 \leftarrow X_2$ and $X_1 \rightarrow X_2$.

Shimizu etc. [55] proposed the Linear non-Gaussian acyclic model (LiNGAM) to achieve causal direction inference. LiNGAM assumes that causal relationships between variables follow a set of conditions [55]: 1) The value of an observed variable is the linear sum of its cause variables, along with a noise variable and a constant. 2) The noise variable of each observed variable follows a non-Gaussian distribution with non-zero variance and the noise variables are independent of each other. Then, Shimizu etc. [55] proposed the ICA-LiNGAM algorithm which uses the Independent Component Analysis algorithm (ICA) to solve LiNGAM.

---

i.e., $X_1 \perp X_3 | X_2$

However, the ICA-LiNGAM algorithm is sensitive to initialisations. Consequently, the ICA-LiNGAM algorithm has a risk of converging a local optimum when the selection of the initial solution is not good [39, 48]. To address this challenge, Shimizu etc. [66] proposed the DirectLiNGAM which determines the causal relationships between variables based on the independence of residuals after linear regression.

LiNGAM-based algorithms are designed to achieve causal direction inference when the causal relationships between variables are linear [55, 66]. To achieve causal direction inference in nonlinear scenarios, Hoyer etc. proposed the Additive Noise Models (ANMs) [67]. ANM assumes that the value of an effect can be represented by a function (which may not be a linear function) of the value of its cause plus an additive noise, and the noise of an effect should be independent of its causes. More specifically, given two variables X_1 and X_2, suppose X_1 is a cause of X_2, ANM assumes that the causal relationship between them can be represented by $X_2 = f(X_1) + n$, where n is the additive noise and n should be independent with X_1. Based on this assumption, ANM inference the causal direction between X_1 and X_2 as follows:

- Step 1: Perform regression analysis on X1 and X2 using $X_2 = f(X_1)+n$, where $f(x)$ is a nonlinear function and n is noise.

- Step 2: Compute residuals n using $n = X_1 - f(X_2)$.

- Step 3: Test the independence between n and X_1.

- Step 4: Determine the causal direction between X_1 and X_2 based on the results of the independence test.

In Step 4, if n is independent of X_1, ANM determines the causal direction $X_1 \to X_2$. Otherwise, $X_2 \to X_1$.

2.2.4 Confounding variable detection

Given a set of variables $\boldsymbol{D} = \{X_1, X_2, \cdots, X_m\}$, a confounding variable is an unobserved variable $X_u \notin \boldsymbol{D}$ which is a cause of at least two variables within $\boldsymbol{D}$. The goal of confounding variables detection is to determine whether there are confounding variables of $\boldsymbol{D}$.

To address the issue that a causal graph cannot represent causal relationships between variables when confounding variables are present, Richardson and Sprites proposed Maximum Ancestral Graphs (MAGs) [68]. Different from causal graphs that only include directed edges, MAG can also include bidirectional edges and undirected edges. If there exists a bidirectional edge between two variables, it indicates that there exist unobserved confounding variables between them.

Based on the concept of MAG, Spirtes et al. [51] proposed the Fast Causal Inference algorithm (FCI) to solve the confounding variables detection problem. Similar as the PC algorithm, FCI also starts with a complete undirected graph and uses the independence between variables to determine the causal relationships between variables. However, the difference between FCI and the PC algorithm is that the output of FCI is a Maximal Ancestral Graph (MAG) not a causal graph. Other Confounding variables detection algorithms include RFCI [69], FTFC [70] etc.

Since this research focuses on how to represent causal relationships between variables within a data set using rules, it only considers cases where there are no confounding variables in the given data set. For this reason, the research focuses on causal graphs rather than MAGs. However, in real-world applications, many data sets do have confounding variables. Future research will focus on how to use MAGs to address this issue.

2.3 Explanations in AI

Across AI, explanations of AI models, depending on their objectives, can be viewed to fall into two categories: *model-based explanations* and *domain explanations* [71]. Model-based explanations seek to explain the operations of AI *models* to human beings, while domain explanations seek to explain the causal relationships behind data, and ideally behind the real-world phenomena in question.

In the XAI area, The term 'explanations' usually represents model-based explanations [72, 73]. According to the ways of generating explanations, model-based explanations can be categorised into intrinsic explanations and post-hoc explanations [7].

- Intrinsic explanations: Intrinsic explanations refers to the explanations provided by a model itself. Such models are self-explainable because they have an inherent ability to be understood. That is, for a trained model, users can comprehend its decision-making process or the rationale behind its decisions without needing additional information. Self-explainable models include decision trees [74] and Mamdani fuzzy systems [75].

- Post-hoc explanations: Post-hoc explanations aim to explain the behaviour of a trained model which is complex and not self-explainable, using another explainable models or methods. Post-hoc explanation generating methods include LIME [10] and mimic learning [11, 12].

According to scope of explanations, model-based explanations can also be categorised into *global explanations* and *local explanations* [72]:

- Global explanations: Global explanations inspect the structures and parameters of models to explain how models work globally, i.e. they seek to explain models' behaviour. For example, the post-hoc explanations provided by the mimic learning method are global explanations [72]. The mimic learning method aims to mimic the behaviour of a complex

model by compressing the function of the complex model into a smaller, faster model with comparable performance [11]. This allows users to understand the behaviour of the original complex model through the compressed model.

- Local explanations: Local explanations focus on explaining an individual output of a model–for a given input. For example, a post-hoc explanation provided by LIME is a local explanation [72], because LIME aims to explain the behaviour of a black-box model for a specific input by training a linear model.

Here, the explanations provided by Mamdani fuzzy systems are further discussed. The rule base of a Mamdani fuzzy system is composed of linguistic rules. The antecedents and consequents of these linguistic rules are defined by linguistic variables, which are composed of linguistic terms and associated fuzzy sets [76]. The rules of a Mamdani fuzzy system can explain the system's behaviour using semantic meaning which humans can understand, without requiring additional models or methods to generate explanations. Thus, Mamdani fuzzy systems can be viewed as self-explainable.

From the global and local points of view, the linguistic rule base of a Mamdani fuzzy system can provide global explanations for a Mamdani fuzzy system, because it shows the relationships between variables captured by the system and demonstrates how the system maps inputs to outputs [14]. Meanwhile, for a Mamdani fuzzy systems, given a specific input, a specific set of its rules would be fired. The fired rules provide local explanations for the system, because those rules show how the system deal with the specific input [76].

Model-based explanations allow users insight into how an AI model works. Ideally however, if AI models can mimic causal processes in the real world, they have the potential to provide domain explanations [77, 78].

2.4 Explainability of fuzzy systems

2.4.1 Semantic meaning of rules

One of the reasons why fuzzy systems have the potential to deliver good explainability is that their rules can explain the system's behaviour to users through their semantic meaning. However, the semantic meaning of the rules is affected by many factors, further discussed next.

The semantic meaning of rules of fuzzy systems is highly related to the type of rules [14]. For example, there is no linguistic terms to the consequents of TSK fuzzy systems [79], which reduces the readability of rules of TSK fuzzy systems. In contrast, Mamdani fuzzy systems are more suitable for applications where the emphasis lies on model explainability, because both antecedents and consequents of their rules are represented by linguistic terms, thus providing model-based explanations to users in a human understandable way [13, 80, 81].

Another factor which relates to the semantic meaning of rules of fuzzy systems is how their 'THEN' parts are interpreted. According to the choice of implication and aggregation operators, their 'THEN' parts can be interpreted in a conjunctive way or an implicative way, corresponding respectively to conjunctive rules and implicative rules [82, 83]. Given an 'IF-THEN' rule 'IF X is low, THEN Y is High', it can be understood as 'The more X is low, the more *possible* Y is High' if it is interpreted in a conjunctive way. Following an implicative interpretation, the same rule can be understood as 'The more X is low, the more Y *is* High'.

The semantic meaning of implicative rules is more precise compared to the semantic meaning of conjunctive rules, which makes implicative rules suitable to represent causal relationships between variables, because as discussed in [2], expressing causal relationships requires minimising uncertainty in the expression. As we discussed in Chapter 1.3, people expect rules reflect causal relationships. Thus, implicative rules are more suitable for XAI. In addition,

more implicative rules lead to a more direct and specific knowledge base, which further enhances their suitability for XAI, because this is in line with a general way of thinking. However, compared to conjunctive rules, there are few practical methods for computing inference with implicative fuzzy rules [82], which restricts the applicability of implicative rules in the real world.

In the literature, most works focusing on the generation of fuzzy rules from data, such as [84, 85], focus on the conjunctive interpretation of rules. This thesis will follow the same approach broadly in order to leverage and compare to these approaches. However, we highlight that in the real-world applications, the selection of conjunctive rules and implicative rules should be problem-depended.

2.4.2 Complexity of fuzzy systems

Apart from the semantic meaning of rules, the explainability of fuzzy systems is also related to their complexity [86]. There is already a comparatively large body of literature focusing on the explainability of fuzzy systems in terms of the complexity of their rule bases [13, 80, 87]. In general, less complex fuzzy systems are considered to have a higher degree of explainability. The complexity of a fuzzy system is affected by many factors, which are further discussed bellow:

1. *Number of fuzzy rules.* A fuzzy system with more rules usually has the capacity to deliver better performance than a fuzzy system with a smaller number of rules. But it will also become more complex.

2. *Number of input variables of a fuzzy rule.* A rule with more input variables has higher complexity. A rule's antecedent complexity and thus the complexity of the rule overall increases with an increasing number of input variables. Respectively, a reduction in the number of input variables will make a rule more concise and potentially increase the interpretability. Thus, it is important to find the most concise set of variables possible,

when generating rules [87].

3. *The fuzzy partition of each input variables.* The complexity of a fuzzy system will increase if the number of fuzzy sets of a fuzzy partition of an input variables increase. Suppose there are m input variables and each input variables is divided into d fuzzy sets. A complete rule base [88] for a fuzzy system needs d^m rules. As d increases, the number of rules will increase, and the complexity of the fuzzy system will increase.

4. *The number of fired rules.* Given an input to a fuzzy system, the rules whose firing strength are not 'zero' are called fired rules. The fired rules provide the explanations of how the output of a given input is obtained by the system. Such explanations also refer to 'local explanations'. A system with a higher average of fired rules is considered to be more complex. Factors affecting the number of fired rules include types of membership functions and logical connectives etc. In general, fuzzy systems using triangle and trapezoidal MFs have a smaller number of fired rules than those using Gaussian MFs. In addition, fuzzy systems using the 'AND' logical connective have a lower number of fired rules than those using the 'OR' logical connective.

5. *The structure of rules.* The example rule of a Mamdani fuzzy system shown in Section 1.3 is a flat rule, i.e., all input variables are shown in the antecedents of the rules [13]. However, when there is a large number of inputs, rules with a hierarchical architecture can have a lower complexity than flat rules [89]. A hierarchical rule is composed of lower-dimensional rules, where each layer of the hierarchical rule corresponds to one (or a set of) lower-dimensional rules, and the output of each layer is the input for the next layer [90].

After reviewing the explainability of fuzzy systems, in the next chapter, we provide an overview of data-driven rule generation approaches for fuzzy sys-

tems.

2.5 Overview of rule generation for fuzzy systems

With the development of data collection techniques, many data-driven rule generation algorithms have been proposed to leverage abundant data. For example, Wand and Mendel [84] proposed the WM algorithm in 1992. The WM algorithm is one of the most traditional data-driven approaches, and it still is widely used as a benchmark algorithm because it provides a consistent basis for comparison [91, 92]. The WM algorithm generates a rule for each training sample and assign a weight for each obtained rule. The assigned weights are further used to remove conflicting rules. Rules which have the same antecedents but different consequent are called conflict. In cases of conflicting rules, the rule with the smaller weight is removed. Here, two or more rules conflict with each other if they share the same antecedent part while having different consequents.

One of the key advantages of the WM algorithm is its simplicity and explainability, which makes it suitable for scenarios where explainability is a key requirement [93]. However, compared to more recent approaches, such as the fuzzy unordered rule induction algorithm (FURIA) [85] (introduced in Section 2.5.3), it does not offer competitive classification performance [94]. Furthermore, when a given data set contains a large number of variables, the WM algorithm often produces a large rule base, which reduces the explainability of the obtained rule base [93, 94].

In 1992, Ishibuchi et al. [95] proposed a weighted fuzzy rule generation approach which is called distributed representation of fuzzy rules (DRFR). The DRFR assigns a weight to each rule using all training samples, which increases the robustness of the obtained fuzzy systems when facing noisy data sets com-

pared to the WM algorithm.

However, the rule bases obtained by data-driven algorithms tend to be very large to achieve a high performance, which increases the complexity of the obtained fuzzy systems and makes them prone to over-fitting [96, 97]. With the increasing demand for the explainability of fuzzy systems in many fields, many data-driven algorithms that aim to reduce the complexity of the resulting rule bases have been proposed. To help the discussion, this thesis categorise the existing methods into three groups based on their mechanisms: 1) rule reduction and compression, 2) sparse or minimal rule generation and 3) iterative rule compression approaches.

2.5.1 Rule reduction and compression approaches

Within this category, this thesis groups algorithms which are based on reducing or compressing an initially established, and usually large, rule base generated by data-driven algorithms which are designed to achieve good performance.

For example, Cordon et al. [98] proposed the accurate linguistic modeling (ALM) algorithm which uses a genetic algorithm (GA) to reduce the number of rules of the initial rule base obtained by the WM algorithm.

Alcala et al. [99] proposed the classification model based on fuzzy association rules for high-dimensional (FARC-HD). FARC-HD also adopts GA for the rule selection. It also adopts a search tree strategy [100] and a subgroup discovery strategy [100] to preselect a subset of rules from the initial rule base before implementing a GA.

Dutu et al. [101] proposed the selection-reduction rule base generation algorithm (SR) to reduce the number of rules in the rule base obtained by the WM algorithm. It first removes each rule in turn from the initial rule base and then tests the performance of the obtained fuzzy system. If the global error is reduced below a pre-defined threshold, the rule is discarded from the initial rule base.

Algorithms within this category often achieve a good trade-off between the number of rules and the performance of the model. However, they require an initial rule base, which can be time-consuming to generate.

2.5.2 Sparse rule generation approaches

Within this category, this thesis groups approaches designed a priori to generate compact rule bases from the data or from experts – without an initially (and usually large) established rule base.

For example, Chi et al. [102] proposed the CHI algorithm which is designed to address redundant rules generated by the WM algorithm by reducing fuzzy partitions through similarity-based merging to create compact rule bases.

Gou et al. proposed the improved Wang–Mendel method based on the FCM algorithm (FCAWM) [88]. FACWM generate compact rule bases by reducing the number of training samples, using clustering algorithms. More specifically, FCAWM adopts the fuzzy C-means algorithm (FCM) [103] to cluster the training data set to obtain a small data set. Then, it generates rules using the WM algorithm from that small data set.

Alonso et al. [104] proposed the highly interpretable linguistic knowledge ++ framework (HILK++). The HILK++ framework integrates the C4.5 feature selection strategy, data-driven rule generation and evolutionary computation strategies to generate a compact rule base while maintaining high model performance.

Unlike rule reduction algorithms, sparse rule generation algorithms do not require an initial rule base beforehand, which saves the cost of generating an initial rule base.

2.5.3 Iterative rule compression approaches

The approaches within this category generate a compact rule base by iterative updating using different heuristic strategies. Compared with the rule reduction

approaches discussed above, the approaches within this category also update each individual rule, including the antecedents or the fuzzy partitions of each variable.

For example, Higgins and Goodman [105] proposed the Higgins & Goodman (HG) algorithm, which generates compact rule bases by controlling the number of fuzzy partitions of each variable. The HG algorithm starts from one fuzzy partition of each variable and adds a new fuzzy set to a variable iteratively until a stopping criterion (e.g. the performance of the obtained fuzzy system is higher than a threshold) is reached. By controlling the stopping criterion, the HG algorithm can avoid generating redundant fuzzy partitions of each variables, thus reducing the risk of generating redundant rules.

Huhn et al. [85] proposed the fuzzy unordered rule induction algorithm (FURIA). FURIA iteratively updates both the parameters of the membership functions and the rule base. It uses a greedy strategy to determine the parameters of the membership functions. To make the model achieve a high performance, in each iteration, the parameters which result in the rules achieving the highest rule 'purity', which is a specialised performance index proposed in [85], are kept. After the determination of the membership functions, FURIA updates the rule bases using a pruning method to remove rules that have a negative impact on the performance of the model.

FURIA avoids fixed fuzzy partitions by learning fuzzy partitions of variables for each rule, which enhances its performance on overlapping and noisy data [106]. Fuzzy systems obtained by FURIA can achieve high classification accuracy with a compact rule base. However, as membership functions are determined at the specific level of each rule, rules generated by FURIA do not offer the same level of semantic explainability of approaches which establish fixed membership functions for linguistic terms a prior– such as the WM approach [84].

Porebski [107] proposed a method which measures the quality of the fit of the membership function to the training data. Then, based on the proposed measurement method, the author proposes the EasIeR algorithm, which iter-

atively updates the rule base using this quality. In each iteration, rules whose quality is higher than a threshold are used to construct a rule base. The final rule base is that which makes the model achieve the highest performance while containing the least number of rules.

However, determining the stopping criteria is not always straightforward. Stopping too early may lead to a suboptimal solution, while stopping too late increases computational costs.

2.6 Summary

This chapter provides background information and an overview of the existing literature that has been used in this thesis.

The first aspect is to provide the background about causal discovery within the context of AI. This chapter first summaries different definitions of causality across different areas. Then this chapter introduces the causal model used in this thesis, i.e., the causal graph model, and discusses different facets of causality can be derived from a causal graph. Finally, this chapter reviews established data-driven causal discovery algorithms for causal structure learning, causal direction inference and confounding variables detection.

The second aspect is about the explanations in AI and the explainability of fuzzy systems. This chapter discusses different types of explanations of AI and discusses the explainability of fuzzy systems from both the semantic meaning aspect and the complexity aspect.

The third aspect is the overview of data-driven rule generation approaches which focus on improving the explainability of fuzzy systems. This chapter first reviews classical data-driven rule generation algorithms. Then, this chapter categorise the existing approaches which focus on improving the explainability of fuzzy systems into three groups based on their mechanisms and review the approaches within each group.

Following this background overview, in the next chapter, a fuzzy rule genera-

tion framework for causal rule generation is established and analysed. Then, a specific variant of the established framework is proposed. The proposed variant leverages information about the dependencies between variables in a causal graph of the given data set to generate rules. The obtained rules show the local causal relationships of the target variable.

Chapter 3

Towards causal fuzzy logic rules by leveraging Markov blankets

As mentioned in Chapter 1, rules of fuzzy systems are expected to reflect causal relationships between variables. However, how to automatically generate causal rules for fuzzy systems from a given data set is still an open problem.

Section 2.1 discusses the causal information that can be extracted from a causal graph of a data set and used to distinguish different facets of causal relationships. Section 2.2 summaries data-driven approaches which can be used to generate a causal graph from a data set. Building on these work, this chapter focuses on how to generate causal rules using a data-driven approach, leveraging causal information extracted from a given data set.

To achieve this goal, this chapter establishes a causal rule generation framework which we refer to as the MArkov BLAnket Rule generation framework (MABLAR), and further introduces a variant of MABLAR, called MABLAR-ST.

MABLAR-ST uses the information of dependencies between variables to generate rules which reflect local causal relationships, which addresses Sub-RO1 of RO3 as shown in Fig. 1.2. It refines the original data set to a subset which only contains variables which are causally related to the variables. In further

chapters, more refined and specific variants are introduced.

The rest of this chapter is organised as follows: Section 3.1 provides an overview of MABLAR. Section 3.2 discusses the motivation of MABLAR-ST, i.e., the standard MABLAR. Section 3.3 introduces the process, rationale and targeted benefits of MABLAR-ST. Section 3.5 shows a series of experiments and associated results, designed to demonstrate that and evaluate how MABLAR-ST can reduce the risk of generating spurious rules. Finally, Section 3.7 presents a discussion and the summary.

3.1 Overview of the MABLAR framework

The MABLAR framework, or in short, MABLAR from here on, is designed to achieve causal rule generation by removing variables which are not causally related to the target variable. Fig. 3.1 shows a high-level view of the comparison between classical data-driven rule generation approaches and the MABLAR framework.

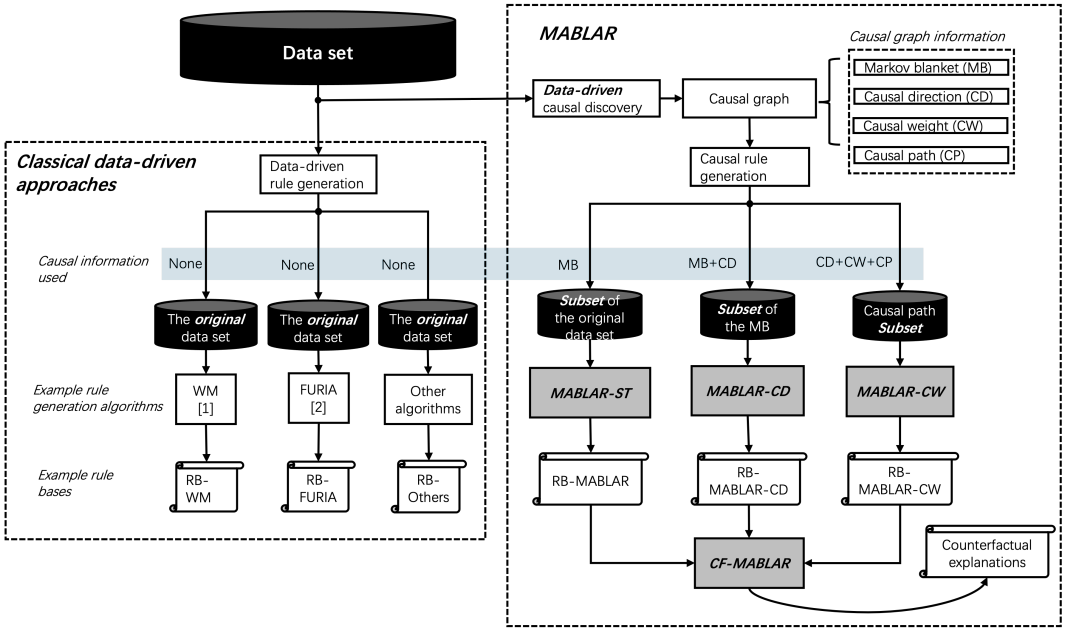


Figure 3.1: *Comparison between classical approaches and the MABLAR framework*

In general, as shown in the dashed rectangle on the left side of Fig. 3.1, classical data-driven rule generation approaches use data sets directly, exploiting

correlations between variables to achieve good performance. As a result, the obtained rule bases tend to be large, modelling correlations between variables, limiting explainability of the resulting rules and risking problems such as overfitting.

To alleviate this issue, the MABLAR framework focuses on causal relationships between variables. More specifically, as shown in the dashed rectangle on the right side of Fig. 3.1, the framework is designed to start by identifying and exploiting causal information from a given data set. Following this, the original data set is transformed, removing variables such that only variables with a causal link to the target output variable(s) remain.

MABLAR has three different variants and one meta-variant (the grey rectangles shown in Fig. 3.1). They are the standard MABLAR (MABLAR-ST), MABLAR-causal direction (MABLAR-CD), MABLAR-causal weight (MABLAR-CW) and counterfactual MABLAR (CF-MABLAR, the mete-variant), respectively. They leverage different types of causal information for generating rules that reflect different facets of causal relationships, which are further explained as follows:

- MABLAR-ST: MABLAR-ST is designed to generate rules which focus on *local causal relationships* of the target variable, i.e., the output variable of the system. To achieve this goal, it leverages *the MB information* to refine the original data set. All non-MB variables of the target variable are removed by MABLAR-ST. MABLAR-ST is further introduced in the rest of this chapter.

- MABLAR-CD: MABLAR-CD is designed to generate rules which focus on *the direct cause-effect relationships* of the target variable. To achieve this goal, it leverages both *the MB information* and *the causal direction information* to refine the original data set. All non-MB variables and non-direct causes of the target variable are removed by MABLAR-ST. MABLAR-CD is further discussed in Chapter 4.

- MABLAR-CW: MABLAR-CW focuses on local explanations provided by fuzzy systems. It is designed to generate rules which reveals *the local causal mechanism* for specific inputs. To achieve this goal, MABLAR-CW leverages the MB information, the causal direction information, the causal weights information and the causal path information to construct different subsets of the original data set. MABLAR-CW is further discussed in Chapter 5.

- CF-MABLAR: CF-MABLAR is designed to generate rules which provide counterfactual explanations to complement the explanations provided by the rules generated by a variant of the MABLAR framework (e.g., MABLAR-ST). It is based on a rule base obtained by a variant of the MABLAR framework (e.g., MABLAR-ST). CF-MABLAR is further discussed in Chapter 6.

As discussed in Section 2.1, there is no universally accepted definition of causality. The MABLAR framework uses data-driven causal discovery algorithms to identify causal relationships between variables, and the definition of causal relationships is inherited from the adopted algorithms. More specifically, throughout the rest of this thesis, the adopted definition of causality is the one from the PC algorithm (introduced in Section 2.2). In the PC algorithm, pairs of variables that satisfy the *causal Markov assumption* (introduced in Section 2.2.1) are viewed as causal structures [2, 3, 51].

In the MABLAR framework, the causal relationships identified by the adopted data-driven causal discovery algorithms are ideally expected not only hold within the given data set but also in the real world. However, it should be highlighted that due to technical or other reasons, the collected data set may exhibit bias or contain incomplete variables, which results in causal relationships discovered by algorithms that may not correspond to actual causal relationships in the real world. Similarly, causal discovery algorithms are not perfect and carry their own assumptions. In other words, different causal dis-

covery algorithms may or may not identify all, and only the correct causal relationships present in a data set [2, 108].

After establishing the overall MABLAR framework, the remainder of this chapter introduces the first variant of MABLAR, i.e., MABLAR-ST, and the next section introduces the motivation of MABLAR-ST.

3.2 Motivation of MABLAR-ST

As mentioned in Section 1.3, many rule-based systems are obtained by data-driven rule generation approaches and the obtained systems usually have good performance. However, as discussed above, classical rule generation approaches have the risk of generating spurious rules. Such rules decrease the explainability of obtained systems, because they increase the complexity of obtained systems. Although many data-driven approaches have been developed to decrease the complexity of obtained systems as introduced in Section 2.5, most of them still focus on the performance of the obtained systems and work looking into the actual nature of rules is still limited.

When referring to the *nature* of rules, as mentioned in Section 1.3, rules are expected to reflect causal relationships between their input variables and output variable(s), because humans' understanding of *rules* is that they *do* reflect causal relationships [109]. Rule bases containing rules which only reflect causal relationships usually have less number of rules compared to those containing rules which conflate causal relationships and correlations. Thus, rules which reflect causal relationships between variables are not only more in line with human way of thinking, but also helpful for reducing the complexity of rule bases, which improves the explainability of the obtained systems.

Rules which reflect causal relationships between variables also have the potential to tell users what is happening in the real-world processes modelled by the AI system. As noted in Section 2.1, causal relationships capture the links between variables in real-world processes. Where data sets exist which include

these variables, it is at least in principle possible to design AI systems which model these causal relationships. When these AI systems are explainable, the explanations of their underlying model will map to explanations of the real-world process which they capture [78]. In other words, such AI systems have the potential to deliver domain explanations to users. For example, Ziatdinov et al. report a case where a causality-based machine learning algorithm uncovers a real-world process in the domain of ferroelectric materials from the data which had not been discovered by domain experts before [77].

However, it is still an open problem how to generate rules which reliably captures causal relationships between the variables within a given data set, not least because of the difficulty of distinguishing between correlation and causality through statistical means. To address this issue, this chapter introduces MABLAR-ST.

MABLAR-ST is designed to leverage the causal information which relates to the dependencies between variables for causal rule generation. More specifically, as shown in Fig. 3.1, MABLAR-ST first identifies the MB of the output variable(s) from the input variables. This is achieved by using the dependencies between variables information shown in a causal graph of a given data set. The input variables within the MB seem are causally related to the output variable(s). Then, MABLAR uses the data within the MB subset to generate rules with the view to articulate the nature of the causal relationships between the variables.

Ideally, MABLAR can minimise the generation of spurious rules (based on spurious correlations of variables) by restricting rule learning to (only) the variables which are causally linked to the system output. MABLAR-ST is designed to address Sub-RO1 of RO3 (see Section 1.4 and Fig. 1.2), as detailed in the following sections.

3.3 MABLAR-ST – The standard Markov blanket rule generation framework

In this section, the details of MABLAR-ST are introduced. The notations and acronyms used in this chapter are shown in Table 1.2

3.3.1 The process of MABLAR-ST

MABLAR-ST is a four-step data-driven rule generation framework which is designed to capture local causal relationships between the system output variable and all variables within its MB, and thus reducing the risk of generating spurious rules.

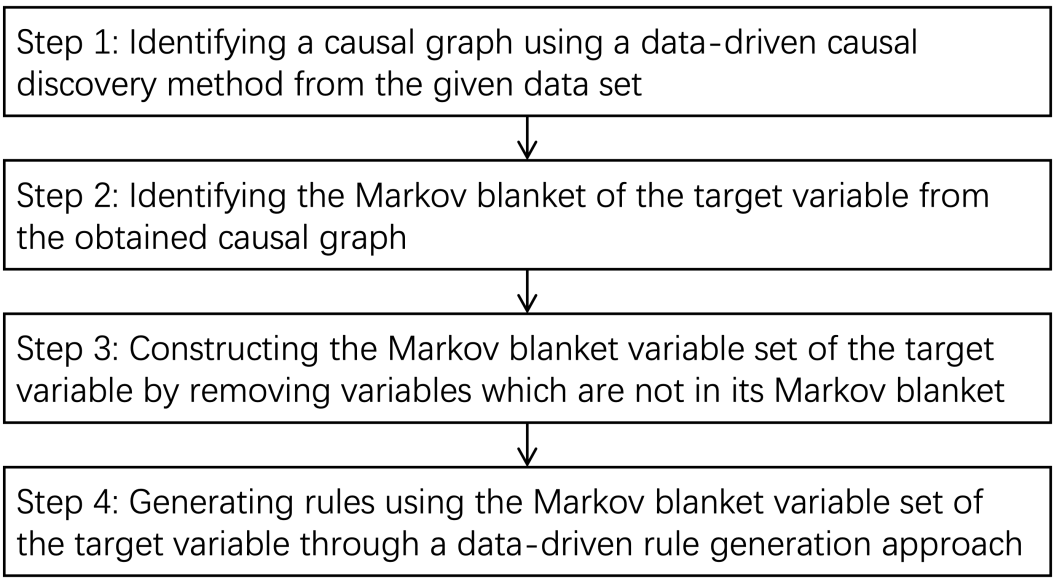


Figure 3.2: *The process of MABLAR-ST*

Fig. 3.2 presents the process of MABLAR-ST. Step 1 of MABLAR-ST is to identify the causal graph of the given data set using data-driven causal discovery approaches. Then, in Step 2, MABLAR-ST identifies the MB of the system output variable from the obtained causal graph. In Step 3, MABLAR-ST constructs $\boldsymbol{D}_{MB}$ by removing non-MB variables from $\boldsymbol{D}$. Finally, in Step 4, using data-driven rule generation approaches (e.g. the WM algorithm), MABLAR-ST automatically generates rules from $\boldsymbol{D}_{MB}$.

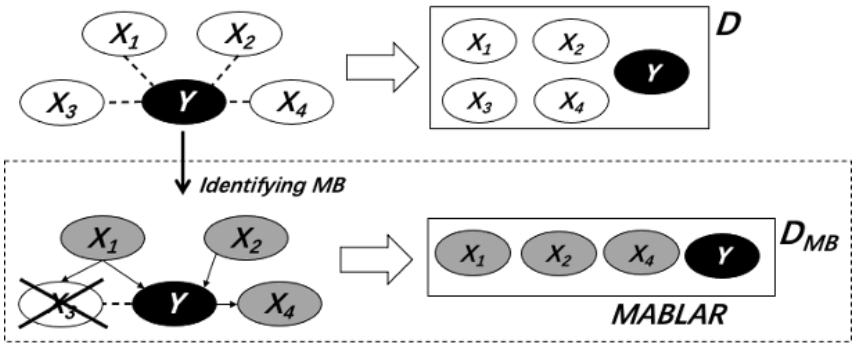


Figure 3.3: *The rationale of MABLAR-ST*

3.3.2 Rationale of MABLAR-ST

MABLAR-ST is designed to generate rules focused on the local causal structure of the output variable, which is achieved by removing variables which are not causally related to the output variable. Fig. 3.3 presents the rationale of MABLAR-ST.

In Fig. 3.3, dashed lines represent correlations between variables and solid arrows between variables represent causal relationships, and the Y is the output variable. All the variables within $\boldsymbol{D}$ are correlated with Y in this example. However, in real-world applications, $\boldsymbol{D}$ often contains variables which exhibit strong correlation with the output variable but which are not actually causally linked with it.

The relationships between such variables and the system output variable are routinely captured as rules by classical data-driven rule generation algorithms (see Section 1.3). Therefore, if a classical data-driven algorithm is directly applied to $\boldsymbol{D}$, the obtained rule base has the risk of containing spurious rules, i.e. rules which reflect spurious correlations between variables.

In contrast, MABLAR-ST removes variables which are not causally related to the output variable by leveraging the actual or approximate[1] causal graph of the data set, resulting rules obtained by MABLAR-ST only contain variables which are causally relevant with the system output variable. Fig. 3.3 presents an example of a causal graph of the original data set. As shown in Fig. 3.3, X_3

[1]In the sense of identified from data but not necessarily being the true real-world causal structure

is removed by MABLAR-ST and all remaining variables are causally related with the system output variable.

To identify the variables which are causally related with the system output variable, MABLAR-ST needs to find the causal graph of D, i.e., implement causal discovery on $\boldsymbol{D}$, in Step 1 as shown in Fig. 3.2. This is achieved by using data-driven causal discovery algorithms, such as the PC algorithm introduced in Section 2.2.

It should be highlighted that most data-driven causal discovery algorithms perform causal discovery based on strong assumptions in respect to the given data set (see Section 2.2) [48]. For example, the PC algorithm assumes that a given data set should satisfy the causal Markov assumptions and performs causal discovery based on the results of tests for conditional independence among variables [51]. However, in real-world applications, the data set may not necessarily satisfy the assumptions of the adopted algorithm, which may result a sub-optimal causal graph. In this thesis, this means that the fuzzy systems obtained by MABLAR-ST have the risk of generating rules which approximate—but may not reflect perfectly—causal relationships between variables, which may reduce model performance.

The design and selection of causal discovery algorithms is in of itself an active research area and out of range of this thesis. See [108, 110] for recent overviews, including on the selection of causal discovery algorithms.

3.3.3 Targeted benefits of MABLAR-ST

Rules obtained through MABLAR-ST can capture (or at least approximate) the local causal relationships between the variables, which is expected to benefit the obtained fuzzy systems in the following ways [92, 111]:

1. ***Reducing rule base complexity***: When fuzzy systems are obtained directly based on correlations, they have the risk of containing *spurious rules* in their rule bases. The antecedents of spurious rules contain vari-

ables that are not in the MB of the system output variable. Hence, spurious rules increase the complexity of rule-based models, which negatively impacts the explainability of fuzzy systems, may increase computational complexity and increase the risk of overfitting.

In contrast, spurious rules are largely absent from the rule bases obtained by MABLAR-ST. Thus, the complexity of fuzzy systems based on MABLAR-ST rules is commonly lower or equal to those obtained by correlation-based approaches as they contain fewer rules.

2. ***Reducing individual rule complexity***: The antecedents of spurious rules tend to contain variables which are correlated–but not necessarily causally related–with the variables in the consequents. This increases the individual rule's complexity. In contrast, MABLAR-ST rules only contain variables from with the MB. Thus, rules obtained by the MABLAR-ST framework are shorter or at least equally short in comparison to traditionally derived rules.

Above mentioned targeted benefits are further verified by the results shown in Section 3.5.2.

3.4 MABLAR-ST illustration – Diabetes diagnosis

Prior to evaluating MABLAR-ST using benchmark data sets (see Section 3.5), a real world data set on diabetes (the Pima Indian diabetes data set) is used to provide an end-to-end illustration of MABLAR-ST's intended use in a real-world context.

3.4.1 Problem and the data set

Predicting an individual's risk of developing diabetes and taking timely preventive measures is an important task [112]. To study how to predict whether women from the Pima Indian community near Phoenix, Arizona, U.S, have diabetes, W. Smith et al. [113] collected data on women over the age of 21 and formed the Pima Indian Diabetes (PID) data set.

The PID data set is available on the Kaggle website [114]. The data set can be used to diagnostically predict whether a Pima Indian woman would develop diabetes Type-2 diabetes [114]. The data set contains 768 samples and 9 variables (8 input variables and 1 binary output variable). The input variables are 'Pregnancies', 'Glucose', 'Blood Pressure', 'Skin Thickness', 'Insulin', 'BMI', 'Diabetes Pedigree Function' (DPF) and 'Age', respectively. More details about the data set (e.g., study population or case selection criteria) can be found in [113].

3.4.2 Rule generation using MABLAR-ST

Step 1 of MABLAR-ST is to identify the causal graph of the PID data set, as shown in Fig. 3.4. In this chapter, the PC algorithm (introduced in Section 2.2.2) is adopted to achieve this goal. The right part of Fig. 3.4 is the obtained causal graph of the PID data set.

It should be noted that the causal graphs obtained may vary when different causal discovery algorithms are used. The selection of causal discovery algorithms should be problem-dependent and is out of range of this thesis. Section 2.2 discusses the advantages and disadvantages of some classic algorithms for reference. The reader is referred to [108, 110, 115] for recent more comprehensive overviews, including the selection of causal discovery algorithms.

Step 2 of MABLAR-ST is to identify the MB of the system output variable, as shown in Fig. 3.5. As introduced in [114], the system output variable of the PID data set is the 'Diabetes' node which is marked as a black node in

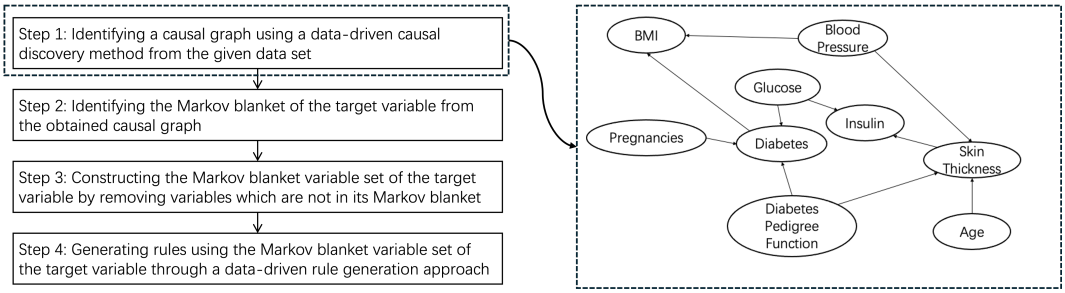


Figure 3.4: *Step 1: Identification of a causal graph of the PID data set*

Fig. 3.5. And according to the definition of MB (see Section 2.1), the MB of the 'Diabetes' node consists of all grey nodes in Fig. 3.5, i.e., 'BMI', 'Blood Pressure', 'Glucose', 'Pregnancies' and 'DPF'.

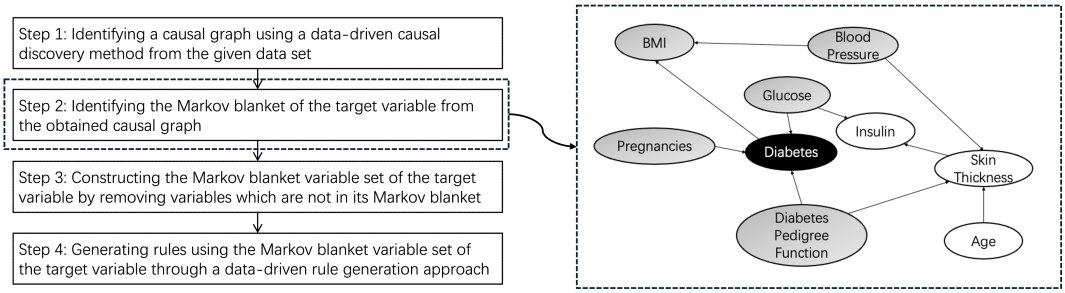


Figure 3.5: *Step 2: Identification of the Markov Blanket of the Diabetes variable*

In Step 3, the identified MB variables construct $\boldsymbol{D}_{MB}$, as shown in Fig. 3.6.The target variable is marked as red in $\boldsymbol{D}_{MB}$ to distinguish it from its MB.

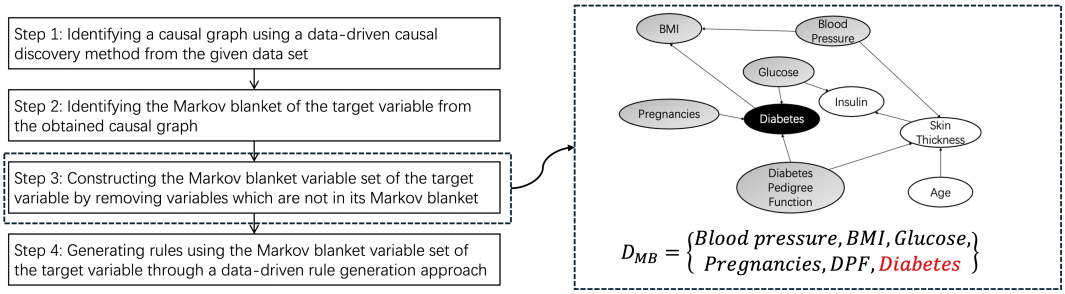


Figure 3.6: *Step 3: Construction of $\boldsymbol{D}_{MB}$ of the PID data set*

Finally, in Step 4, as shown in Fig.3.7, MABLAR-ST generates rules from $\boldsymbol{D}_{MB}$ using a rule generation algorithm. In this example, the Mamdani fuzzy system is adopted as the base model and the WM algorithm is adopted for rule generation.

Trapezoidal and triangular membership functions are adopted in this example, because they are considered to be suitable for scenarios that focus model

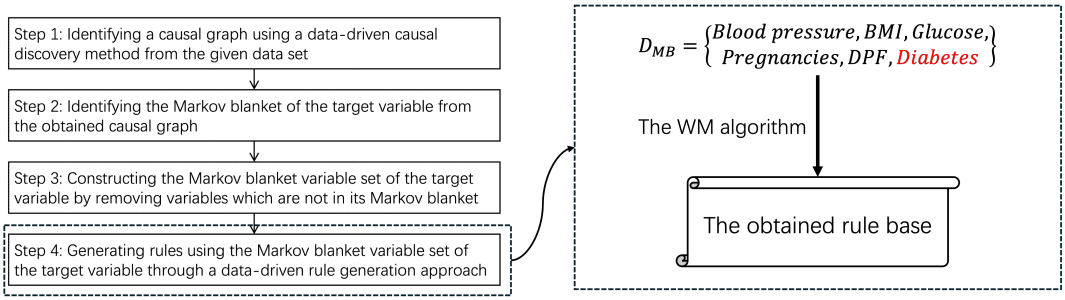


Figure 3.7: *Step 4: Generation of causal rules for the PID data set*

explainability [13]. Fig. 3.8 shows the fuzzy partition of the 'Blood Pressure' variable as an example. Note that all variables have been normalized to [0,1]. For the convenience of showing rule bases, all variables are divided into three fuzzy sets, which also facilitates the assignment of linguistic labels to each fuzzy partitions.

The FCM algorithm [103] is used to estimate the parameters of membership functions. More specifically, for each data set, FCM is first used to cluster all data points into three clusters on each variable as each variable is divided into three fuzzy sets. The value of the corresponding cluster centres ranked from low to high are noted as v_1, v_2 and v_3, respectively. Then, the membership functions in Fig. 3.8 are defined as follows:

$$\mu_{low}(x) = \begin{cases} 1, & x \leq v_1 \\ \frac{x-v_2}{v1-v2}, & v1 < x < v_2 \\ 0, & v_2 < x \end{cases} \tag{3.1}$$

$$\mu_{medium}(x) = \begin{cases} 0, & x \leq v_1 \\ \frac{x-v_1}{v2-v1}, & v1 < x \leq v_2 \\ \frac{x-v_3}{v2-v3}, & v_2 < x \leq v_3 \\ 1, & v_3 < x \end{cases} \tag{3.2}$$

$$\mu_{high}(x) = \begin{cases} 0, & x \leq v_2 \\ \frac{x-v_2}{v3-v2}, & v2 < x < v_3 \\ 1, & v_3 < x \end{cases} \tag{3.3}$$

In this example, the linguistic labels '*low*', '*medium*' and '*high*' are assigned to each fuzzy partition incrementally 'from left to right', where (3.1), (3.2) and (3.3) are the corresponding membership functions. For discrete variables, the singleton membership function is adopted and the linguistic label of each singleton fuzzy partition is the same as the original linguistic label. In this example, the output variable of the PID data set is a discrete variable and its original linguistic labels are 'Yes' and 'No'. Thus, the corresponding linguistic labels are 'Yes' and 'No'.

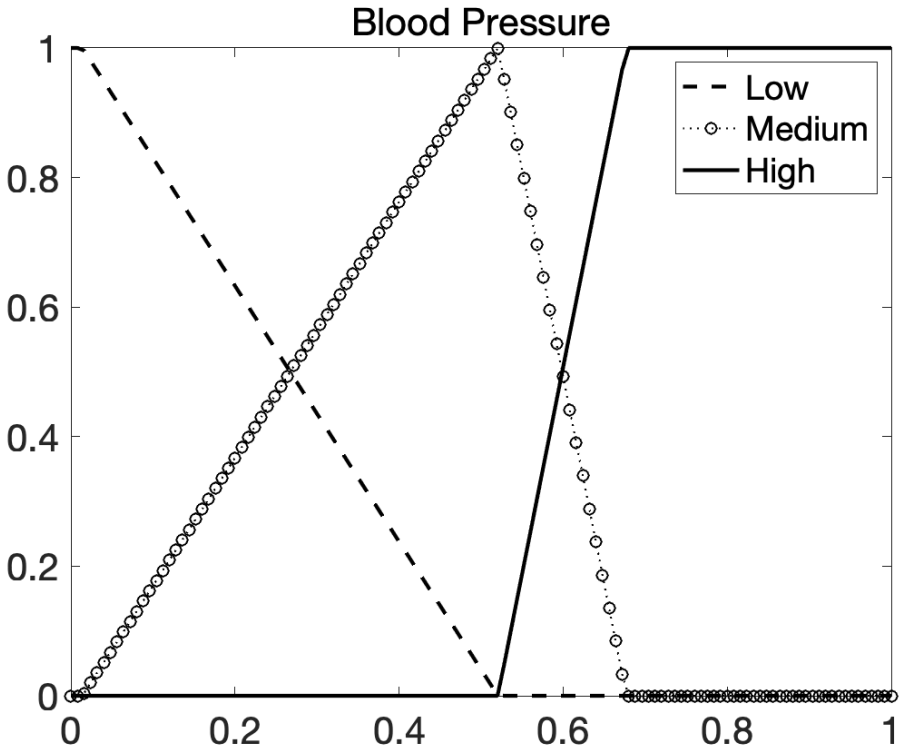


Figure 3.8: *The fuzzy partition on the Blood Pressure variable*

3.4.3 Performance of MABLAR-ST on the Pima Indian diabetes data set

This subsection shows the performance of MABLAR first, because the system obtained by MABLAR-ST should not only provide better explanations but also has comparable performance to classical rule generation approaches. The average classification accuracy over 5-fold cross-validation is adopted as the performance index, with the obtained system obtained achieving an accuracy of 0.6843 ± 0.0172(mean $\pm$ std). As a reference, the system obtained using the WM algorithm alone achieves an accuracy of 0.6927 ± 0.0332. Thus, the classification performance of MB-WM is comparable to WM.

For convenience, in the rest of this chapter, MB-WM and WM are used to represent the system obtained by MABLAR-ST and the WM algorithm, respectively.

3.4.4 Model-based explanations of the Pima diabetes classification fuzzy system

To facilitate further description and discussion, Table 3.1 shows 4 rules selected from the rule base of MB-WN obtained in the first fold. Among them, two rules have the consequent 'Yes', while the other two have the consequent 'no'. Moreover, for ease of comparison, the antecedents of rules with the same consequent differ by only one condition.

Table 3.1: Rules in MB-WM on the PID data set

	Pregnancies	Glucose	Blood pressure	BMI	DPF	Diabetes
R1	High	High	High	High	Medium	Yes
R2	Medium	High	High	High	Medium	Yes
R3	Low	Low	Low	Low	Low	No
R4	Medium	Low	Low	Low	Low	No

From the global explanation aspect, rules in Table 3.1 show that the obtained system uses 'Pregnancies', 'Glucose', 'Blood Pressure', 'BMI', and 'DPF' to

determine how a sample should be classified. By reading the rules, it is possible to get good insight of how the obtained Mamdani fuzzy system works. In our case, conjunction (i.e. *and*) is adopted as the logical connective when using the WM algorithm to generate rules. Thus the linguistic connection the between different variables shown in the rules is shown as 'AND'.

The system obtained by MABLAR-ST also provides an explanation for the prediction of each individual sample. Given a sample, e.g. $x_{example} = [0, 60, 40, 18, 0.14]$, the output of the system–in this case: 'No diabetes within 5 years' can be generated. It can be traced that rule R3 in Table 3.1 is fired for this sample (For convenience, in this example, only the rule with the highest firing strength is discussed). By reading the fired set of rules, a strong sense is gained of why the given sample is classified into 'No'. Here, it can be known that the sample is classified into 'No' because the value of all its variables are low. In this case, the system provides the local explanation of the sample $\boldsymbol{x}_{example}$ to users through R3 in Table 3.1.

The model-based explanations provided by MB-WM are further compared with those based on the WM approach. Table 3.2 shows the number of rules in the rule base obtained in each fold of MB-WM and WM and Table 3.3 shows 4 rules selected from the rule base obtained by WM in the first fold. Table 3.2 highlights that, although the rule bases obtained by MB-WM still containing a relatively large number of rules, the number of rules in the rule bases of MB-WM is substantially lower than that of WM across all folds. Table 3.1 and Table 3.3 show that the number of inputs of MB-WM (5 inputs) is also lower than for WM (8 inputs). Thus, MB-WM provides both better global, and local model-based explanations than WM, because MB-WM has lower complexity than WM on both rule base and individual rule levels (See Section 3.3).

Table 3.2: Number of rules in each fold

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Mean (std)
MWM	137	145	140	143	143	141.6 (3.13)
WM	398	408	401	392	398	379.6 (48.6)

Table 3.3: Rules in WM on the Pima Indian diabetes data sets

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DPF	Age	Diabetes
R1	High	High	Medium	Low	Low	Low	Medium	Low	Yes
R2	Low	Medium	Medium	High	Medium	High	High	Medium	Yes
R3	Low	Low	Medium	Medium	Low	Low	Low	Low	No
R4	Low	Low	Medium	Medium	Low	Medium	Low	Low	No

3.4.5 Domain-based explanations of the Pima diabetes classification fuzzy system

The MB leveraged as part of MABLAR-ST and shown in Fig. 3.5 indicates that, given the data set, the Type-2 diabetes status of a Pima Indian woman is causally related to her number of pregnancies, glucose, blood pressure, BMI and DPF–i.e. it is these variables which one should consider to determine whether or not an individual would develop diabetes within 5 years.

Conversely, the MB indicates that the value of insulin, skin thickness and age are not needed to be measured, because they are not found to be causal factors for Type-2 diabetes. Among the measured variables, take the domain knowledge embedded in R1 of Table 3.1 as an example–it reflects that if the values of pregnancies, glucose, blood pressure and BMI are very high and the value of DPF is medium, the individual probably would develop Type-2 diabetes within 5 years. It is noted that there is potential here for developing MABLAR-ST to a hierarchical rule-based system generation approach [89]–something expected to be considered in the future.

The explanations provided by MB-WM and WM are further compared from the domain explanation aspect. According to [116–118] (and some public resources e.g., [119]), the recognized risk factors of type-2 diabetes, which are also included in the actual Pima Indian diabetes data set are blood pressure, BMI, DPF[2] and age. Table 3.1 indicates that:

1. Although the variable 'age' is not included in the rules of MB-WM, most

[2]The DPF is a function proposed by the authors of [113] which is used to evaluate the family diabetes history. The higher the DPF value of a person, the more relatives of theirs have developed Type-2 diabetes. More detail can be found in [113]

variables recognized to increase the risk of type-2 diabetes are included in the antecedent of MB-WM (3 out of 4).

2. Compared with rules in WM (as shown in Table 3.3), the variables 'Insulin', 'Skin thickness' are removed from the rules in MB-WM, as expected based on recognized risk factors.

3. It is noted that the rules in MB-WM do still contain two variables 'Pregnancies' and 'Glucose' which are not recognized risk factors of type-2 diabetes, however, MB-WM contains less non-recognized risk variables than WM.

Overall, while not perfect, MB-WM provides a clear improvement. It is also worth highlighting that the variable 'Glucose', which represents the plasma glucose concentration at 2 Hours in an oral glucose tolerance test, is usually used to diagnose whether a person *has* type-2 diabetes. According to [120, 121], diabetes has been defined as a plasma glucose concentration greater than 200 mg/dl over two hours, following the ingestion of 75 gm of a carbohydrate solution. In other words, although the variable 'Glucose' is not a risk factor, it is causally related to the 'Diabetes' variable.

Note that, in Fig. 3.5, the causal relationship between 'BMI', 'Glucose' and 'Diabetes' is 'Glucose $\rightarrow$ Diabetes $\rightarrow$ BMI'. However, according to [117], the correct causal relationship should be 'BMI $\rightarrow$ Diabetes $\rightarrow$ Glucose'. A possible reason for this phenomenon is that these two causal structures belong to the same Markov equivalence classes, which cannot be determined by the adopted PC algorithm (See Section 2.2).

To sum up, although MABLAR-ST is designed to generate rules which reflect the local causal structure of the output variable, the domain knowledge users learned from the above MABLAR-ST rules may of course not be the perfect domain knowledge. However, rules obtained by MABLAR-ST still help users approximate it efficiently.

Following this illustration of deploying MABLAR-ST in a real-world context, a series of formal experiments is conducted to evaluate MABLAR-ST, both in terms of statistical accuracy, and explainability, in the next section.

3.5 Experiments

The previous section uses the PID data set as an example to show how MABLAR-ST works and compares it with a classical approach, i.e., the direct application of the WM algorithm to the data set. This section compares MABLAR-ST with the WM algorithm on more data sets to further verify whether MABLAR-ST can provide the benefits discussed in Section 3.3.3 to the obtained fuzzy systems.

3.5.1 Experiment setting

All data sets are from the UCI machine learning repository [122]. To enable detailed analysis and discussion, the data sets used in this chapter all have a small number of features. The numbers of samples, input variables, and classes of each data set are shown in Table 3.4. The average classification accuracy over five-fold cross-validation with different data partitions is used as the performance index. To keep the original class proportions in the training data set of each fold, stratified sampling is used.

Table 3.4: Data sets used in our experiments

Data set	# Samples	# variables	# Classes
HTRU2	17898	9	2
Breast	699	10	2
Wine	4898	12	2
Iris	150	5	3
Dry bean	13611	17	6

Trapezoidal and triangle membership functions are adopted as they facilitate explainability as discussed in [13]. These are distributed in the same way as

detailed in Section 3.4.2. Their linguistic labels are also assigned in the same way as detailed in Section 3.4.2. One can adopt different membership functions for specific problems. All variables have been normalized into [0,1].

The WM algorithm is selected as the benchmark rule generation approach and the rule generation approach for MABLAR-ST, because the WM algorithm is among the first data-driven rule generation method and is one of the benchmark methods in the field [123]. Since this chapter focuses on classification problems, the system output variable is used to represent the class in the rest of this chapter. For causal discovery, the PC algorithm is adopted to identify the causal graph of each data set.

3.5.2 Performance and explanations

This section shows the results obtained by MB-WM and WM. The numbers of obtained MB variables of each output variable in each data set are shown in Table 3.5. Note that the MB is determined based on the full data sets. The systems obtained by MABLAR-ST have a 'MB-' prefix in Table 3.5 (Similar for the rest tables of this chapter).

Table 3.5: Number of inputs of MB-WM and WM

	MB-WM	WM
HTRU2	4	8
Breast	6	9
Wine	5	11
Iris	3	4
Dry bean	5	16

Table 3.5 shows that the number of input variables is substantially reduced for all data sets. Thus, the rules obtained using the MB will have a smaller number of input variables in the antecedent part than using the original data set, which means the rules using the MB will be more concise and have lower complexity.

Of course, lower complexity of rules on its own is not necessarily valuable.

In other words, a system may have a very small number of rules but also very low performance, making it unusable. The goal of MABLAR-ST is to maximise explainability while achieving strong performance. The latter is defined as performance which is comparable to the performance achieved by the standard rule base optimisation approaches, excluding a focus on the MB and causality. The performance of the fuzzy systems obtained by the WM algorithm and MABLAR-ST is compared. The performance index used in this section is the same as Section 3.4.3, i.e., the average classification accuracy over 5-fold cross-validation. All features were normalized $[0, 1]$.

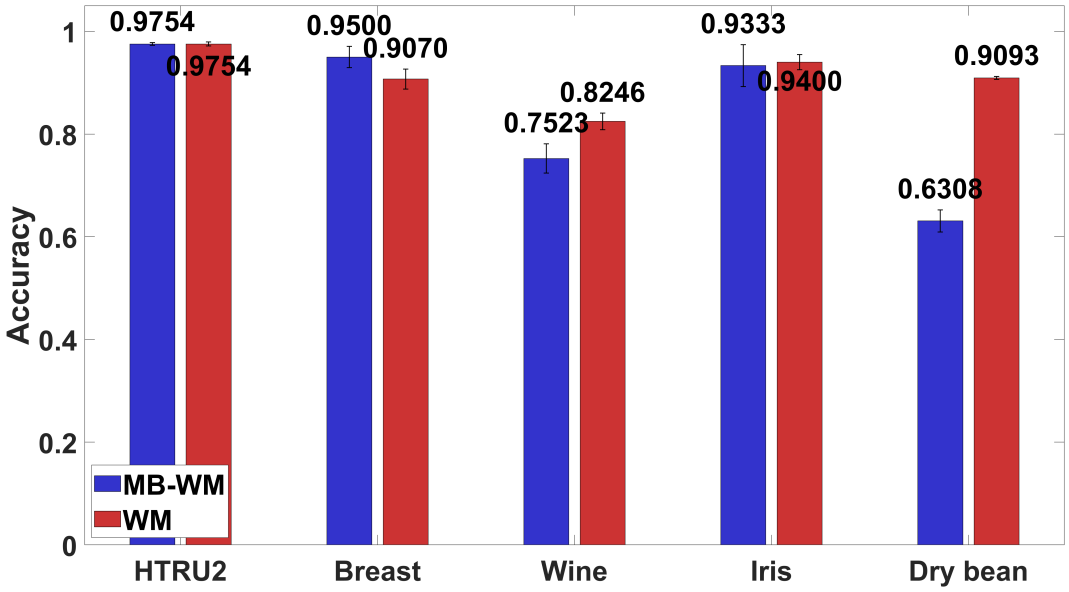


Figure 3.9: *Classification accuracy of MB-WM and WM*

The experiment results are shown in Fig. 3.9. Fig. 3.10 shows the average number of rules of each model during the five-fold cross-validation. It should be clarified that the fuzzy partitions of all datasets are determined before cross-validation, but the rule base gets generated (using the WM algorithm) each time. Thus, the fuzzy sets are the same, but the number of rules may be different for each fold. For example, the numbers of rules in WM rule bases of the Iris dataset during the five-fold cross validation were $\{22, 20, 23, 21, 22\}$, respectively. In Fig. 3.10, it is shown as $[20, 23]$. This is because the training datasets used in each fold are slightly different.

From Fig. 3.10 and Fig. 3.9 the following observations can be made: 1)

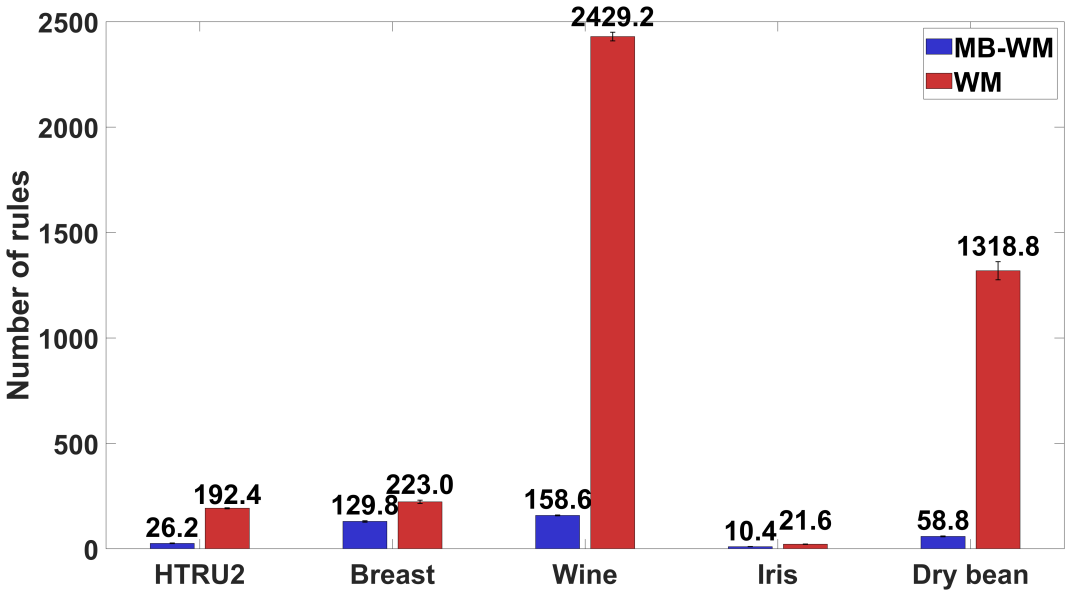


Figure 3.10: *Average number of rules of MB-WM and WM*

According to Fig. 3.9, the systems obtained by MABLAR-ST show comparable classification performance with the system obtained by the WM algorithm in most data set. To further compare the performance of MB-WM and WM, the T-test was used to check if there a significant difference of performance between two models. The P-value is 0.3315, which means the different in performance between both approaches is not statistically significant. 2) As shown in Fig. 3.10, the systems obtained by MABLAR-ST has fewer fuzzy rules than those obtained by the WM algorithm.

3.6 Discussion

The results show that the systems obtained by MABLAR-ST have generally comparable classification performance with the systems obtained by the WM algorithm. Thus, as expected, leveraging the subset of variables which are within the MB rather than the full set does not result in a negative impact of model performance. In addition, in the Breast data set, the performance of MB-WM is even higher than WM. This observation shows that using the MB to replace the original feature sets can help improve the model performance. The case is expected because by relying only on variables which are causally

related to the output, the MB-WM avoid the modeling of spurious correlations, i.e. they avoid spurious rules which in turn will minimize overfitting.

Rule bases of MB-WM have a smaller number of rules than those of WM across all data sets, indicating that the use of the MB can result in the generation of smaller rule bases. MABLAR-ST reduces the number of input variables, which compacts the input feature space. Thus, the fuzzy systems obtained by MABLAR-ST can use fewer rules to cover *relevant* points in the input feature space.

It should be highlighted that in the Wine data set and the Dry bean data set, the performance of MB-WM has a significant decrease. A possible reason for this observation is that the causal graph of these datasets is not the ground truth causal graph. In other words, it seems that the adopted PC algorithm was unable to determine the accurate causal relationships between variables within these data sets using the data provided. As a result, the identified MB of the output variable of these data sets is not the ground truth MB, which decrease the performance of the systems obtained MABLAR-ST. This will be further investigated in the future.

3.7 Summary

This chapter provides an overview of the MABLAR framework and its first variant developed in this thesis, i.e., MABLAR-ST. The MABLAR framework is designed to generate rules which reflect different facets of causal relationships by removing variables which are not causally related to the target variable. More specifically, the MABLAR framework first identifies causal information from a given data set and then uses the identified information to refine the original data set to achieve causal rule generation. The MABLAR framework contains there variants, which are MABLAR-ST, MABLAR-CD, MABLAR-CW, and one meta-variant which is CF-MABLAR. They are designed to leverage different types of causal information to achieve different

goals.

MABLAR-ST is further introduced in this chapter, and the rest three variants are further introduced in the following chapters. MABLAR-ST is designed to generate rules which capture the relationships between the system output variable and the input variables that are (potentially) causally related to it, i.e., the local causal relationships of the output variable. MABLAR-ST uses the information of the dependencies between variables in a causal graph obtained from a given data set to refine the original data set into the MB subset of the system output variable. By only using the MB subset for rule generation, MABLAR-ST ensures that all rule antecedents are based on variables which are causally related to the target variable.

The experiment results show that the rules generated by MABLAR-ST have less variables in their antecedents and the obtained fuzzy system contains fewer rules. The rules generated by MABLAR-ST model specifically the causal relationships between the system inputs and output(s), i.e. the framework avoids *spurious rules*. Through this mechanism, MABLAR-ST can reduce the complexity of fuzzy systems and improve the explainability of the obtained fuzzy system, while maintaining performance.

A challenge of MABLAR-ST is to find the correct MB for the target variable. Finding the correct MB is based on establishing the correct causal relationships between variables–the causal discovery research challenge as discussed in Section 2.2. In this chapter, the PC algorithm is adopted as one of the most popular methods for a variety of problem. Different causal discovery algorithms can be used depending on the specific application.

As rules obtained by this way focus on the local causal structure of the output variable, they may conflate the causes and the effects of the system output variable in their antecedents. However, in some cases, rules are expected to *only* contain the (direct) *causes* of the system output variables. For example, when the problem to be resolved is "What are the risk factors that lead to diabetes?", the rule antecedents are expected to only contain the (direct) causes

of diabetes.

The next chapter broadens this discussion to explain the differences in explanations produced by these two types of rules, as well as the scenarios in which each is applicable. Additionally, the next chapter introduces another variant of the MABLAR framework, which is called Markov blanket rule generation using causal direction (MABLAR-CD). MABLAR-CD is designed to generate rules whose antecedents contain *only* the direct causes of the output variables.

Chapter 4

Towards causal fuzzy system rules using causal direction

As mentioned in the previous chapter, rules obtained by MABLAR-ST may not be suitable for the cases where rules are expected to only represent the relationships between the output variable and the input variables which are the direct causes of it. This chapter discusses the cases where rules are expected to represent only the relationships between the output variable and its input variables which are the direct causes (See Section 2.1.2 for the term 'direct causes'). This chapter explains why the rules obtained by MABLAR-ST may not be suitable in such scenarios.

To address this, MABLAR-Causal Direction (MABLAR-CD) is introduced in this chapter, addressing Sub-RO2 of RO3, i.e., using causal direction to generate rules which reflect 'direct cause-effect relationships'. In addition, a detailed comparison of MABLAR-ST and MABLAR-CD is provided, establishing and demonstrating the scenarios in which each variant of MABLAR is suitable.

The rest of this chapter is organised as follows: Section 4.1 explains the motivation of MABLAR-CD. Section 4.2 shows the training process and the rationale of MABLAR-CD. Section 4.2.3 provides a detailed comparison between MABLAR-ST and MABLAR-CD. Section 4.3 shows the experiment results to show the difference between rules obtained by MABLAR and MABLAR-CD.

Section 4.4 provides the discussion of the experiment results. Finally, Section 4.5 summaries the contributions and the conclusions of this chapter.

4.1 Motivation of MABLAR-CD

As mentioned in Section 1.4, human are cause-effect thinkers. We reason through the causal relationships between things (or events) [1]. In many cases, we not only infer a result from its direct causes, but also infer a result from its effects or both its effects and direct causes. For example, when diagnosing whether a patient has the flu, doctors do not only rely on whether the patient is infected with the influenza virus (the direct cause of the flu) but also consider symptoms such as fever (the effects of the flu) for diagnosis. However, in some cases, only the direct causes should be considered. For example, when a researcher aims to identify the causes of lung cancer, they will only consider the direct causes that may lead to lung cancer, such as smoking or cell carcinogenesis due to genetic mutations (potential direct causes of lung cancer), rather than symptoms (the effects of lung cancer) like coughing that are caused by lung cancer.

The previous chapter introduces MABLAR-ST. MABLAR-ST uses the information about dependencies between variables from a causal graph of a given data set to narrow the original data set to the MB subset of the system output variable for rule generation. It is designed to avoid generating spurious rules by restricting rule learning to the input variables which are within the MB of the system output variable. Rules obtained by MABLAR-ST reflect local causal relationships of the output variable, because, as explained in Section 2.1, variables within the MB of a target variable describes the local causal structure of the target variable.

However, the MB of a target variable does not only contain its direct causes, but also contains its direct effects and the direct causes of its direct effects [43]. When the later two types of variables are present in a given data set

and MABLAR-ST successfully identifies them, rules obtained by MABLAR-ST do not distinguish the direct causes and direct effects of the system output variable in their antecedents. In that case, rules obtained by MABLAR-ST may not be suitable for the scenarios where the focus is primarily on the direct causes of a target variable, such as studies investigating the causes of lung cancer.

As discussed in Section 1.3, the rules of a fuzzy system can represent the relationships between variables captured by the model. In addition, the 'IF-THEN' structure applied by rules of fuzzy systems implies to their users that 'IF *cause*, THEN *effect*' [109]. Thus, if the rules of a fuzzy system only contain the direct cause variables of the system output variable and are understandable by the users, such a system has the potential to provide explanations in the scenarios which require a primary focus on the direct causes of the system output variable. These explanations (potentially) can help users find the direct causes of certain results in the real world.

For example, consider a data set used to predict whether patients will develop lung cancer, which includes many variables, some of which are factors causing lung cancer, but it is unknown which variables are the actual causes. If a fuzzy system can be trained using this data set through a data-driven approach, ensuring that its rules include only the direct cause variables of whether patients will develop lung cancer (i.e., the system output variable), then rules of this system can help users identify the causes of lung cancer among the patients in the dataset. Additionally, targeted experiments can be conducted to validate the causes identified by the system, thereby determining whether these variables are the causes of lung cancer not only in the data set, but also in the real world.

This chapter proposes MABLAR-CD which moves beyond MABLAR-ST in focusing not only on the subset of variables described by the MB, but also further narrows this set of variables to those which are direct causes. This chapter explains why MABLAR-CD can capture *'direct cause-effect relationships'*, and

provides a detailed comparison of MABLAR-ST and MABLAR-CD, showing the scenarios in which each variant of the MABLAR framework is suitable. MABLAR-CD is designed to address Sub-RO2 of RO3 (see Section 1.4 and Fig. 1.2), as detailed in the following sections.

4.2 MABLAR-CD – Markov blanket rule generation using causal direction

This section introduces the details of MABLAR-CD, including the training process and the rationale of MABLAR-CD. The notations and the acronyms used in this chapter are shown in Table 1.2.

4.2.1 The process of MABLAR-CD

MABLAR-CD uses both information about the dependencies between variables and information about causal direction to generate rules which are expected to only contain the direct cause variables of the system output variable in their antecedents. Fig. 4.1 shows the process of MABLAR-CD.

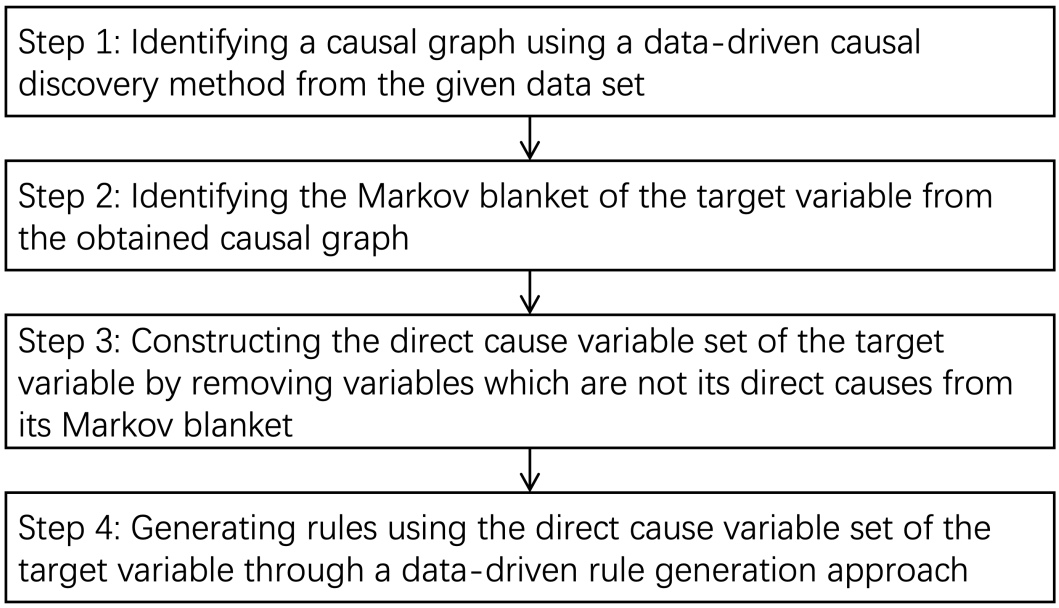


Figure 4.1: *The process of MABLAR-CD*

There are 4 steps in MABLAR-CD:

- Step 1: Generating the causal graph of the given data set using causal discovery algorithms.
- Step 2: Identifying the MB of the output(target) variable from the obtained causal graph and generating $\boldsymbol{D}_{MB}$ by removing non-MB variables from $\boldsymbol{D}$.
- Step 3: Narrowing $\boldsymbol{D}_{MB}$ obtained in step 2 to $\boldsymbol{D}_{CD}$ using the causal direction between variables within the MB of the output(target) variable.
- Step 4: Generating rules using a data-driven rule generation algorithm from $\boldsymbol{D}_{CD}$ obtained in step 3.

4.2.2 Rationale of MABLAR-CD

Fig. 4.2 presents the rationale of MABLAR-CD. In Fig. 4.2, dashed lines represent correlations between variables and solid arrows between variables represent causal relationships. Grey nodes and grey striped nodes represents the MB variables and direct cause variables of the output variable, respectively.

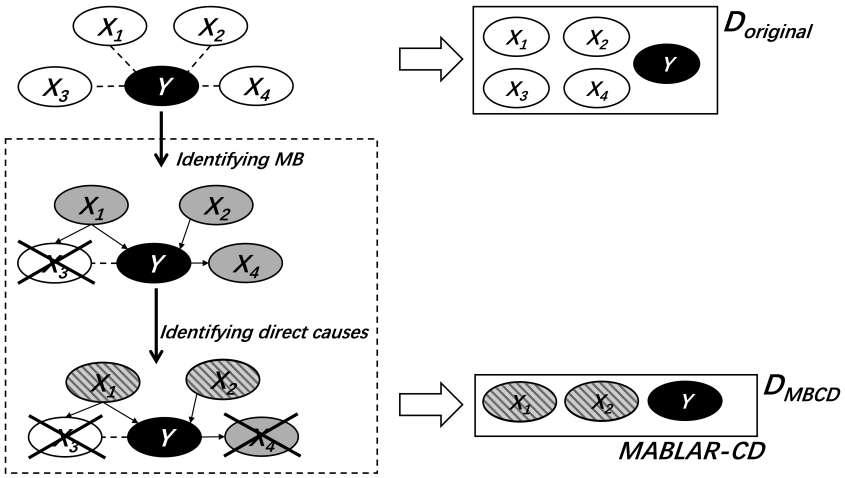


Figure 4.2: *The rationale of MABLAR-CD*

Similar to MABLAR-ST, MABLAR-CD first removes all variables which are correlated but not causally related to the system output variable from a given data set. This is achieved by leveraging information about the dependencies between variables from the causal graph of the given data set.

Then, different from MABLAR-ST which uses all remaining variables to generate rules, MABLAR-CD further identifies the direct cause variables of the system output variable from the remaining variables and removes all non-direct cause variables. This is achieved by leveraging the causal direction information. As discussed in Section 2.1.2, causal direction between two variables indicates cause-effect relationships between them. Nowadays, many data-driven approaches can be used to determine the causal direction between two variables as reviewed in Section 2.2.

Finally, only the direct cause variables are retained for rule generation, ensuring that the antecedents of rules obtained by MABLAR-CD contain only the direct cause variables of the system output variable. This is expected to benefit the obtained fuzzy system from the following 2 aspects:

1. Reduction of the complexity of each rule. In general, rules with lower complexity are more explainable [87, 124]. Rules obtained by MABLAR-CD only contain the variables which are the direct cause variables of the output(target) variable, and the number of such variables is lower (or equal) to the number of variables in the original data set and the MB subset of the system output variable. Thus, the antecedent parts of rules obtained by MABLAR-CD contain less (or equal) variables, which means rules obtained by MABLAR-CD have less complexity and better explainability. Experiment results shown in Section 4.3.3 verify this benefits.

2. Highlighting the 'direct cause-effect' relationship between variables in the antecedent and the consequent of each rule. The antecedents of rules obtained by MABLAR-CD only contain the direct cause variables of the system output variable. Thus, MABLAR-CD helps to generate rules which maintain the 'IF *direct causes*, THEN *effect*' structure. Such rules are more in line with the human way of thinking, which is helpful for users to understand rules and associated fuzzy systems [109]. This

benefit is verified by the experiment results shown in Section 4.3.4.

MABLAR-CD is dependent on eliciting direct-cause variables. However, in real-world applications, the node of the target variable may not have any parent node in a causal graph obtained from the given data set, which will be further shown in Section 4.3.3. Consequently, MABLAR-CD cannot be implemented in those cases, which is a limitation of MABLAR-CD.

4.2.3 Comparison between causal relationships captured by MABLAR-ST and MABLAR-CD

As noted in the previous sections, rules obtained by MABLAR-ST and MABLAR-CD represent different facets of the causal relationship shown in a causal graph, because MABLAR-ST and MABLAR-CD use different subsets of the variables for the rule generation.

MABLAR-ST uses $\boldsymbol{D}_{MB}$ for the rule generation. The variables within $\boldsymbol{D}_{MB}$ provide the local causal structure of the system output variable, because $\boldsymbol{D}_{MB}$ containing both the direct cause variables and the direct effect variables of the output variable.

In contrast, MABLAR-CD is designed to use $\boldsymbol{D}_{CD}$ for rule generation, focusing only on the direct cause variables of the system output variable. Thus, the rules obtained by MABLAR-ST capture local causal relationships between the system output variable and all variables within its MB, while the rules obtained by MABLAR-CD capture the relationships between the system output variable and its direct cause variables.

In real-world applications, whether one selects MABLAR-ST or MABLAR-CD should depend on the problem. The rules obtained by MABLAR-ST map the variables within the input variables which have causal relationships with the output variable. Moreover, the fuzzy systems which consist of rules obtained by MABLAR-ST are likely to exhibit better performance than those consisting of rules obtained by MABLAR-CD for most data sets, because MABLAR-CD

uses less information from the data set.

This reduced utilisation of the data may lead to reduced performance, especially when the overall information content of the data set is crucial for achieving optimal results. In other words, while MABLAR-CD is designed to focus on the direct cause variables only, completing, for example, a classification task within this setting can be much more challenging than when also considering variables representing causally linked direct effects. For example, it is often harder to diagnose the presence of a disease based on the presence of a given virus and the degree of infection of the patient than it is by considering the symptoms the patient is exhibiting. Conversely, establishing what the disease is exactly may not substantially benefit from considering the symptoms.

4.3 Experiments

This section presents the results of experiments on both synthetic and real-world datasets, comparing the explainability and the performance of rules obtained from classical data-driven rule generation approaches with those obtained from MABLAR-ST and MABLAR-CD.

4.3.1 Experiment setting

Two established data-driven rule generation approaches are used as benchmarks, and compared with the rule generation approaches of both MABLAR-ST and MABLAR-CD. The first approach is the WM algorithm, which, as arguably the most traditional data-driven approach, provides a consistent basis for comparison [91, 92]. The other one is FURIA, representing a popular, high-performance technique, which already by designing narrows the original variable sets to improve the explainability of the fuzzy system—albeit without focusing on causal aspects, as is done in this thesis.

Specifically, unlike MABLAR-ST and MABLAR-CD which use causal information, FURIA narrows original data sets by focusing on the correlations be-

tween the variables as discussed in Section 2.5. Thus, the comparison between MABLAR-ST, MABLAR-CD and FURIA highlights some of the differences between rules derived when focusing specifically on causal relationships vs rules derived from correlations.

A traditional fuzzy set design is purposely adopted, using trapezoidal and triangular membership functions for the WM algorithm. The parameters of the membership functions of the WM algorithm are estimated using the same way as described in Section 3.4.2. The ideal number of fuzzy sets of a fuzzy partition for a variable is generally problem dependent. However, this chapter uses the same number of fuzzy sets (i.e., three fuzzy sets) of fuzzy partitions for all variables to facilitate comparisons.

The *python-weka-wrapper3* package is adopted, which is available in [125], to implement FURIA. All parameters are set to their default settings. Trapezoidal membership functions are adopted for FURIA in-line with the algorithm. In FURIA, the parameters of the membership functions and the fuzzy partitions of each variables are automatically determined by FURIA itself.

In terms of causality, the PC algorithm [51] is adopted for all data sets for learning the causal structure, and the ANM algorithm [67] (reviewed in Section 2.2) to infer causal direction which cannot be determined by the PC algorithm. The two adopted algorithms are implemented using the *causal-learn* Python library, which is published in [126] and is available in [127]. In practice, causal discovery algorithms suitable for the intended application should be chosen. However, this research focuses on the above as the universal setup to facilitate experiments and comparisons.

More details about the experiment setting including the parameter settings of the causal discovery algorithms used in this chapter are provided in Appendix A.

4.3.2 Evaluation on synthetic data sets

In this subsection, three experiments are conducted using different synthetic data sets to evaluate the performance, explainability, and robustness of the fuzzy systems obtained by the different approaches. As the variables in the synthetic data sets do not have a physical meaning, here, only the model-based explanations of the fuzzy systems obtained by the different approaches are evaluated. The potential domain-based explanations provided by the fuzzy systems will be evaluated in the next subsection using a real-world data set where each variable has a physical meaning.

The generation of synthetic data sets

Three synthetic data sets were generated to simulate different real-world scenarios, in order to test different rule generation approaches. All synthetic data sets are available at https://github.com/ztcsrookie/Sim_Data_sets_for_MABLAR.git.

The first synthetic data set, marked as $\boldsymbol{D}_{SO}$ ('SO' is short for 'synthetic and original'), is generated to simulate a scenario where the ground truth causal relationships between different variables are known. $\boldsymbol{D}_{SO}$ was generated by the method shown in [128]. The details of this generation are as follows:

$$
\begin{aligned}
X_1 &= \sin(X_4^2) + X_2^2 + \cos(X_7) + E_1, && E_1 \sim U(-0.1, 0.1) \\
X_2 &= X_3 + E_2, && E_2 \sim U(-0.5, 0.5) \\
X_3 &= E_3, && E_3 \sim U(-1.0, 1.0) \\
X_4 &= \sin(X_2) + \sin(2X_3) + E_4, && E_4 \sim U(-0.5, 0.5) \\
X_5 &= \tanh(X_6 + X_7 + X_2) + E_5, && E_5 \sim U(-0.3, 0.2) \\
X_6 &= \sin(X_2) + \cos(2X_4) + E_6, && E_6 \sim U(-0.5, 0.5) \\
X_7 &= \cos(X_6 + X_3) + E_7, && E_7 \sim U(-0.3, 0.3) \\
Y &= \mathbb{I}(X_3 + X_1^2 - 1.5 + \epsilon > 0), && \epsilon \sim N(0, 1)
\end{aligned}
\tag{4.1}
$$

In (4.1), U represents the uniform distribution and N represents normal distribution. For example, $E_3 \sim U(-1.0, 1.0)$ represents E_3 is uniformly distributed in the range $[-1.0, 1.0]$. 5000 data points are generated for $\boldsymbol{D}_{SO}$ using the function relationships shown in (4.1). In other words, $\boldsymbol{D}_{SO}$ contains 5000 samples and the system output variable of $\boldsymbol{D}_{SO}$ is Y.

The second synthetic data set, marked as $\boldsymbol{D}_{SD}$ ('SD' is short for 'synthetic and different'), is generated to simulate a scenario where the testing data set has a different distribution with the training data set. $\boldsymbol{D}_{SD}$ contains 5000 samples which are generated by changing '*sin*' to '*cos*', '*cos*' to '*sin*', and '*tanh*' to '*tan*' in the functions shown in (4.1).

The third synthetic data set, marked as $\boldsymbol{D}_{SS}$ ('SS' is short for 'synthetic and same'), is generated to simulate the scenario where the testing data set has the same distribution of the training data set. $\boldsymbol{D}_{SS}$ contains 5000 samples which are generated using the same way as $\boldsymbol{D}_{SO}$, but with a different Python random seed.

The function relationships shown in Equation (4.1) are adopted as the ground truth causal relationships, and these relationships are represented by the causal graph shown in Fig. 4.3.

Fig. 4.4 shows the causal graph obtained by causal discovery algorithms using $\boldsymbol{D}_{SO}$. Table 4.1 shows the $\boldsymbol{D}_{MB}$ and $\boldsymbol{D}_{MBCD}$ obtained by the causal discovery algorithms adopted in this chapter and the ground truth, respectively.

Table 4.1: $\boldsymbol{D}_{MB}$ and $\boldsymbol{D}_{MBCD}$ of the system output variable on the synthetic data set

	PC-ANM	Ground truth
$\boldsymbol{D}_{MB}$	{$X1$,$X2$,$X3$,$X4$,$X6$,$X7$}	{$X1$, $X3$}
$\boldsymbol{D}_{MBCD}$	{$X1$, $X3$ }	{$X1$, $X3$}

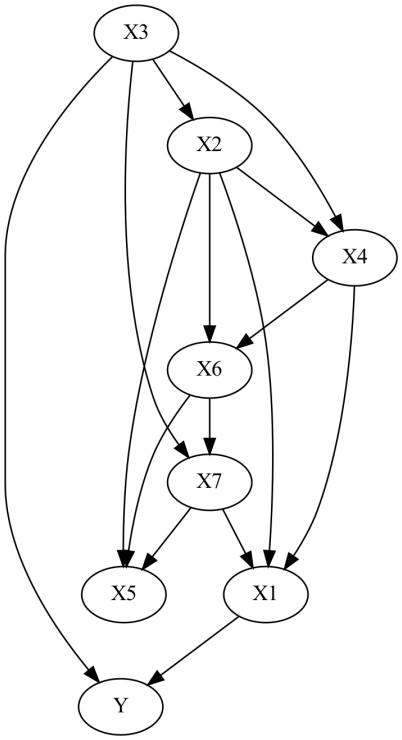


Figure 4.3: *The ground truth causal graph of the simulated data set*

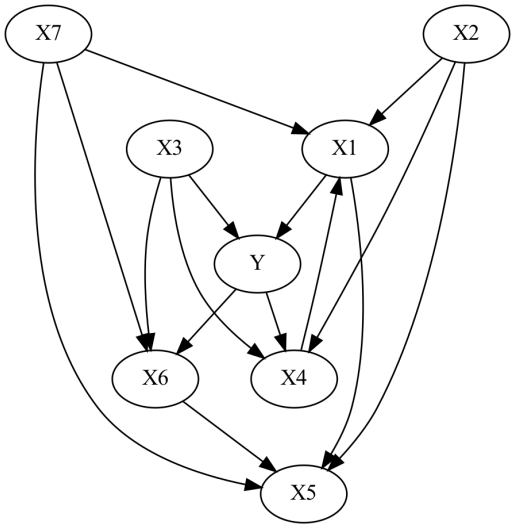


Figure 4.4: *The causal graph of the simulated data set obtained by adopted causal discovery algorithms*

Experiment 1: Evaluation on performance and model-based explanations

The performance and the model-based explanations of the fuzzy systems obtained by these different approaches are first evaluated. This experiment is conducted on $\boldsymbol{D}_{SO}$. The average of the classification accuracy over 5-fold cross-validation is adopted as the performance index. The index is better when higher.

Stratified sampling is adopted to keep the original class proportions in the training data set of each fold the same.

The number of rules is adopted as the evaluation index of the model-based explanations, because in the setting of this synthetic experiment, the only factor that differs between the fuzzy systems obtained by the different approaches and has an impact on the model-based explanations of the system is the number of rules.

Table 4.2 shows the results of each fuzzy system during the 5-fold cross-validation. '# Rules' in Table 4.2 represents 'number of rules'. An algorithm with a prefix of 'MB' (resp., 'MBCD') indicates that the fuzzy system is obtained using MABLAR-ST (resp., MABLAR-CD).

Table 4.2: Performance and # rules of different fuzzy systems

	WM	MB-WM	MBCD-WM	FURIA	MB-FURIA	MBCD-FURIA
Accuracy(Std)	0.7972(0.0180)	0.7998(0.0187)	**0.8140(0.0085)**	0.7998(0.0011)	0.8015(0.0062)	0.8057(0.0063)
# Rules	323.8(3.3)	188.8(1.1)	**9(0)**	7.6(1.4)	5.4(2.0)	**4.6(1.1)**

When interpreting the results, it should be cautioned that these are illustrative only, as they are based on one synthetic model and its associated data set. From Table 4.2, the following can be observed: MB-WM, MBCD-WM have a higher performance and smaller number of rules than WM. Similarly, MB-FURIA and MBCD-FURIA have a higher performance and smaller number of rules than FURIA. This observation indicates that MABLAR-ST and MABLAR-CD can improve the model-based explanations of the obtained fuzzy

systems while improving the performance.

Experiment 2: Evaluation on robustness in distribution shift scenarios

Experiment 2 is conducted to evaluate the robustness of the fuzzy systems obtained by the different approaches when the distributions of the training and testing data are different.

In Experiment 2, $\boldsymbol{D}_{SO}$ is used as the training set. For each rule generation approach, a fuzzy system is trained using the training set. The obtained fuzzy system are tested on both $\boldsymbol{D}_{SD}$ and $\boldsymbol{D}_{SS}$. As introduced before, $\boldsymbol{D}_{SD}$ and $\boldsymbol{D}_{SS}$ are generated using methods different from and the same as $\boldsymbol{D}_{SO}$, respectively. Therefore, testing on $\boldsymbol{D}_{SD}$ simulates a scenario where the distributions of the training set and test set are different, while testing on $\boldsymbol{D}_{SS}$ simulates a scenario where the distributions of the training set and test set are the same.

In theory, if a model still has good performance (with classification accuracy used as the performance metric in this section) when faced with different distributions, it is considered to have good robustness in distribution shift scenarios[8]. Thus, in Experiment 2, if the performance of a system on $\boldsymbol{D}_{SD}$ is closer to its performance on $\boldsymbol{D}_{SS}$, the system is considered to have better robustness in distribution shift scenarios. Ideally, systems based on causal relationships should have better robustness than those based on correlations [8, 108].

The results of Experiment 2 are shown in Fig. 4.5. In Fig. 4.5, the blue bars represent the testing results obtained on $\boldsymbol{D}_{SS}$, i.e., the testing set generated in the same way as $\boldsymbol{D}_{SO}$ (the training set), while the red bars represent the testing results obtained on $\boldsymbol{D}_{SD}$, i.e., the testing set has different distribution with $\boldsymbol{D}_{SO}$.

From Fig. 4.5, the following observations can be made : 1) The performance of the fuzzy systems obtained by WM and FURIA is significantly worse when the distribution of the testing set is different from that of the training set. In

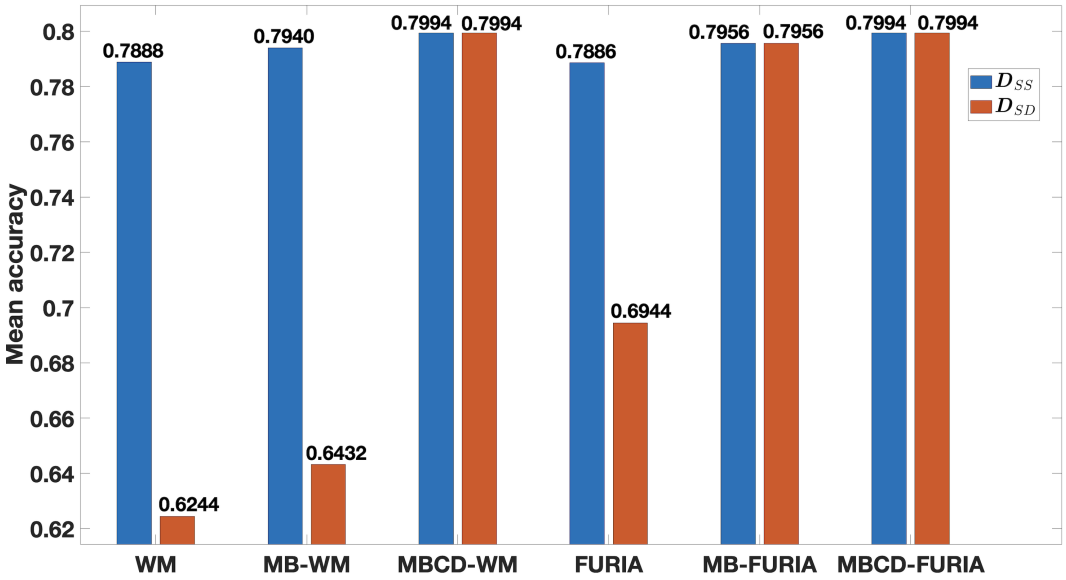


Figure 4.5: *Results of Experiment 2*

contrast, the performance of the fuzzy systems obtained by MBCD-WM and MBCD-FURIA does not worsen. 2) The performance of the fuzzy systems obtained by MB-WM also worsens significantly when the distribution of the testing set is different from that of the training set. However, compared to the results of WM, MB-WM shows less performance degradation.

The above observations indicate that fuzzy systems whose rules only capture the causal relationships between the variables are more robust than those which also capture correlations between the variables when facing a distribution shift between the training set and the testing set. Another way of interpreting this is that the focus on the causal relationships makes the systems robust to overfitting.

Experiment 3: Evaluating the importance of the causal variables for the performance of the model

Experiment 3 is designed to evaluate the importance of the variables which have causal relationships with the system output variable.

In Experiment 3, for each rule generation approach, a fuzzy system is trained using $\boldsymbol{D}_{SO}$. Then, each obtained fuzzy system is tested on $\boldsymbol{D}_{SS}$. More specifically, the variables $X1$–$X7$ are replaced one by one in $\boldsymbol{D}_{SS}$ in turn with random

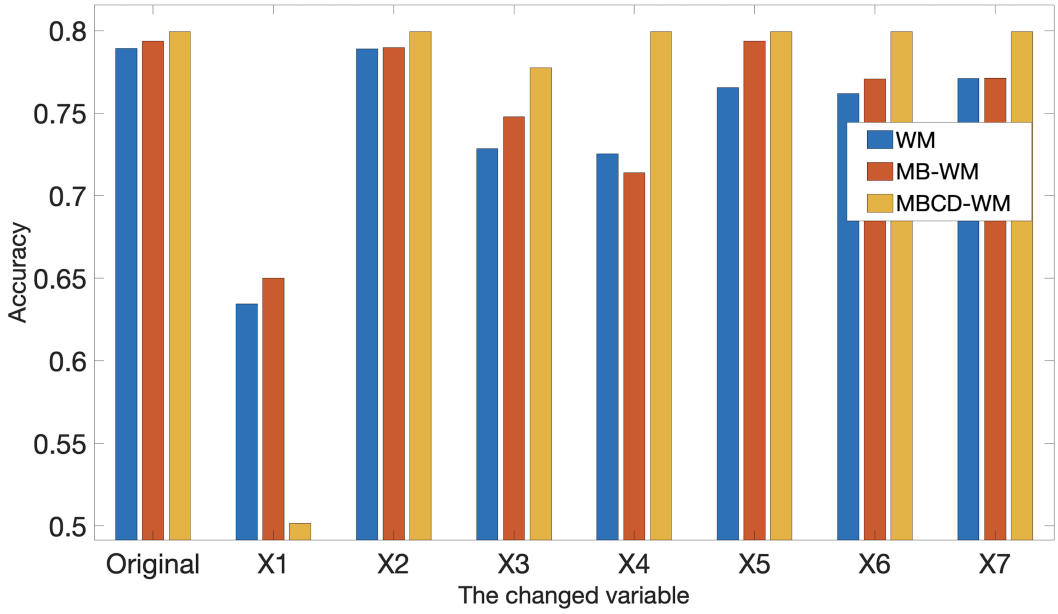
noise (generated from a standard normal distribution, i.e. $\mu(0, 1)$) and tested the fuzzy system on each of these modified testing sets. If the performance of the system drops more significantly after a variable is replaced with random noise, that variable is considered to be more important.

In Fig. 4.6, $X1$–$X7$ on the x-axis represent the variable which is replaced by the random noise. The 'original' on the x-axis represents the performance of the model on the unmodified $\boldsymbol{D}_{SS}$ (i.e. no variable is replaced by random noise).

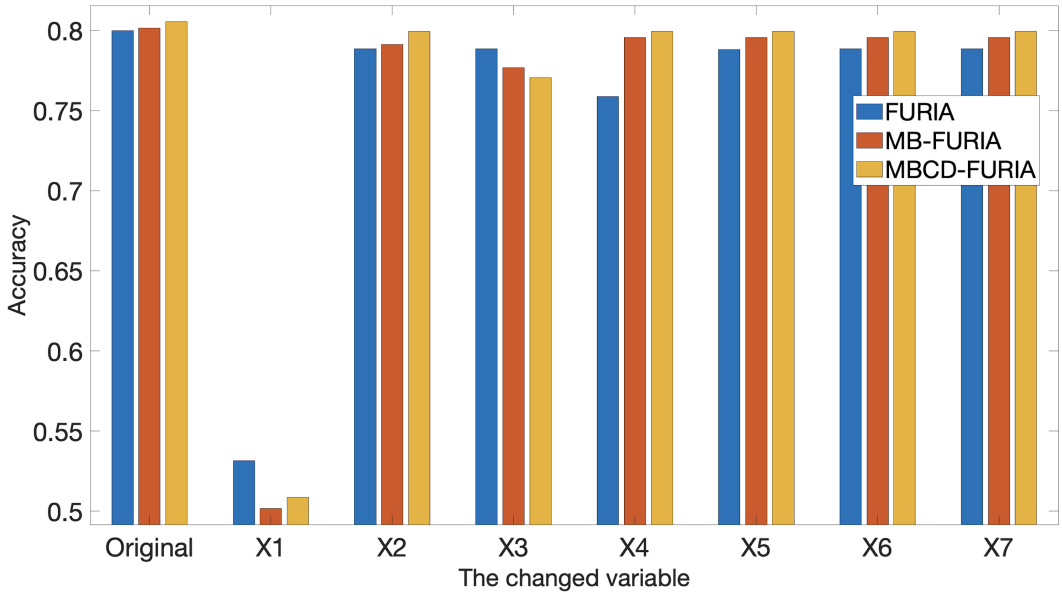
From Fig. 4.6 the following can be observed: 1) The performance of all obtained fuzzy systems decreases the most when '$X1$' is replaced, and the performance of most obtained fuzzy systems decreases the most when $X3$ is replaced. 2) The performance of MB-WM and MB-FURIA show the most significant performance decrease when $X1$ and $X4$ are replaced, while those of MBCD-WM and MBCD-FURIA decrease the most when $X1$ and $X3$ are replaced with noise. This indicates that, from the aspect of improving the model's performance, the model should focus more on variables that have causal relationships with the target variable rather than variables which are only correlated with it.

4.3.3 Evaluation on real-world data sets

The evaluation of the fuzzy systems obtained by the different approaches is conducted on 17 real-world data sets from the UCI data repository [122] in terms of both performance and explainability. These 17 datasets are selected because they are widely used as benchmarks for rule generation. Also, these datasets has varying sample sizes (both small and large) as well as varying number of variables (few vs. many variables). By conducting experiments on additional data sets beyond those used in Section 3.5, this section provides a more comprehensive comparison between MABLAR-ST and MABLAR-CD. Table 4.3 shows the number of samples, and the number of classes of each data



(a) *The WM algorithm*



(b) *The FURIA algorithm*

Figure 4.6: *Experiment results with Experiment Setting 4*

set, the number of variables in different subsets and the percentage of samples in each class.

Table 4.3: Data sets used in this chapter[1]

Name	Samples	Class	$\|D\|$	$\|D_{MB}\|$	$\|D_{CD}\|$	Percentage of each class
Authorship	841	4	70	36	4	38%, 35%, 7%, 21%
Breast	699	2	9	6	3	66%, 34%
Dry bean	13611	7	16	2	0	10%, 4%, 12%, 26%, 14%, 15%, 19%
Ecoli	336	8	7	2	0	43%, 23%, 1%, 1%, 10%, 6%, 1%, 15%
Glass	214	6	9	7	1	33%, 36%, 8%, 6%, 4%, 14%
Haberman	306	2	3	1	1	74%, 26%
HTRU2	17989	2	8	7	0	91%, 9%
Iris	150	3	4	3	0	33%, 33%, 33%
Mammographic mas	830	2	6	4	2	51%, 49%
Page-blocks	5473	5	10	6	0	90%, 6%, 1%, 2%, 2%
Pendigits	10992	10	16	10	0	10%, 10%, 10%, 10%, 10%, 10%, 10%, 10%, 10%, 10%
Pima Indian Diabetes	768	2	8	6	3	65%, 35%
Sonar	208	2	60	1	0	53%, 47%
Vehicle	846	4	18	10	0	26%, 25%, 26%, 24%
Vowel	990	11	12	7	2	55%, 45%
Waveform-5000	5000	3	40	11	0	34%, 33%, 33%
Wine	178	3	13	8	2	33%, 40%, 27%

[1] The sum of percentage of each class for some data sets is not 100% because of rounding.

The performance and explainability of the model-based explanations of the fuzzy systems obtained by the different approaches are first evaluated.

Performance on real-world data sets

Considering some data sets have extreme unbalance classes (e.g., the HTRU2 data set), the average of F1 scores over 5-fold cross-validation is adopted as the performance indices. The F1 score is calculated as follows [129]:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{4.2}$$

where

$$Precision = \frac{TP}{TP + FP}, \tag{4.3}$$

$$Recall = \frac{TP}{TP + FN}, \tag{4.4}$$

and

- TP represents the number of true positive cases
- TN denotes the number true negative cases;
- FP represents the number false positive cases;

- FN is the number of false negative cases.

For multi-class data sets, the Macro F1 scores is adopted as it has been widely used as an extension of F1 scores in multi-classes problem. For the data set with N_c classes ($N_c > 2$), the Macro F1 scores is calculated as follows[130]:

$$Macro - F1 = \frac{\sum_{i=1}^{N_c} F1_i}{N_c}, \tag{4.5}$$

where $F1_i$ is the F1 score of the i the class. Both the accuracy and the (Macro) F1 scores are better when higher. In addition, the performance of the random forest model [131] is also provided as a baseline of classical black box classifiers for reference.

The performance indices are shown in Table 4.4, with the best indices in bold [1]. MABLAR-CD generates no result for several data sets in Table 4.4 due to the limitation discussed in in Section 4.2.2. The data sets which have no result obtained by MABLAR-CD are specifically included here to highlight this aspect of the variants of MABLAR such as MABLAR-CD which are dependent on eliciting direct-cause variables.

[1] In these two tables, only the best index among all fuzzy approaches are bold to highlight the comparison between fuzzy systems obtained by different approaches

Table 4.4: The F1 scores of different fuzzy systems on each data set

	WM		MB-WM		MBCD-WM		FURIA		MB-FURIA		MBCD-FURIA		Random Forest	
	F1-mean	Std	F1-mean	Std	F1-mean	Std	F1-mean	Std	F1-mean	Std	F1-mean	Std	F1-mean	Std
Authorship	0.0853	0.0011	0.1022	0.0190	0.7615	0.0297	0.8392	0.0356	**0.8624**	0.0242	0.7401	0.0442	0.9855	0.0001
Breast	0.9545	0.0126	0.9558	0.0205	0.9279	0.0299	**0.9575**	0.0097	0.9518	0.0047	0.9495	0.0098	0.9573	0.0007
Dry-bean	**0.9254**	0.0031	0.2532	0.0258			0.9197	0.0018	0.4940	0.0116			0.6105	0.0497
Ecoli	**0.4841**	0.0392	0.1945	0.0230			0.4521	0.0319	0.2642	0.0182			0.7196	0.0110
Glass	**0.5868**	0.1167	0.5302	0.0830	0.1522	0.0636	0.4395	0.0639	0.4361	0.0609	0.2346	0.0082	0.6665	0.0018
Haberman	0.5527	0.0261	0.4552	0.0565			0.8215	0.0295	**0.8230**	0.0298			0.5626	0.0065
HTRU2	0.9221	0.0087	0.9110	0.0172			0.9874	0.0008	**0.9879**	0.0003			0.9322	0.0001
Iris	0.9254	0.0388	0.9389	0.0445			0.9257	0.0274	**0.9415**	0.0132			0.9531	0.0015
mammographic	0.7541	0.0808	0.7907	0.1109	0.7598	0.0864	**0.8439**	0.0145	0.8416	0.0157	0.7779	0.0192	0.7869	0.0016
Page_blocks	**0.6502**	0.0091	0.5276	0.0516			0.1654	0.0208	0.1546	0.0347			0.7728	0.0066
Penditgits	**0.9852**	0.0028	0.9583	0.0074			0.9493	0.0034	0.9169	0.0023			0.9906	0.0000
Pima_diabetes	0.6750	0.0289	0.6610	0.0205	0.6917	0.0442	0.8064	0.0168	0.8057	0.0120	**0.8091**	0.0132	0.7236	0.0027
Sonar	0.4369	0.0280	0.3419	0.0529			**0.6792**	0.0231	0.5269	0.0932			0.7079	0.0154
Vehicle	0.6614	0.0328	**0.6635**	0.0455			0.5908	0.0500	0.5777	0.0571			0.7469	0.0002
Vowel	0.1336	0.0260	0.1233	0.0041			0.5402	0.0387	**0.5410**	0.0278			0.6148	0.0072
Waveform_5000	0.1670	0.0063	0.7533	0.0195			**0.8027**	0.0061	0.7650	0.0046			0.8474	0.0002
Wine	0.7624	0.0608	**0.9008**	0.0243	0.8248	0.0773	0.2168	0.0078	0.2425	0.0118	0.1992	0.0266	0.9673	0.0009

From Table 4.4, the following observations can be made: 1) Among all fuzzy approaches, the fuzzy systems obtained by WM and FURIA achieve the highest F1 scores on 10 data sets, while the fuzzy systems obtained by MABLAR-ST or MABLAR-CD achieve the highest accuracy on 7 data sets. 2) The random forest model achieves the highest F1 scores in most data sets (11 data sets), however, MABLAR-based fuzzy systems achieve comparable performance in most data sets.

The above observations indicate that MABLAR-ST and MABLAR-CD cannot always improve the performance for a given data set, i.e. based on the data available and parameters such as the causal discovery technique adopted, they may not be able to identify and/or leverage the set of causal variables to achieve performance levels which are higher than traditional rule generation approaches. However, the performance of the systems obtained by MABLAR-ST and MABLAR-CD is still comparable with those obtained by traditional rule generation approaches in most data sets.

In addition, as expected, the selected black-box model, i.e., the random forest, shows better performance than the causal rule-based systems which are obtained under the specific experiment setting (for example, the number of fuzzy sets of each fuzzy partition) in this section. However, it is worth noting that causal rule-based systems have the potential to deliver higher performance by increasing its model complexity, such as by increasing the number of fuzzy sets of each fuzzy partition. However, this often negatively impacts the explainability of the rule-based systems. In other words, the measurement of performance is useful in context, but not in absolute terms.

We further apply the Friedman test [132, 133] to estimate whether there are statistically significant differences in performance among the models. The Friedman test is a non-parametric method and has been widely used in the machine learning area to estimate the significance of differences between multiple classification models evaluated over multiple data sets [134, 135]. The null hypothesis in the Friedman test states that there is no statistically significant

difference in classification performance among all models [135].

Since MBCD-WM and MBCD-FURIA do not have results on data sets such as Dry-bean and Ecoli, we conduct the Friedman test using the results of all models on the Authorship, Breast, Glass, mammographic, Pima_diabetes, and Wine data sets. The p-value is 0.0862, which is greater than 0.05. Therefore, the null hypothesis is not rejected. In other words, there is no statistically significant difference in the performance of all models on these data sets. However, as discussed above, the causal rule-based systems still have the potential to achieve good performance by increasing their complexity.

This chapter also compares the execution time of different approaches as a supplement of the performance comparison. However, since MABLAR-ST and MABLAR-CD are not primarily designed to improve execution efficiency, the results of this experiment are provided in the Appendix B.1.

Model-based explainability on real-world data sets

The number of rules is adopted as the evaluation index of the model-based explanations on real-world data sets, because similar as the Experiment 1 on the synthetic data set, the only factor that differs between the fuzzy systems obtained by the different approaches and has an impact on the model-based explanations of the system is the number of rules. The result is shown in Table 4.5.

From Table 4.5, the following observations can be made: For most data sets, the number of rules of a fuzzy system obtained by MABLAR-ST or MABLAR-CD is lower than that of one obtained by the original approach. This observation indicates that MABLAR-ST and MABLAR-CD can improve the explainability of model-based explanations by reducing the complexity of obtained fuzzy systems.

Table 4.5: The number of rules of different fuzzy systems for each data set

	WM		MB-WM		MBCD-WM		FURIA		MB-FURIA		MBCD-FURIA	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Authorship	672.8	0.4	672.8	0.4	70.4	1.3	8.0	1.1	9.8	1.2	**7.4**	1.2
Breast	230.4	4.2	130.4	5.2	24.8	1.1	10.2	1.2	10.6	1.9	**9.4**	2.2
Dry_bean	1338.8	24.6	**9.0**	0.0			33.4	1.5	12.0	1.7		
Ecoli	81.2	5.2	9.0	0.0			6.8	1.0	**4.6**	0.5		
Glass	87	6.8	61.2	4.5	3.0	0.0	5.0	1.1	4.0	1.3	**1.4**	0.5
Haberman	25.4	0.5	3.0	0.0			3.2	2.1	**1.4**	0.5		
HTRU2	159.4	3.4	104.8	2.2			**12.6**	2.8	14.2	2.6		
Iris	19.6	1.8	10.2	0.4			3.4	0.5	**3.0**	0.0		
Mammographic	77.4	5.6	53	5.8	9	0	**4.2**	1.3	4.4	1.4	4.6	1.6
Page_blocks	187.4	5.5	72.6	7.1			17.0	2.3	**16.0**	1.9		
Penditgits	3960.8	37.7	2153.0	9.6			**204.0**	6.2	212.0	13.6		
Pima_diabetes	396.6	8.1	330.8	6.0	9.0	0.0	9.0	2.1	10.8	1.5	**7.4**	1.0
Sonar	166.2	0.4	3.0	0.0			4.2	1.0	**2.6**	0.8		
Vehicle	537.6	5.4	418.4	8.0			**10.2**	3.1	11.8	1.9		
Vowel	539.6	7.0	240.8	35.1			42.6	4.5	**40.4**	4.5		
Waveform_5000	4000	0.0	2920.2	14.9			158.6	5.2	**73.6**	13.2		
Wine	140.4	1.1	113.8	1.6	8.0	0.0	8.2	0.7	7.8	1.2	**6.6**	2.1

4.3.4 Evaluation of domain-based explanations

This section uses the Pima Indian Diabetes (PID) data set as an example to show the difference between the domain-based explanations provided rules obtained by different approaches, highlighting how the rules obtained by MABLAR-ST capture local causal relationships, and how the rules obtained by MABLAR-CD capture causal relationships.

Fig. 4.7 shows the causal graph of the PID data set obtained using the experiment settings provided in Section 4.3.1 and Appendix A. In Fig. 4.7, the outcome variable (marked as black) is the target variable and it represents whether or not a Pima Indian woman has type-2 diabetes [114]. The MB of the outcome consists of all grey nodes, and the striped grey nodes are the direct cause variables of the outcome.

It should be noted that Fig.4.7 is different from the causal graph shown in Section 3.4. Both causal graphs are obtained by the PC algorithm using the 'causal-learn' Python library. However, Fig.3.6 is obtained using the default parameter setting, while Fig.4.7 is obtained using the parameter setting shown in Appendix A. Changing the parameter setting in this section is intended to simulate the scenarios where relevant background knowledge can assist in causal discovery. In this section, the background knowledge regarding the

causal relationships between variables within the PID dataset comes from [136] and [137].

Table 4.6 shows the variables used in each fuzzy system based on Fig. 4.7. For this illustrative case, the number of fuzzy sets of the fuzzy partition for the BMI and age variables is adjusted from 3 to 4 to reflect the medical and World Health Organisation guidelines on key categories [138, 139]. The linguistic terms for the BMI fuzzy partitions are underweight, healthy, overweight, and obese, respectively. And the linguistic terms for the age fuzzy partitions are young, middle, young old and old. The complete rule base of each fuzzy system is shown in Appendix C and Appendix D.

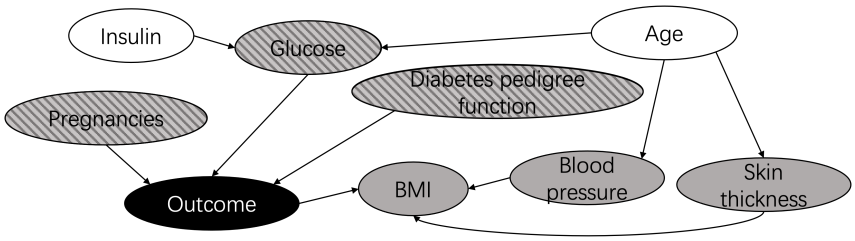


Figure 4.7: *The causal graph of the PID dataset*

Table 4.6: Variables used in different fuzzy systems

	Pregnancies	Glucose	***Blood Pressure***	Skin Thickness	Insulin	***BMI***	***Diabetes Pedigree Function***	***Age***
WM	✓	✓	✓	✓	✓	✓	✓	✓
MB-WM	✓	✓	✓	✓		✓	✓	
MBCD-WM	✓	✓					✓	
FURIA		✓	✓			✓	✓	✓
MB-FURIA		✓				✓	✓	
MBCD-FURIA	✓	✓					✓	

Local causal relationships captured by MABLAR-ST

According to [140], the diagnosis of a disease depends on the symptoms that are causally related to the disease. As discusses in Section 3.3.2, rules obtained by MABLAR-ST only contain variables which are causally related to the target variable. Thus, the rules generated by MABLAR-ST from the PID dataset provide users with a guideline on how to diagnose type-2 diabetes.

The resulting causal links are highly influenced by the causal graph obtained by the adopted causal discovery algorithm. As shown above, the causal graph shown in Fig. 4.7 and the causal graph shown in Fig. 3.6 are obtained by

the PC algorithm using different parameter settings. As a result, compared to rules obtained by MB-WM based on Fig. 3.6 (See Table 3.1), rules obtained by MB-WM based on Fig. 4.7 (See Table 4.6) suggests that the skin thickness variable is also causally linked to the target variable. Thus, in practice, the resulting causal links can be viewed as suggestions which may guide experts – who may proceed to experimentally verify them.

Causal relationships captured by MABLAR-CD

The rules obtained by MABLAR-CD provide explanations for the etiology of type-2 diabetes, i.e. the factors that cause type-2 diabetes. As shown in Table 4.6, both MBCD-WM and MBCD-FURIA indicate that type-2 diabetes is caused by the variables pregnancies, glucose and Diabetes Pedigreee Function (DPF).

In the real-world, the etiology of type 2 diabetes remains an unresolved issue. However, there are many risk factors being considered as the potential causes of type 2 diabetes. According to [116–118] (and some public sources, e.g. [119]), the recognized risk factors of type-2 diabetes, which are also included in the actual Pima Indian diabetes data set, are blood pressure, BMI, DPF and age. Among these four variables, DPF is the only variable considered both as a risk factor and included in the rules generated by MABLAR-CD, which suggests that users can be more concerned with the DPF variable when aiming to prevent type-2 diabetes.

In addition, the rules in MBCD-WM and MBCD-FURIA both include the glucose variable; however, it is not recognized as a risk factor. In Fig. 4.7, the causal relationship between 'BMI', 'Glucose' and 'Outcome' is 'Glucose $\rightarrow$ Outcome $\rightarrow$ BMI'. However, according to [117], the correct causal relationship should be 'BMI $\rightarrow$ Outcome $\rightarrow$ Glucose'. A possible reason for this is that these two causal structures belong to the same Markov equivalence classes, which cannot be correctly determined by the adopted PC algorithm (See the discussion of the PC algorithm in Section 2.2).

If the algorithm could correctly identify this causal relationship from the given data set [2], the glucose variable in the rules obtained by MABLAR-CD would be replaced by the BMI variable. Consequently, the rules obtained by MABLAR-CD would provide better explanations for the etiology of type-2 diabetes, more consistent with the domain knowledge – and with the potential to guide medical experts in less researched contexts.

Overall, while not perfect, MABLAR-CD provide important improvements in generating rules with the capacity to provide direct cause explanations.

4.4 Discussion of the experiment results

MABLAR-ST and MABLAR-CD are designed to generate rules which capture causal relationships between the input and output variables, which benefits fuzzy systems in terms of performance, robustness and explainability. Table 4.2 and Table 4.5 support the conclusion that both MABLAR-ST and MABLAR-CD can reduce the complexity of the obtained fuzzy systems and improve the explainability of the obtained fuzzy systems in terms of model-based explanations. However, Table 4.2 and Table 4.4 show MABLAR-ST and MABLAR-CD may not always improve the performance of the obtained fuzzy systems.

As explained in Section 4.2.3, the performance of fuzzy systems which are obtained by MABLAR-ST should be the highest among all adopted approaches if the appropriate causal relationships are represented in the data set *and* are identified. In practice, as experiments on both the simulated data set and the real world data sets show, this is not necessarily the case, not least because the data sets may not fulfil the latter condition, and even when they do, the adopted causal discovery algorithm may not identify the appropriate relationships correctly.

One possible reason for this is that the causal discovery algorithms used is

[2] It is noted that a data set itself may not allow this to be done, i.e. it may contain insufficient information.

not suitable for the data set, which results in the obtained Markov blanket not being perfect, i.e. the obtained Markov blanket is not the ground truth Markov blanket. When the obtained Markov blankets are imperfect, the rules obtained by MABLAR-ST or MABLAR-CD run the risk of capturing non-causal relationships, which worsens the performance of the obtained fuzzy system and risks generating misleading rules/explanations.

As shown in Table 4.2, MBCD-WM and MBCD-FURIA perform better than WM and MB-WM. Table 4.1 indicates that, on the simulated data set, the MB adopted by MABLAR-CD is equal to the ground truth MB. Thus, given the ground truth MB, it is clear that MABLAR-ST and MABLAR-CD can be helpful for improving the performance of the obtained fuzzy system.

Also, the results shown in Table 4.2 and Table 4.4 show that even when MABLAR-ST and MABLAR-CD adopt an imperfect MB, they are still helpful for improving the performance of the obtained fuzzy system in several cases. The discussion about how to select causal discovery algorithms is outside the scope of this research. See [108, 110] for recent overviews, including on the selection of causal discovery algorithms.

The results in Section 4.3.4 indicate that, although not perfect, the rules obtained by MABLAR-ST and MABLAR-CD provide better domain-based explanations than do rules obtained by a classical data-driven rule generation algorithm. Also, even while not perfect, the rules obtained by MABLAR-ST and MABLAR-CD are still helpful for increasing users' understanding of the real-world processes.

The explanations provided by the rules obtained by MABLAR-CD is further discussed. In many real-world applications, the real cause(s) of a phenomenon are not known. In these cases, even where the rules obtained by MABLAR-CD capture imperfect direct cause-effect relationships between the variables, they are still helpful for increasing the users' understanding of the real world. To confirm the direct cause of a real-world phenomenon, experts (e.g. in biology or medicine) traditionally engage in large-scale analysis and/or experiments.

The rules obtained by MABLAR-CD can reduce the search scope for the direct cause(s).

As shown in Section 4.3.4, doctors have narrowed the search scope for the causes of type-2 diabetes, which is also included in the PID data set, to the blood pressure variable, the BMI variable, the DPF variable and the age variable. The rules obtained by MABLAR-CD further narrow the search scope to the DPF variable, which highlights a mechanism enabling doctors to gather insight into the etiology of diseases such as type-2 diabetes more directly based on what is learnt by the fuzzy system obtained by MABLAR-CD.

In general, the experiment results support the conclusion that the rules obtained by MABLAR-ST and MABLAR-CD benefit fuzzy systems in terms of both performance and explainability. However, as discussed in Section 4.2.2, the implementation of MABLAR-CD depends on the causal graph obtained from the given data set. In real-world applications, for the same data set, adopting a different causal discovery approaches may result in a different causal graph.

In addition, the given data set may not contain direct causal variables for the target variable, which could also result in the target variable node having no parent nodes in the obtained causal graph. Thus, the implementation of MABLAR-CD is highly influenced by the causal discovery algorithms adopted and the quality of the data set.

4.5 Summary

This chapter discusses MABLAR-CD which is another variant of MABLAR, and further compares the rules obtained by MABLAR-ST and MABLAR-CD. The experiment results shown in this chapter indicate that MABLAR-CD benefits fuzzy systems in A) providing domain-based explanations which highlight the 'direct cause-effect' relationship between variables in the antecedent and the consequent of each rule and B) improving the model-based explainability

due to the reduction of both the number of rules and the complexity of the rules vis-a-vis classical data-driven rule generation methods. In addition, the experiment results on the synthetic data sets show that the systems based on the causal relationships between variable deliver strong robustness when facing distribution shift.

This chapter has also shown how causal relationships capture by MABLAR-CD different from those captured by MABLAR-ST, and has highlighted their respective suitability for different problems. Similar as MABLAR-ST, the quality of the causal graph obtained from the given data set also affects the quality of the rules generated by MABLAR-CD. Rules generated by MABLAR-CD based on a perfect causal graph, i.e. a ground truth causal graph, can perfectly reflect the 'direct cause-effect' relationships between variables. However, the experiment results indicate that even imperfect causal graphs, which may be obtained by data-driven causal discovery algorithms, contribute to some extent in generating rules that reflect the 'direct cause-effect' relationships between the variables.

So far, two different variants of the MABLAR framework which aims to capture different types of causal relationships have been introduced. Both MABLAR-ST and MABLAR-CD are designed to focus only one causal mechanism of the system output variable. However, in the real-world, an effect is generally related to multiple causal mechanisms. For example, the causes of a person's fever could be the flu or COVID-19. The next chapter introduces MABLAR-Causal Weight (MABLAR-CW) which is designed to focus on local explanations of fuzzy systems by considering potential causal mechanisms.

Chapter 5

Towards locally relevant explanations using causal rule discovery

The previous two chapters introduce MABLAR-ST and MABLAR-CD, respectively. Rules obtained by MABLAR-ST and MABLAR-CD focus on modelling a single causal mechanism (See Section 2.1.2 for the term 'causal mechanism') of the target variable. For example, MABLAR-ST focuses on modelling the causal mechanism between the target variable and the variables within its MB. As a result, rules obtained by MABLAR-ST or MABLAR-CD use the same causal mechanism to provide explanations for different samples.

However, in the real-world, an effect often arises via multiple causal mechanisms [141]. For example, infection with COVID-19 or flu virus can lead to COVID-19 or flu respectively, and both COVID-19 and the flu can cause a fever. In this scenario, the *fever* is the effect and there are two causal mechanisms of fever: 1) 'infection with the COVID-19 virus' $\rightarrow$ 'COVID-19' $\rightarrow$ 'fever' and 2) 'infection with the flu virus' $\rightarrow$ 'flu' $\rightarrow$ 'fever'.

In this light, this chapter proposes Markov blanket rule generation-causal weight (MABLAR-CW) which is a variant of the MABLAR framework, addressing Sub-RO3 of RO3. The rest of this chapter is organised as follows:

Section 5.1 explains the importance of identifying possible causal mechanisms, as well as the significance of recognising the local relevant causal mechanisms for a specific outcome. Section 5.2 provides the details of MABLAR-CW. Section 5.3 conducts a series of experiments in terms on performance and explainability, exploring in particular its potential for delivering improved local explainability. Finally, Section 5.4 presents a discussion and summary.

5.1 Motivation of MABLAR-CW

As mentioned above, in the real-world an effect often arises via multiple causal mechanisms. Thus, for different samples, the specific causal mechanism leading to their respective effects, referred to as the *locally relevant* causal mechanism, may be different. In the aforementioned example, if a patient with a fever has COVID-19 but does not have flu, the locally relevant causal mechanism leads to their fever is most likely the first causal mechanism. In contrast, for another patient with a fever who only has the flu and not COVID-19, the locally relevant casual mechanism leads to their fever is the second causal mechanism.

It is critical to distinguish the *locally relevant* causal mechanism of a given effect from other *possible* but situationally unrelated causal mechanisms. Returning to the above fever example, when a doctor explains to the patient with flu why they have a fever, the doctor usually explains, 'Because you were infected with the flu virus, it resulted in you contracting flu, and flu subsequently led to your fever.' rather than saying, 'Because you were infected with the flu virus and not with the COVID virus, you contracted flu and not COVID-19. Flu caused your fever, not COVID-19.' In other words, explanations generally focus on the locally relevant causal mechanism for the given specific, i.e. 'local'–in XAI terms–case.

In addition, only by identifying the locally relevant causal mechanism can doctors implement appropriate treatment measures to cure the patient. In

other words, identifying the locally relevant causal mechanism of a given effect enables suitable treatment and is critical to enabling meaningful explanations in the context of AI.

As mentioned at the beginning of this chapter, although both MABLAR-ST and MABLAR-CD can generate rules which capture causal relationships between variables, they are designed to model a single causal mechanism. It is still a challenging area to generate fuzzy systems that capture different possible causal mechanisms of the target variable from a given data set [57]. Beyond generating working rule bases and explanations, an area which has seen comparatively little attention is the generation of fuzzy systems with the capacity to provide local explanations that distinguish the locally relevant causal mechanism for each specific effect from the set of possible captured causal mechanisms–as alluded to above.

To address the aforementioned problems, this chapter introduces MABLAR-CW. Given a causal weighted graph which is generated from a given data set, MABLAR-CW is designed to generate rules for fuzzy systems leveraging information about causal strength among variables with causal relationships. The obtained rules are designed to capture different possible causal mechanisms between the input variables and the target variable as discovered by data-driven causal discovery algorithms for the given data set. For a given sample, i.e., a specific set of inputs, the obtained fuzzy system provides local explanations which distinguish the locally relevant causal mechanism(s) of its effect from other possible, but not applicable causal mechanisms, thus avoiding both overly simplistic single-cause and exhaustive–potentially misleading explanations.

5.2 MABLAR-CW – Markov blanket rule generation using causal weight

Focusing on the classification problem, this section provides the details of MABLAR-CW, including its rationale, and its process for rule generation, prediction and local explanation generation.

5.2.1 The process of MABLAR-CW

Fig. 5.1 shows the process of MABLAR-CW. Similar to the process of MABLAR-ST and MABLAR-CD (see Fig. 3.2 and Fig. 4.1), MABLAR-CW also uses data-driven causal discovery approaches to identify a causal graph among variables within the given data set. However, different from MABLAR-ST and MABLAR-CD, it does not require identifying the MB of the target variable, because MABLAR-CW is designed to identify the locally relevant causal mechanism of the target variable. Instead, MABLAR-CW focuses on identifying possible causal mechanisms of the target variable.

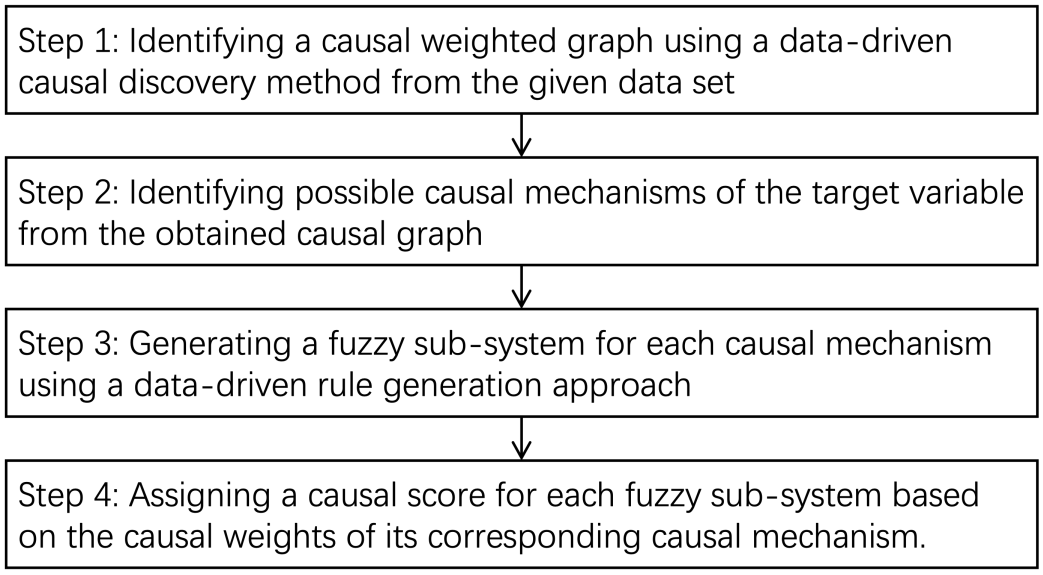


Figure 5.1: *The process of MABLAR-CW*

Although MABLAR-CW does not make use of the MB of the target variable, the principle of MABLAR-CW is consistent with that of MABLAR-ST and MABLAR-CD. That is, it refines the original data set using causal information

from a causal graph of a given data set, thereby achieving causal rule generation. Therefore, the prefix 'MABLAR-' is still used in MABLAR-CW to maintain consistency with the naming convention of the previously introduced variants.

To understand the strategy of MABLAR-CW, consider a data set $\boldsymbol{D}_{example} = \{X_1, X_2, \cdots, X_7\}$ which contain seven variables with a target variable X_5, representing the class of a sample, with values being either C_1 or C_2, i.e. $X_5 \in \{C_1, C_2\}$. Following the process shown in Fig. 5.1, Fig. 5.2 provides an illustrative example of the process of MABLAR-CW. As shown in Fig. 5.2, the

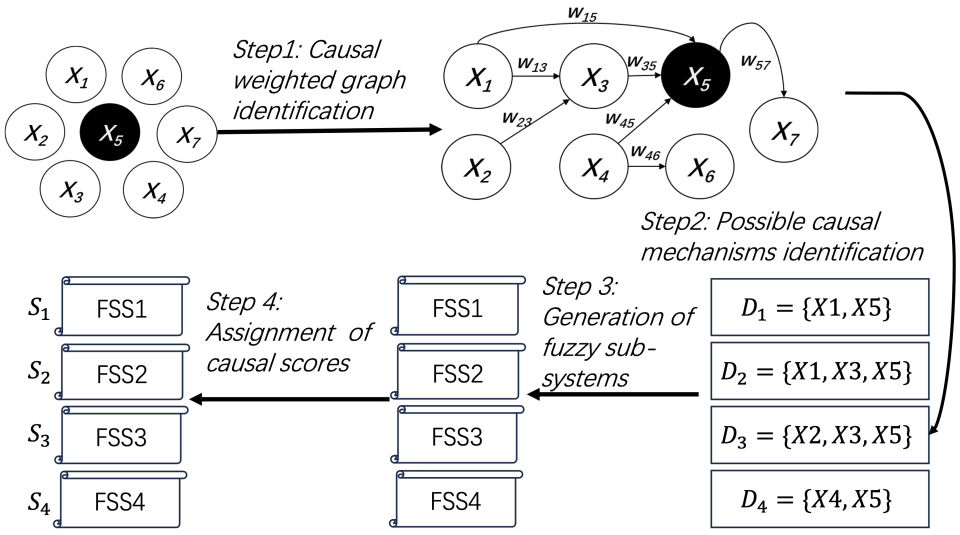


Figure 5.2: *An illustrative example of MABLAR-CW*

fuzzy system generation process of MABLAR-CW proceeds in the following steps:

Step 1: Causal weighted graph identification. In this step, MABLAR-CW generates a causal weighted graph from $\boldsymbol{D}_{example}$ using data-driven causal discover approaches. In this chapter, the DirectLiNGAM algorithm (which is introduced in Section 2.2) is adopted to achieve this goal as it offers good robustness in respect to initialisation conditions [48]. In other words, given the same data set, the results obtained by the DirectLiNGAM algorithm remain stable across different runs.

The stability of the DirectLiNGAM algorithm benefits MABLAR-CW by re-

sulting in the generation of the same local domain-based explanations for the same effect—if MABLAR-CW were to be executed multiple times on the same data set, i.e., improving the stability of local explanations. This stability of local explanation improves the trustworthiness of the generated local explanations, conditional on the correctness of the explanation itself [142]—it is noted that that trustworthiness in this space is a broad and complex question in of itself–beyond the remit of this research.

Step 2: Possible causal mechanisms identification. In this step, MABLAR-CW first identifies J complete causal paths, which is introduced in Section 2.1, from the inputs to the target variable in the causal weighted graph of the given data set, where J is the number of complete causal paths of the target variable. For the example shown in Fig. 5.2, four complete causal paths, i.e., $J = 4$ in this example, of X_5 are identified:

1. $X_1 \rightarrow X_5$
2. $X_1 \rightarrow X_3 \rightarrow X_5$
3. $X_2 \rightarrow X_3 \rightarrow X_5$
4. $X_4 \rightarrow X_5$

Then, for j belonging to $\{1, \cdots, J\}$, MABLAR-CW constructs a variable subset $\boldsymbol{D}_j$ for the jth complete causal path of the target variable. $\boldsymbol{D}_j$ contains all the variables within the jth path. For the example shown in Fig. 5.2, four subsets, i.e., $\boldsymbol{D}_1$-$\boldsymbol{D}_4$, are constructed.

Step 3: Generation of fuzzy sub-systems. In this step, MABLAR-CW generates one fuzzy sub-system from each subset obtained in Step 2 using data-driven approaches, such as the WM algorithm and FURIA (which are introduced in Section 2.5), which makes sure each identified possible causal mechanism of the target variable is modelled by one fuzzy sub-system. In this chapter, the WM algorithm is adopted, because it is simple but effective and

provides reasonably good performance [93]. For the example shown in Fig. 5.2, four fuzzy sub-systems are generated. They are 'FSS1' - 'FSS4'.

Step 4: Assignment of Causal Scores. In this step, each fuzzy sub-system is assigned a causal score based on the weights of its corresponding complete causal path of the target variable. In the example shown in Fig. 5.2, four causal scores are obtained, denoted as S_1, S_2, S_3, and S_4, respectively. The calculation and rationale of the causal scores are further introduced in Section 5.2.2.

So far, the final fuzzy system has been represented by the set of fuzzy sub-systems. The following rule is the kth rule of the jth fuzzy sub-system:

$$\begin{aligned} &\text{If } x_1 \text{ is } A_{1j}^k \text{ and } x_2 \text{ is } A_{2j}^k \text{ and ... and } x_d \text{ is } A_{dj}^k \\ &\text{Then } Class \text{ is } C_j^k \text{ with causal impact } S_j, \end{aligned} \tag{5.1}$$

where $\boldsymbol{x} = [x_1, \ldots, x_d]$ is a training sample with d input variables, A_{dj}^k is the antecedent fuzzy set of the k-th rule for the d-th input variable in the jth fuzzy system, C_j^k is the consequent class of k-th rule in the jth fuzzy system, and S_j is the causal score obtained by (5.2). For this example, $C_j^k \in \{C_1, C_2\}$ and the second rule of FSS1 in Fig. 5.2 can be 'If x_1 is A_{11}^2, Then Class is C_1 with causal impact S_1'

5.2.2 The rationale of MABLAR-CW

MABLAR-CW identifies different causal mechanisms of the target variable by leveraging the information of causal path within the causal weighted graph. In the causal weighted graph of a given data set, each complete causal path shows the relationships between the target variable and one set of its direct and/or indirect causes. Thus, each complete causal path can be viewed as one possible causal mechanism of the target variable, as represented by the causal weighted graph. By identifying all complete causal paths of the target variable in the causal weighted graph, MABLAR-CW captures different causal

mechanisms of the target variable.

The fuzzy system obtained by MABLAR-CW is constructed as a set of fuzzy sub-systems akin to an ensemble. For each identified complete causal path, MABLAR-CW generates a fuzzy sub-system using a variable subset which only contains variables on that path. By modelling complete causal paths using a set of fuzzy sub-systems, the fuzzy system obtained by MABLAR-CW captures different possible causal mechanisms shown in the causal weighted graph.

Different causal mechanisms may have different causal impacts on the target variable. As discussed in Section 2.1, in a causal weighted graph, the causal weight between two variables represents the causal strength between them. Therefore, to capture the causal impact strength of different causal mechanisms on the target variable, MABLAR-CW assigns a causal score to each fuzzy sub-system by leveraging the causal weight information between variables. The causal score is calculated as follows:

$$S_j = \sum_{i=1}^{E} \frac{W_i}{i}, \tag{5.2}$$

where S_j is the causal score of the jth fuzzy sub-system, E is the number of edges in its causal path, i is the index of the edge within the path from the path's endpoint to its starting point, and the W_i is the weight of the ith edge. For example, Fig. 5.3 shows a causal path with four variables, namely V_1 to V_4. Below each edge are their corresponding index numbers, while above each edge are their corresponding weights. In this case, the target variable is V_4 and the causal score of this path is $\frac{W_{12}}{3} + \frac{W_{23}}{2} + W_{34}$.

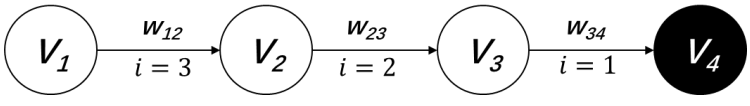


Figure 5.3: *The causal path from V_1 to V_4*

Equation (5.2) is motivated by the consideration that causal impacts of indirect causal variables on the target variable often occur through the sequential

transmission of causal effects onto the target variable via intermediate variables [47]. Here, intermediate variables refer to the variables along a causal path between the target variable and its indirect inputs. For example, in Fig. 5.3, V_3 is the intermediate variable between V_4 and V_2.

Considering that the further indirect causal variables are from the target variable, the more intermediate variables are in play to transmit their causal impact. This leads to a more pronounced attenuation of causal effects. As shown in Fig. 5.3, V_1 requires two intermediate variables while V_2 only requires one. Thus, MABLAR-CW assigns a smaller weight for the variable which is further away from the target variable in the causal path.

This thesis adopts the LiNGAM model, which assumes an effect is a linear sum of its direct/indirect cause variables [55], for causal discovery in MABLAR-CW. Thus, (5.2) sums the weighted causal impact of all variables in the corresponding causal path. However, as discussed in Section 2.1, the meaning of causal weights obtained by different methods can be different. Thus, in future research, exploring new methods to integrate causal weights would be worthwhile.

After explaining the rationale of the rule generation process of MABLAR-CW, the next two subsections demonstrate how a fuzzy system obtained by MABLAR-CW predicts a given sample and provides a local explanation for it.

5.2.3 The prediction and local explanation generation process of MABLAR-CW

Using a specific input sample $\boldsymbol{x_{test}} = [x_1^{te}, x_2^{te}, x_3^{te}, x_4^{te}, x_6^{te}, x_7^{te}]$, Fig. 5.4 illustrates the prediction process of MABLAR-CW following the same example as Fig. 5.2. Additionally, to provide a better description, Fig. 5.4 also shows each step of FSS1.

Given an input sample, MABLAR-CW generates a prediction based on the locally relevant causal mechanism as shown in the steps below.

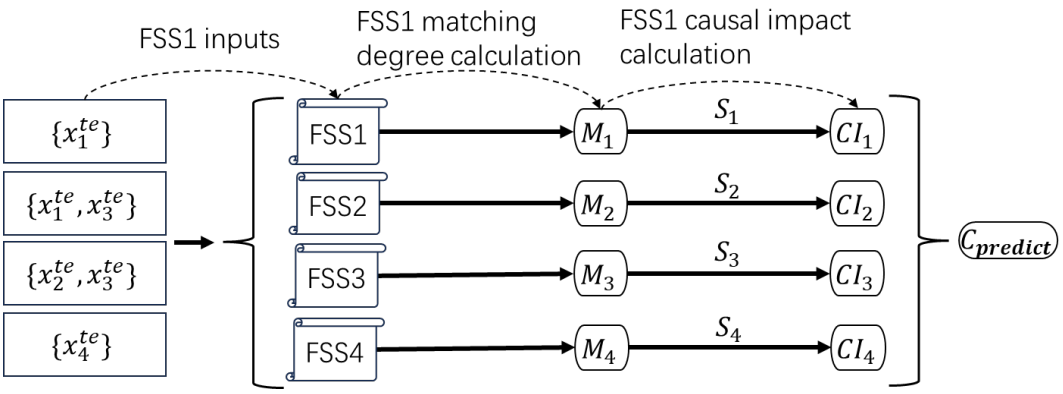


Figure 5.4: *The prediction process of MABLAR-CW*

Step 1: Matching degree calculation. The matching degree describes the extent to which the given sample matches, or is related to, a causal mechanism modelled by a fuzzy sub-system. In this step, for a given sample, the matching degree for each fuzzy sub-system is calculated in respect to the relevant subset of inputs relevant to it, as illustrated Fig. 5.4. The matching degree of each fuzzy sub-system is calculated as follows:

$$M_j = \max_k \prod_{d=1}^{D_j} \mu_{dj}^k(x_d), \tag{5.3}$$

where M_j is the matching degree of the jth fuzzy sub-system to its corresponding inputs, x_d is the dth input of the jth fuzzy sub-system, $\mu_{dj}^k(x_d)$ is the membership degree of x_d to A_{dj}^k, and K is the number of rules in the jth fuzzy sub-system.

In other words, the matching degree is the maximum firing strength of the given sample for each sub-system. The rule with the maximum firing strength of the given sample for the jth sub-system is marked as R_j^{max}. Note that this thesis uses the product t-Norm to compute the given rule's firing strength, because t-Norm preserves more information by combining all input values without discarding any, which makes it suitable for improving the prediction accuracy of the model [13].

The rationale for this is that the firing strengths indicate how well a set of inputs matches the given rule. Thus, as shown in Equation (5.3), for each fuzzy

sub-system, MABLAR-CW calculates the firing strength of each rule for its corresponding inputs. Considering that the locally relevant causal mechanism should highly match the sample and facilitate the comparison between different causal mechanisms, in Equation (5.3), only the highest firing strength is selected to represent the matching degree of each causal mechanism modelled by the corresponding fuzzy sub-system. As shown in Fig. 5.4, in this case, 4 matching degrees are obtained, i.e., M_1-M_4, corresponding to R_1^{max}-R_4^{max}. Considering only the rule with the highest firing strength may result the obtained fuzzy system to lose information that could be meaningful for the overall decision, especially when several rules have very similar maximum values. This could decrease the performance of the obtained system, such as reducing the classification accuracy of the obtained system. Currently, the winner-takes-all strategy is adopted to maintain the simplicity of the local explanations generated by MABLAR-CW. In future research, methods to calculate matching degrees which retain more information while maintaining simplicity will be explored.

Step 2: Locally relevant causal impact calculation. As discussed in Section 5.1, a locally relevant causal mechanism is one which matches the input sample, i.e., it is possible given the inputs, and it should have a strong causal impact on the effect. Thus, in this step, MABLAR-CW calculates the locally relevant causal impact CI_j of each jth fuzzy sub-system using $CI_j = M_j \cdot S_j$, i.e. the product of the causal score arising from the causal weighted graph and the firing strength of the given input set for R_j^{max}, i.e., the matching degree M_j.

Step 3: Final prediction via voting. The final prediction for the given sample is selected as the consequent of R_j^{max} which produces the highest CI_j. MABLAR-CW uses the 'winner takes all' voting strategy to reflect the intention to focus on the most locally relevant causal mechanism. For this example, suppose the consequent of R_j^{max} with the highest CI_j is C_1, the the final prediction for $\boldsymbol{x_{test}}$, i.e., $C_{predict}$ in Fig. 5.4, would be C_1. Of course, alterna-

tive approaches could explore combining a set of best-fitting rules and causal mechanism reflecting that in the the real world frequently multiple causal mechanisms come together to underpin an effect. This is worth studying in the future.

Finally, given a sample, the local explanation of its effect predicted by the fuzzy system obtained by MABLAR-CW is R_j^{max} which results the highest CI_j, because, as explained above, CI_j measures the locally relevant causal impact of the jth fuzzy sub-system on the predicted effect.

Having established both the prediction and explanation stages of MABLAR-CW, the next section proceeds to experiment evaluation and analysis.

5.3 Experiment Results

5.3.1 Experiment settings

MABLAR-CW is compared with 3 different rule generation approaches: MABLAR-ST, MABLAR-CD and the WM algorithm, respectively. As discussed in Section 5.1, MABLAR-ST or MABLAR-CD only models a single causal mechanism of the target variable and uses the modelled causal mechanism to provide local explanations for different samples. In contrast, MABLAR-CW models different causal mechanisms of the target variable and provides local explanations for different effects by identifying the locally relevant causal mechanism of the specific sample. Thus, the comparison between MABLAR-ST, MABLAR-CD and MABLAR-CW highlights the difference between different ways of using causal mechanisms for generating local explanations. The WM algorithm, as arguably the most classical data-driven rule generation approach, is adopted as the baseline correlation-based rule generation approach, because it provides a consistent basis to compare all approaches evaluated [123]. The WM algorithm is also adopted in the corresponding rule generation step of MABLAR-ST, MABLAR-CD and MABLAR-CW to maintain consistency be-

tween different frameworks.

As the number of variables increases, the resulting causal graph may include more complete causal paths to the target variable, which significantly increases the runtime of MABLAR-CW. To ensure the feasibility of repeated experiments and to maintain reproducibility, we selected datasets with smaller number of variables. Thus, five real-world data sets are selected, which are also widely used as benchmarks for rule generation, from the UCI data repository [122] and the Kaggle website [143]. Table 5.1 shows the details of the selected data sets.

Table 5.1: Data sets used in this chapter

Name	Samples	Class	$\|D\|$	$\|D_{MB}\|$	$\|D_{CD}\|$
Breast[122]	699	2	9	6	6
Iris[122]	150	3	4	4	1
Mammographic mass (MAM) [143]	830	2	6	4	4
Pima Indian Diabetes (PID)[143]	768	2	8	8	6
Wine[122]	178	3	13	13	5

Trapezoidal and triangle membership functions are adopted as they facilitate explainability [13]. One can adopt different membership functions for specific problems. Here, each variable is divided into 3 fuzzy partitions for the convenience of facilitating the assignment of linguistic labels. The parameters of membership functions are estimated by the K-means clustering method. More details can be found in Section 3.5 which follows the same approach. All variables are normalized to [0,1].

The average classification accuracy over 5-fold cross-validation is adopted as the performance index. Stratified sampling is adopted to keep the original class proportions in the training data set of each fold the same. In order to keep the consistency of causal weighted graphs and fuzzy partitions across different data set partitions during the cross-validation process, the causal discovery and the determination of fuzzy partitions are implemented before cross-validation.

More specifically, DirectLiNGAM is first implemented on the original data set to obtain the causal weighted graph. Then K-means is implemented on the original data set to determine the fuzzy partitions of each variable. Finally, the WM algorithm, MABLAR-ST, MABLAR-CD, and MABLAR-CW are applied using the initially determined causal weighted graph and partitions as applicable.

5.3.2 Performance evaluation

MABLAR-CW is designed to maximise the local explainability of fuzzy systems. As discussed in Section 5.2.3, MABLAR-CW adopts the 'winner-takes-all' strategy to maintain the simplicity of the generated local explanations, which has the risk of decreasing the performance of the obtained systems. Thus, to evaluate the impact of adopting the 'winner-takes-all' strategy on performance, the first comparison is the performance of the fuzzy systems obtained by the different frameworks.

Table 5.2: Performance of different frameworks

	WM	MABLAR-ST	MABLAR-CD	MABLAR-CW
Breast	0.9128	0.9571	0.9571	**0.9686**
Iris	**0.9667**	**0.9667**	0.6667	0.6667
MAM	0.8145	0.8277	0.8277	**0.8386**
PID	0.6850	**0.6889**	0.6276	0.6238
Wine	0.7521	0.7578	0.8711	**0.9160**

Table 5.2 shows the results of each fuzzy system. It is noted that the fuzzy systems obtained by MABLAR-CW achieve higher or comparable performance with other variants in most data sets, however, the performance of MABLAR-CW has a significant decrease in the Iris data set, as well as MABLAR-CD– which warrants further investigation.

Fig. 5.5 shows the causal weighted graph of the Iris data set obtained by DirectLiNGAM in this chapter. According to [144], both 'petal length' and 'petal width' should be used to classify an iris flower. However, as shown in

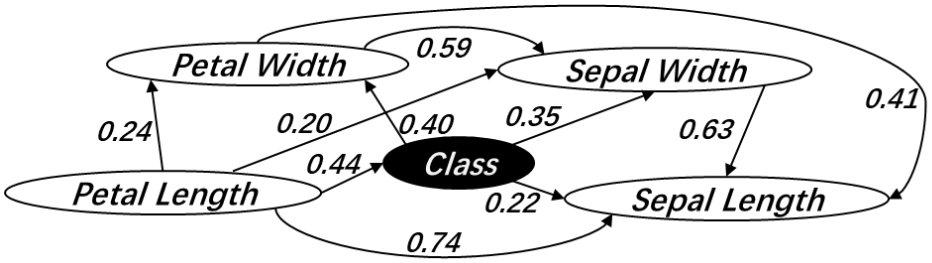


Figure 5.5: *The weighted causal weighted graph of the Iris data set*

Fig. 5.5, only the 'petal length' variable is identified by the algorithms as a direct cause variable for the target variable (i.e., the 'Class' node in Fig. 5.5), which indicates that DirectLiNGAM does not perform well on the Iris data set. The sub-optimal causal weighted graph leads to insufficient information for classification in MABLAR-CD and MABLAR-CW, resulting in a decrease in the model performance.

This may be a result of the Iris dataset not meeting the conditions for LiNGAM. As introduced in Section 2.2, LiNGAM assumes the causal relationships between variables should be linear, while the causal relationships between variables within the Iris dataset may be non-linear, which results in a suboptimal causal weighted graph. In real-world applications, causal discovery algorithms suitable for a given data set should be chosen (See [108, 110] for recent overviews, including the selection of causal discovery algorithms). However, the example highlights that the quality of the MABLAR framework are dependent on the quality of the causal weighted graph.

Overall, the performance is considered to be satisfactory and sufficient to meaningfully consider the explanations generated by the approach in most data sets. Focusing on the MAM data set, the next subsection compares local explanations provided by the different variants of the MABLAR framework. The MAM data set is selected because the causal weighted graph complexity of the MAM data set is moderate, which is suitable for visualisation.

5.3.3 Local explanations — a case study

The MAM data set is used to predict the severity (benign or malignant) of a mammographic mass [145]. Fig. 5.6 shows the causal weighted graph of the MAM data set obtained by the DirectLiNGAM algorithm in this chapter.

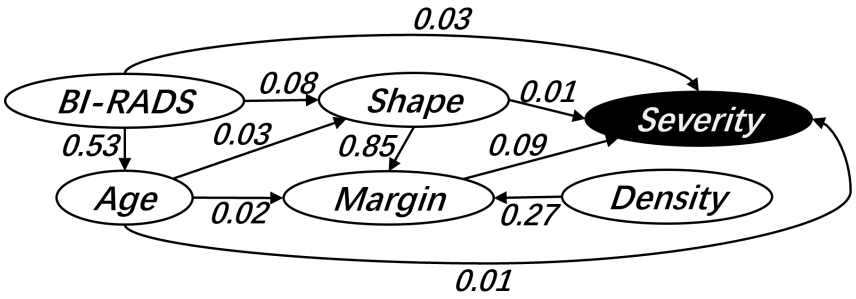


Figure 5.6: *The weighted causal weighted graph of the MAM data set*

From Fig. 5.6, 8 complete causal paths to the target variable *Severity* are identified:

1. BI-RADS $\xrightarrow{0.03}$ Severity
2. BI-RADS $\xrightarrow{0.53}$ Age $\xrightarrow{0.01}$ Severity
3. BI-RADS $\xrightarrow{0.53}$ Age $\xrightarrow{0.02}$ Margin $\xrightarrow{0.09}$ Severity
4. BI-RADS $\xrightarrow{0.53}$ Age $\xrightarrow{0.03}$ Shape $\xrightarrow{0.85}$ Margin $\xrightarrow{0.09}$ Severity
5. BI-RADS $\xrightarrow{0.53}$ Age $\xrightarrow{0.03}$ Shape $\xrightarrow{0.01}$ Severity
6. BI-RADS $\xrightarrow{0.08}$ Shape $\xrightarrow{0.85}$ Margin $\xrightarrow{0.09}$ Severity
7. BI-RADS $\xrightarrow{0.08}$ Shape $\xrightarrow{0.01}$ Severity
8. Density $\xrightarrow{0.27}$ Shape $\xrightarrow{0.01}$ Severity

In addition, in this example, the MB of the target variable (i.e. 'Severity') are the variables 'BI-RADS', 'Age', 'Shape' and 'Margin', and the direct causes of 'Severity' are also the same variables, because the target variable has no child or spouse nodes in Fig. 5.6.

Three samples are selected from the MAM data set to help illustrate key difference between the local explanations provided by the different variants of the MABLAR framework for each sample. The input samples marked as ①, ②, ③, respectively, as well as the associated output class are shown in Table 5.3. Note that all scalar values have been normalised to $[0, 1]$:

Table 5.3: The tested samples

	BI-RADS	Age	Shape	Margin	Density	Class
①	0.0909	0.5128	1	1	0.6667	Malignant
②	0.0909	0.7436	0	0.75	0.6667	Malignant
③	0.0909	0.6282	1	1	0.6667	Malignant

Table 5.4-5.7 show the local rule-based explanations provided by the different variants of the MABLAR framework for different samples in table form. Note that all rules use the AND logical connective exclusively, implemented as the product t-Norm as mentioned above.

Table 5.4: Local explanations provided by the WM algorithm

	BI-RADS	Age	Shape	Margin	Density	Class
①	Mid	Mid	High	High	High	Malignant
②	Mid	High	Low	Mid	High	Malignant
③	Mid	High	High	High	High	Malignant

Table 5.5: Local explanations provided by MABLAR-ST

	BI-RADS	Age	Shape	Margin	Class
①	Mid	Mid	High	High	Malignant
②	Mid	High	Low	Mid	Malignant
③	Mid	High	High	High	Malignant

In Table 5.7, the symbol '×' represents that the corresponding variable is no included in the antecedents of the local explanation. From Table 5.4-5.7, the following can be observed: The local explanations provided by the WM algorithm, MABLAR-ST or MABLAR-CD have the same variables in their

Table 5.6: Local explanations provided by MABLAR-CD

	BI-RADS	Age	Shape	Margin	Class
①	Mid	Mid	High	High	Malignant
②	Mid	High	Low	Mid	Malignant
③	Mid	High	High	High	Malignant

Table 5.7: Local explanations provided by MABLAR-CW

	BI-RADS	Age	Shape	Margin	Class
①	Mid	Mid	High	High	Malignant
②	Mid	High	Low	Mid	Malignant
③	Mid	×	High	High	Malignant

antecedents for different samples. More specifically, the WM approach use all variables to provide local explanations for different samples, while MABLAR-ST and MABLAR-CD uses only the variables 'BI-RADS', 'Age', 'Shape' and 'Margin' to provide local explanations for different samples. In contrast, local explanations provided by MABLAR-CW have different antecedents for different samples.

This observation supports the conclusion that the system obtained MABLAR-CW can capture different causal mechanism of the target variable, and the local explanations provided by MABLAR-CW are based on the locally relevant causal mechanisms of the effect of different samples.

The comparison between the local explanations for ① and ③ provided by the different variants of the MABLAR framework is further highlighted. Note that, as shown in Table 5.3, all values for ③ are the same as for input ①, except the age value, which is greater for ③. Having only one variable with differing values makes the comparison easier.

As shown in Table 5.4-Table 5.6, for these two samples or input-effect pairs, the local explanations in each case use the same causal mechanism to explain why ① and ③ are malignant. For example, the local explanation provided by MABLAR-CD indicates that 'BI-RADS', 'Age', 'Shape' and 'Margin' are the

causes for the 'Malignant' of both ① and ③.

However, according to [146], the authors support the idea that a mammographic mass in middle-aged women is more likely to be malignant than in older women. In other words, in the causal mechanism that leads to a malignant mammographic mass for a middle-aged woman, "Age is middle" should be included. In contrast, in the causal mechanism that leads to a malignant mammographic mass for an elderly woman, "Age is high" should not be considered. Thus, it is intuitive for that 'Age is Mid' should be included in the local explanations for the effect of ①. However, 'Age is High' should be excluded in the local explanations for the effect of ③ as it is an unnecessary component in the local explanations for ③.

Noting this, reviewing Table 5.4 - Table 5.7 reveals that local explanations for ③ provided by WM, MABLAR-ST and MABLAR-CD all include 'Age is high'. In contrast, 'Age is high' is not present from the local explanation provided by MABLAR-CW for ③. This observation supports that MABLAR-CW focuses explanations on the most locally relevant causal path. More specifically, the local explanations shown in Table 5.7 indicate that MABLAR-CW provides the local explanation for ① based on the causal mechanism 'BI-RADS $\xrightarrow{0.53}$ Age $\xrightarrow{0.03}$ Shape $\xrightarrow{0.85}$ Margin $\xrightarrow{0.09}$ Severity', while provides the local explanation for ① based on the causal mechanism 'BI-RADS $\xrightarrow{0.08}$ Shape $\xrightarrow{0.85}$ Margin $\xrightarrow{0.09}$ Severity'.

However, of course, this is only an illustrative example and broader evaluation will be necessary to underpin firmer conclusions, including comparing local explanations across more samples on additional data sets and conducting user studies. As the evaluation of the quality of explanations is significantly more challenging than the evaluation of system performance, the latter will not be a straightforward or quick process, but will required sustained research efforts in the future.

5.4 Summary

This chapter proposes MABLAR-CW, which is an variant of MABLAR. MABLAR-CW is designed to focus on locally-relevant causal mechanisms rather than correlations between variables–both for the generation of outputs, and for producing meaningful locally relevant explanations.

Given a data set and its corresponding causal weighted graph, MABLAR-CW uses causal path information from the associated causal weighted graph to generate a fuzzy system constructed as a set of fuzzy sub-systems. The obtained fuzzy system captures possible causal mechanisms of the target variable. Given an input sample, MABLAR-CW uses causal weight information from the causal weighted graph to identify the most causally relevant mechanism and generate an output–and explanation.

Initial experiments indicate that MABLAR-CW has the capacity to produce fuzzy systems which can provide locally relevant explanations by identifying the locally relevant causal mechanism for a given sample, while maintaining comparable performance compared to classical rule generation approaches.

Finally, it should be noted again that MABLAR-CW is dependent on the quality of the causal weighted graph and thus generally, the quality and suitability of the causal discovery algorithm used. In addition, a fuzzy system obtained by MABLAR-CW is constructed as a set of fuzzy sub-systems akin to an ensemble with each fuzzy sub-system model on causal mechanism in a causal weighted graph of a given data set. When the causal weighted graph of the given data set shows a large number of possible causal mechanisms, MABLAR-CW has the risk of generating a fuzzy system which contains a large number of fuzzy sub-system, which increases the risk of contain a large number of rules in the final obtained system. Thus, MABLAR-CW is suggested to be used when the task is focused on the local explanations of a specific output.

As discussed above, MABLAR-CW adopts winner-takes-all strategy to make final predictions, which has the risk of losing information that could be mean-

ingful for the overall decision. In future, predictive approaches which combine a set of best-fitting rules and causal mechanisms will be explored, aiming to better utilize information useful for prediction while maintaining the simplicity of the generated local explanations.

So far, three different variants of MABLAR have been introduced, i.e., MABLAR-ST, MABLAR-CD and MABLAR-CW, respectively. They are designed to provide different types of domain-based explanations in different scenarios. Explanations provided by above mentioned variants of MABLAR focus on factual explanations. In other words, such explanations are based on the given input. However, in real-world applications, users may not only want to know why a given output was reached, but also to know why other potential outputs were not. For example, if a person is considered to be ineligible for a loan by an AI loan system, they may not only want to know why the AI system made that decision but also what they can do to get a loan in the future. To supplement the explanations provided by fuzzy system with counterfactual explanations, the next chapter introduces counterfactual Markov blanket rule generation framework (CF-MABLAR), which is an variant of MABLAR designed to generate counter factual rules.

Chapter 6

Towards counterfactual rule generation for fuzzy systems

The previous chapters show how rules obtained by the MABLAR framework can capture causal relationships between variables. These variants leverage different types of causal information extracted from a given data set. Rules obtained by the different variants provide explanations which reveal different facets of causal relationships between variable.

This chapter goes a step further. To complement an explanation provided by the rules obtained by the MABLAR framework, this chapter proposes Counterfactual Markov blanket rule generation (CF-MABLAR) which is a meta-variant of the MABLAR framework, addressing RO4. CF-MABLAR first generates rules using a variant of the MABLAR-framework to approximate the causal links between inputs and output(s) of fuzzy systems. Then it generates counterfactual (CF) rules based on rules obtained by a variant of the MABLAR-framework (such as MABLAR-ST). Uniquely, the obtained CF rules not only provide a basic model-based explanation, but can also articulate how the given inputs would need to be changed to generate a different output, crucial for lay-user insight, verification and sensitivity-evaluation of XAI systems, for example in decision support around credit risk, cyber security and medical assistance.

The rest of this chapter is organised as follows: Section 6.1 discusses why CF explanations are meaningful for XAI. Section 6.2 analyses the question which is expected to be answered by a CF explanation generated by CF-MABLAR, and shows the details of CF-MABLAR. Section 6.3 uses the Iris data set as an example to illustrate how CF-MABLAR works. Section 6.4 conducts experiments on several real-world data sets to validate the effectiveness of CF-MABLAR in generating counterfactual explanations. Finally, Section 6.5 summaries the work of this chapter.

6.1 Motivation of CF-MABLAR

As discussed in Section 2.3, rules of fuzzy systems can provide model-based explanations for a specific input through a set of executed, or fired, rules. For example, 'This flower is Iris Setosa, because its sepal length, sepal width, petal length and petal width are all low' is a model-based explanation as can be derived from a rule of an Iris classification fuzzy system. Broadly speaking, such model-based explanations can also be called *factual explanations* [23], because they describe which variables contributed to the model's output based on the given input.

Beyond the 'factual' explanations used in AI systems, human explanations are often counterfactual (CF) [147]. When a model gives a particular output, users benefit from not only knowing why a particular output is obtained, but also why other outputs are not obtained. For example, in the above Iris case, why the flower is classified to Iris Setosa, and not Iris Versicolour? To explain this, explanations need to contain CF elements. In the above case, a potential CF explanation would be 'This flower would be classified as Iris Versicolour if its sepal width and sepal length were medium'. Here, 'sepal width and sepal length were medium' is a CF element, because the given inputs of sepal width and sepal length are low, and not medium. The explanations which contain CF elements are called CF explanations [23].

Recently, aiming at the problem of generating CF explanations for rule-based models, many data driven CF explanation generation methods have been proposed. For example, Stepin et al. [148] proposes a CF explanation method, marked as the correlation-counterfactual method (Cor-CF), for decision tree models and fuzzy rule-based models. Stepin et al. [149] also proposed a CF explanation generation method based on the fuzzy unordered rule induction algorithm [85]. However, these approaches do not focus on the challenge that where models are correlation-based, explanations, including CF explanations, reflect correlations, rather than causal relationships.

Humans expect rules of fuzzy systems to reflect causal relationships [1]. Consequently, there is a desire for fuzzy systems to provide CF explanations based on their rules which reflect causal relationships. In addition, as discussed in Section 3.1, rules of fuzzy system which reflect causal relationships have the potential to tell users not only what happened in the system, but also what happens in the real-world. Thus, CF explanations based on causal rules have the potential to guide users not only in changing inputs to provoke a change in the output of the system, but also in changing relevant variables in the real world, thereby achieving corresponding changes in *the real-world*.

For example, someone whose application for a bank loan was rejected may be able to get specific insight on how to manage their income/spending patterns to increase their chanced for receiving a loan. However, how to design an approach to generate CF explanations based on causal relationships between variables is still an open problem. To address this issue, this chapter introduces CF-MABLAR, as detailed in the next section.

6.2 CF-MABLAR – Counterfactual Markov blanket rule generation

This section first defines the question expected to be answered by CF explanations generated by CF-MABLAR, then shows the details of CF-MABLAR.

6.2.1 The question to be answered by CF-MABLAR

The CF explanations generated by CF-MABLAR are expected to answer the following question:

- Intervention Question: Suppose a fuzzy system F has already been obtained. Given a test sample $\boldsymbol{x}_{test}$ with d inputs, i.e., $\boldsymbol{x}_{test} = [x_1, x_2, ..., x_d]$, the output of $\boldsymbol{x}_{test}$ obtained from F is A. How should one intervene, in other words, how should $\boldsymbol{x}_{test}$ be changed to make the output of the system become another output B?

Here, B is another output of F, and the term 'intervened' means changing values. For example, the intervention question in the Iris case can be 'What should the intervention of the input variables be for the given flower to be classified as Iris Versicolour'.
The above intervention question is further divided into the following two sub-questions:

- Sub-Question 1 (SQ1): *Which input variables should be changed?*
- Sub-Question 2 (SQ2): *How and how much should they be changed (e.g. could be increase or decrease)?*

In real world applications, a data set for a machine learning task usually contain many variables. However, variables within data sets for machine learning tasks usually have high correlation with the output variable, but not all of them have a causal relationship with the output variable. Interventions on

these variables may lead to changes in the output of the model, however, these interventions will not result in corresponding changes in *the real world*.

For example, considering the case discussed by Freitas [150]. Here, a prediction model is trained to discriminate between enemy and friendly tanks using photos of tanks. However, the users found that the model makes a prediction based on the weather shown in the photo, because all photos of friendly tanks were taken on sunny days, while all photos of enemy tanks were taken on overcast days. In this example, suppose a tank in a photo is an enemy tank and is correctly classified by the model as such. Altering the weather in the photo can lead the model classify the tank as friendly. In the real world, however, the tank remains an enemy tank after altering the weather in the photo, because the relationship between the weather in the photo and the category of a tank is a correlationship rather than a causal relationship.

As mentioned in Section 6.1, the CF explanations provided by CF-MABLAR are expected to guide users not only obtain the desired output from a system, but also to obtain a desired outcome in the real world. Thus, for SQ1, CF-MABLAR should focus on variables which are causally related to the output variable.

In addition, in the tank classification example, the weather shown in the photo could be 'changed from overcast to sunny using image-editing software' or 'overcast to rainy' etc. Thus, it is important to decide which one should be selected. According to the Occam's razor, for SQ2, CF-MABLAR seeks to obtain the expected output with the minimal intervention/change on the inputs.

6.2.2 The process of CF-MABLAR

The previous subsection defines the question expected to be answered by the CF explanations generated by CF-MABLAR. Focusing on Mamdani fuzzy systems and classification problems, this subsection shows the details of CF-

MABLAR and explains the rationale of CF-MABLAR.

Given a data set and a test sample (i.e., a sample to be explained), CF-MABLAR generates a counterfactual explanation for the test sample by the process shown in Fig. 6.1.

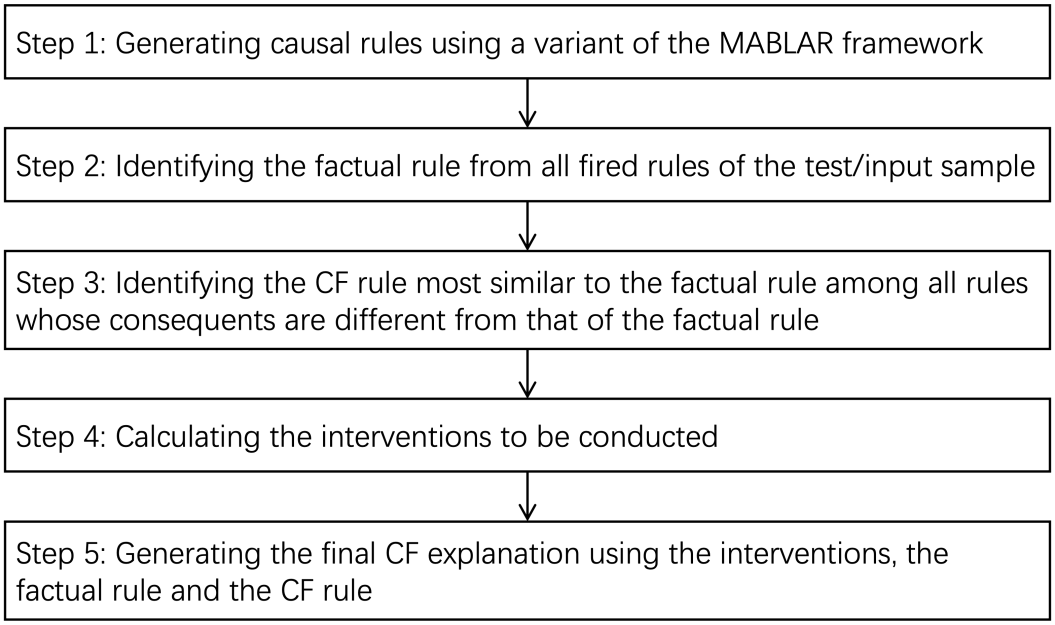


Figure 6.1: *The process of CF-MABLAR*

The steps of CF-MABLAR in Fig. 6.1 are detailed as follows:

Step 1: Generating causal rules using the MABLAR framework. In this step, CF-MABLAR generates a fuzzy system from the given data set using a variant of the MABLAR framework. In this step, the selection of the MABLAR variant should depend on the real-world application's need for factual explanations. As introduced in the previous chapters, rules of different variants of the MABLAR framework reflect different types of causal relationships between variables. The applicable scenarios for the factual explanations provided by different variants can refer to the previous chapters where the corresponding variants are introduced.

Step 1 is to solve SQ1. In Step 1, all rules are obtained by a variant of the MABLAR framework, such as MABLAR-ST. As mentioned in Section 3.1, all input variables in rules obtained by the MABLAR framework should have a causal relationship with the output(s). Thus, the intervention on the inputs

of rules obtained by the MABLAR framework should affect the output(s) not only within the system, but also in the real-world.

Step 2: Identifying the factual rules. In this step, CF-MABLAR first collects all fired rules of the test sample, then identifies the rule with the highest firing strength of the test sample, marking it as *the factual rule* for the test sample. The factual rule of the test sample is used to form the factual part of the CF final explanation (an example of a final CF explanation will be shown in Section 6.3.2).

As discussed in Section 6.1, all fired rules of the test sample can be used to provide factual explanations for the test sample. However, to keep the final CF explanation concise, only the rule with the highest firing strength is considered in CF-MABLAR.

Step 3: Identifying the CF rule most similar to the factual rule. In this step, CF-MABLAR collects all rules which have a different consequent with the factual rule, and calculates their similarity to the factual rule. In addition, when there is more than one consequent different from the factual rule, such as in multi-class classification problems, the user should specify a desired consequent.

In CF-MABLAR, the Similarity-Rule index (the SR index) proposed by Garcia et al. [151] is adopted to measure the similarity between two rules due to its simpleness, which is calculated as follows:

$$SR(k_1, k_2) = \sum_{i=1}^{D} S(A_i^{k_1}, A_i^{k_2}), \tag{6.1}$$

where $SR(k_1, k_2)$ represents the similarity between rule k_1 and k_2. D is the number of inputs. $A_i^{k_1}$ and $A_i^{k_2}$ are the antecedent fuzzy sets of the i-th input for rule k_1 and k_2, respectively. $S(A_i^{k_1}, A_i^{k_2})$ is the similarity between $A_i^{k_1}$ and $A_i^{k_2}$. CF-MABLAR adopts the Jaccard ratio [152] to calculate the similarity

between two fuzzy sets. Thus,

$$S(A_i^{k_1}, A_i^{k_2}) = \frac{\int_{x \in X} \min(\mu_{A_i^{k_1}}(x), \mu_{A_i^{k_2}}(x))}{\int_{x \in X} \max(\mu_{A_i^{k_1}}(x), \mu_{A_i^{k_2}}(x))}. \tag{6.2}$$

As discussed in Section 6.2.1, CF-MABLAR is designed to obtain the expected output with the minimal intervention. Thus, the rule with the highest SR index is selected as the CF rule for the test sample, and is used to form the factual part of the final CF explanation.

Step 4: Calculating the interventions to be conducted. In this step, CF-MABLAR first finds the value x_i^{cf} which can make the i-th input have the highest membership degree in rule R^{CF} on $A_i^{k_{CF}}$:

$$x_i^{cf} = \arg\max_{x_i} \mu_{A_i^{k_{CF}}}(x_i) \tag{6.3}$$

where $\mu_{A_i^{k_{CF}}}(x_i)$ is the membership degree of x_i to $A_i^{k_{CF}}$ in rule R^{CF}. R^{CF} is the counterfactual rule whose consequent part is the expected output and has the highest similarity with the factual rule.

Then, CF-MABLAR calculates the difference between x_i and x_i^{cf} in Step 4:

$$diff_i = x_i^{cf} - x_i \tag{6.4}$$

Here, $diff_i$ is the interventions for x_i and is used to form the conclusion part of the final CF explanation.

Step 4 is used to solve SQ2. As shown above, CF-MABLAR makes (slight) interventions on all variables in the test sample which are included in the antecedent of the factual rule. However, in real-world applications, it is sometimes unnecessary to intervene on all variables causally related to the target variable; instead, a substantial intervention may be performed on only a subset of these variables, such as significantly changing one variable. In this thesis, to ensure that the system produces an output corresponding to the consequent of the CF rule when predicting the intervened test sample, CF-MABLAR is

designed to intervene all variables causally related to the target variable. Future research will focus on further reducing the number of variables requiring intervention.

Step 5: The final CF explanation generation. In this step, CF-MABLAR generates the final CF explanation by combining the factual rule, '$diff_i$' and the CF factual rule. The final CF explanation consists of a factual part, a CF part and a CF conclusion part, and an example of a final CF explanation generated by CF-MABLAR will be shown in Section 6.3.

6.2.3 Properties of CF explanations obtained by CF-MABLAR

CF-MABLAR are designed to generate CF explanations. CF explanations are helpful for complementing a meaningful explanation from at least the following two aspects [23]:

- CF explanations can help users–commonly expert-users–to assess whether an AI is making decisions as expected. An example here is tank classification example above, where a CF explanation can help highlight that the discrimination strategy of the AI is not reasonable.

- Where validating the correct functionality is not the focus, CF explanations can guide users to understand how they can change inputs to provoke a change in output. For example, CF explanations can help a credit-applicant understand that their high-spending behaviour is the reason their loan application was rejected and inform them to what degree they need to change this behaviour in order to be considered for a loan.

The CF explanations generated by CF-MABLAR are based on causal relationships between variables which can improve the explainability of a Mamdani fuzzy system by reducing the complexity of CF explanations. More specifically,

existing CF explanation generation methods, such as Cor-CF[149], tend to use all variables to generate CF explanations. However, usually, not all variables in a data set have causal relationships with the output variable. CF-MABLAR only uses the variables which are causally related to the output variable(s) to generate CF explanations.

A set of variables which contains only those causally related to the target variable often contains fewer variables than the whole data set. For example, the MB of the output variable(s) usually contains a smaller (or equal) number of variable than the whole data set [111]. Thus, CF explanations obtained by CF-MABLAR contain less elements than the CF explanation obtained by correlation-based methods.

In addition, it should be noted that the set of variables used by CF-MABLAR for CF explanation generation depends on the variant used for the factual explanation generation. For example, if CF-MABLAR uses MABLAR-ST to generate factual explanations, the CF explanations generated by CF-MABLAR will only focus on the variables within the Markov blanket (MB) of the output variable.

6.3 CF-MABLAR illustration – Iris classification

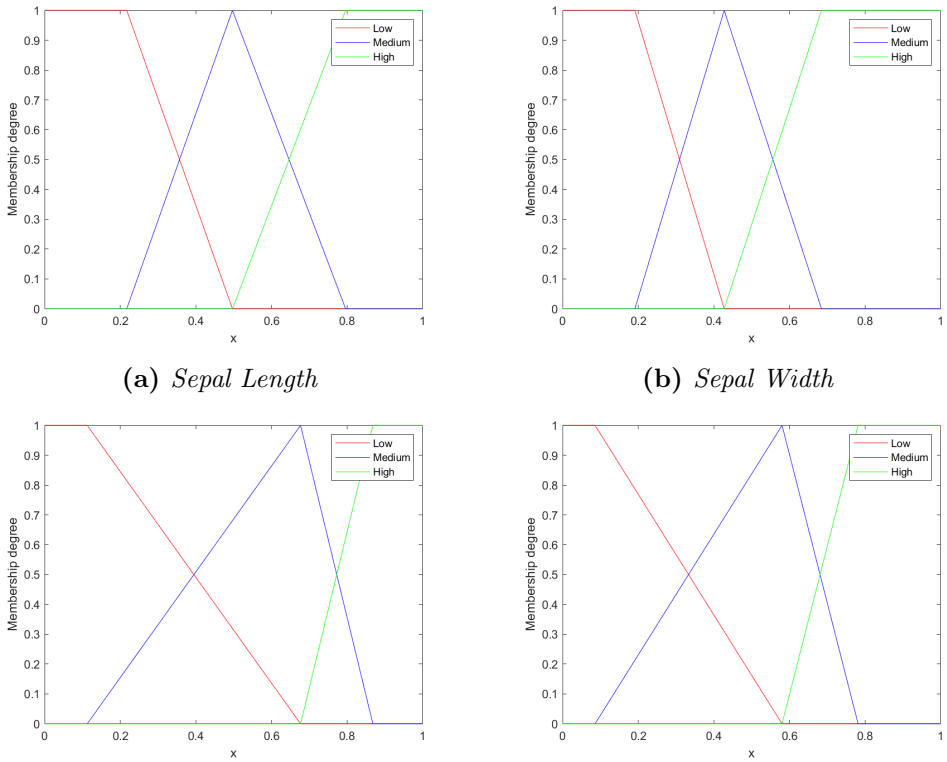
This section uses the Iris data set as an example to show how CF-MABLAR works.

6.3.1 Problem description and settings

The Iris data set [122] can be used to predict which species of iris flower a sample belongs to based on its variables. The data set contains five variables, i.e., inputs of the fuzzy system. The five variables are 'Sepal Length'(SL), 'Sepal Width'(SW), 'Petal Length'(PL), 'Petal Width'(PW) and 'class', respectively.

The 'class' variable is the output variable while the rest four variables are the input variables.

Fig. 6.2 shows the fuzzy partitions of each feature and the corresponding linguistic term. Note that all variables have been normalized to [0,1]. To keep rules concise, each input variable is divided into three fuzzy sets, also facilitating the assignment of linguistic labels to each fuzzy partitions. For discrete variables, the singleton membership function is adopted and the linguistic label of each singleton fuzzy partition is the same as the original linguistic label. In this example, the output variable of the Iris data set is a discrete variable and its original linguistic labels are 'Iris Setosa', 'Iris Versicolour', and 'Iris Virginica', respectively. Thus, the corresponding linguistic labels are 'Iris Setosa', 'Iris Versicolour', and 'Iris Virginica', respectively.



(a) *Sepal Length* **(b)** *Sepal Width*

(c) *Petal Length* **(d)** *Petal Width*

Figure 6.2: *Fuzzy partitions of each feature*

For simplicity, all the samples of 'Iris Setosa' and 'Iris Versicolour' are selected. Thus, the problem becomes a binary classification problem and the data set contains 100 samples.

6.3.2 Illustrative example of the CF-MABLAR framework

This example is demonstrated following the steps shown in Fig. 6.1.

Step 1: Generating causal rules using the MABLAR framework. Here, MABLAR-ST is adopted because this example focuses on how to classify Iris species, rather than exploring the direct causes of differences in Iris species or the locally relevant causal mechanisms that make a particular Iris flower belong to a certain species. In real-world applications, as discussed in Section 6.2.2, the selection of MABLAR variants should be problem-dependent. In this step, the PC-algorithm [51] and the Wang-Mendel algorithm [153] are used to generate the causal graph and generate rules, respectively. One can also use other algorithms according to specific applications.

The obtained MB of the target variable contains 'Sepal length', 'Petal length' and 'Petal width'. Note that, the obtained causal graph may still not be perfect. According to [144], biologists actually only use 'Petal length' and 'Petal width' to classify Iris, which indicates that botanists consider only these two variables are causally related to the categories of Iris. Thus, only these two variables are expected to be included in the MB of the output variable in the Iris data set. Part of rules of the obtained Mamdani fuzzy system are shown in Table 6.1.

Table 6.1: Part of rules of the obtained fuzzy system

	Sepal length	Petal length	Petal width	Class
R1	Low	Low	Low	Setosa
R2	Medium	Low	Low	Setosa
R3	Medium	Medium	High	Versicolour

Step 2: Identifying the factual rule. To make sure the selection is unbiased, a sample is randomly selected from the data set. The selected sample belongs to the Setosa class. The feature value of the selected sample is *sepal length* = 0.2963, *petal length* = 0.0976 and *petal width* = 0.0588, re-

spectively. The sample is correctly classified into the Setosa class and the factual rule of the sample is '*The sepal length is low and the petal length is low and the petal width is low, then class is Setosa*'.

Step 3: Identifying the CF rule most similar to the factual rule. The consequent of the factual rule is Setosa. Thus, all rules of which the consequent part is Versicolour, i.e., NOT Setosa, are collected. In total, there are nine rules which have Versicolour consequent part in the obtained Mamdani rule base, which is shown in Table 6.2.

Table 6.2: Counterfactual rules of the selected sample

	Sepal Length	Petal Length	Petal Width	Class
R1	High	High	High	Versicolour
R2	Medium	Medium	High	Versicolour
R3	Medium	High	High	Versicolour
R4	***Low***	***Medium***	***Medium***	***Versicolour***
R5	Low	Medium	High	Versicolour
R6	Medium	Medium	Medium	Versicolour
R7	High	Medium	High	Versicolour
R8	High	High	Medium	Versicolour
R9	Medium	High	Medium	Versicolour

In this example, R4(italicized and bolded) in Table 6.2 has the highest SR index with the factual rule and the corresponding SR index is 0.9362. Thus, the CF rule of the selected sample is R4.

Step 4: Calculating the interventions to be conducted. The x_i^{cf} calculated by Equation (6.3) of R4 in Table 6.2 are 0.2172, 0.6770 and 0.5804, respectively. Thus, using Equation (6.4), the difference between each input of the test sample and its corresponding x_i^{cf} can be calculated as -0.0791, 0.5795 and 0.5216, respectively.

Step 5: The final CF explanation generation. So far, the final CF explanation of the test sample is obtained. As explained in Section 6.2.2, the final CF explanation contains three parts: the factual part, the CF part and the CF conclusion part:

- The factual part: The test sample is *Iris Setosa*, because its *sepal length* is **low** and its *petal length* is **low** and the *petal width* is **low**.
- The CF part: To obtain *Iris Versicolour*, its *sepal length* should be **decreased by 0.0791** and its *petal length* should be **increased by 0.5795** and its *petal width* should be **increased by 0.5216**.
- The CF conclusion part: In that case, its *sepal length* would be **low** and its *petal length* would be **Medium** and its *petal width* would be **Medium**, and it would be classified *Iris Versicolour*.

This thesis exclusively focuses on conjunctive rules (i.e. using the *and* logical connective). Considering disjunctive rules employing *or*, or indeed a mix of both is possible, however increases complexity and is left to future work.

6.4 Experiments

This section compares CF-MABLAR with Cor-CF [149] to validate the effectiveness of CF-MABLAR across different data sets. This chapter selects Cor-CF as a comparison method because it represents a type of CF explanation generation method specifically for fuzzy systems.
More specifically, the key difference between Cor-CF and CF-MABLAR is that Cor-CF uses all variables within the data set to generate CF explanations, while CF-MABLAR only focuses on variables which are causally related to the target variable. Thus, the comparison with Cor-CF highlights the difference between the CF explanations generated based on correlations between variables and the CF explanations based on causal relationships between variables.

6.4.1 Experiment settings

Four real-world data sets are selected, which are widely used as benchmarks for rule generation, from the UCI data repository [122] and the Kaggle website [143]. They are Breast [122], Iris [122], Mammographic mass (MAM) [143] and

Pima Indian Diabetes (PID)[143], respectively. Table 6.3 shows the details of the selected data sets.

Table 6.3: Data sets used in this chapter

Name	Samples	Class
Breast	699	2
Iris(2 classes)	100	2
MAM	830	2
PID	768	2

To simplify the problem and reduce the computation time of the program, all selected data sets are binary classification data sets. For the Iris dataset, as mentioned in the previous section, only samples of 'Iris Setosa' and 'Iris Versicolour' were selected. To distinguish it from the original Iris data set, the Iris subset containing only 'Iris Setosa' and 'Iris Versicolour' samples is marked as Iris (2 classes) in Table 6.3.

Two fuzzy systems are trained for each data set using the settings shown in Section 6.3.1. More specifically, one Mamdani fuzzy system is trained for the causality-focused CF-MABLAR, and another for the standard Cor-CF approach. Then, a CF explanation is generated from both frameworks for each sample in a data set. Thus, each sample has two CF explanations (one from CF-MABLAR and one from Cor-CF).

As discussed in Section 6.2.2, the variant of the MABLAR framework used in CF-MABLAR should be problem-dependent. In this section, MABLAR-ST is adopted for the generation of factual explanations of CF-MABLAR for the similar reason explained in Section 6.3.1. That is, all factual explanations here are expected to focus on how to perform classification rather than exploring the direct causes of a specific class or the locally relevant causal mechanism that make a specific sample belong to a certain class. The WM algorithm is selected for rule generation of both Cor-CF and MABLAR-ST.

6.4.2 Evaluation metric

The following three indices are adopted in this chapter to evaluate CF-MABLAR and Cor-CF:

- Number of inputs (#inputs): The number of inputs of each rule. This index is used to measure the complexity of rules. A rule with more inputs can be viewed as having higher complexity, and thus lower explainability [111, 124]. Although the complexity of rules is affected by multiple factors, such as the number of fuzzy partitions of each input, structure of rules etc. [124], the number of inputs is one intuitive index. All the remaining factors are the same for both frameworks.

- Average Minimal Intervention (AMI): A better CF explanation should have a lower degree of intervention [154], which means a good CF explanation should change the inputs as little as possible. Thus, an AMI index is defined as follows:

$$AMI = \frac{\sum_{j=1}^{n} \sum_{i=1}^{d} (diff_i^j)^2}{n}, \tag{6.5}$$

  where n is the number of samples to be intervened (for example, $n = 699$ for the Breast data set) and d is the number of inputs.

- Validity: Validity is the percentage of samples which obtain a different output (e.g. an output desired by a user) after intervention [154].

6.4.3 Results

Table 6.4 shows the number of inputs, AMI and validity of CF-MABLAR and Cor-CF on different data sets.

From Table 6.4 the following can be observed:

1. The validity indices on all data sets of both approaches are 1, which means both two frameworks can obtain the desired output by the inter-

Table 6.4: Results on different data sets

Data set	#Input		AMI		Validity	
	Cor-CF	CF-MABLAR	Cor-CF	CF-MABLAR	Cor-CF	CF-MABLAR
Iris(2 classes)	4	3	1.0605	0.8238	1	1
Breast	9	6	0.8525	0.5269	1	1
MAM	5	4	0.3551	0.3851	1	1
PID	8	7	0.1519	0.1264	1	1

ventions provided by their CF explanations. As this is a low-complexity, binary classification task, this is expected.

2. Compared with Cor-CF, CF-MABLAR has a smaller number of inputs in its rules on all data sets. Thus, rules generated by CF-MABLAR have less complexity, which means a better explainability as discussed in Section 6.2.3. This behaviour is in-line with results originally presented in Section 3.5.

3. The AMI of CF-MABLAR is smaller than Cor-CF on most data sets (3 out of 4). This is because for Cor-CF intervention affects a larger number of inputs–including those which do not share a causal relationship with the output. Specifically, this also means that CF-MABLAR can avoid some redundant interventions, thereby improving the explainability of its counterfactual explanations.

   For example, as discussed in Section 6.3.2, ideally, the MB of the output variable in the Iris data set should not contain the variable 'sepal width', indicating that 'sepal width' does not have a causal relationship with the output variable. Thus, the intervention on 'sepal width' is a redundant intervention, which in turn limits the explainability of CF explanation obtained by Cor-CF, because redundant interventions increase the complexity of a CF explanation.

   Also, redundant intervention results in potentially misleading explanations, indicating to users that changes in the 'sepal width' will result in a different classification, even though this is not the case—for the causal

system.

6.5 Summary

This chapter proposes CF-MABLAR which is designed to generate CF explanations based on causal relationships between variables, complementing the factual explanations obtained by a variant of the MABLAR framework (e.g., MABLAR-ST), making them more meaningful.

More specifically, CF-MABLAR generates rules using one variant of the MABLAR framework to capture (or at least approximate) the causal relationships between inputs and output(s) for fuzzy systems – and it enables the generation of CF explanations leveraging these causal relationships. The CF explanation provided by CF-MABLAR can tell users how to intervene, i.e. how to change given input variables of a sample to obtain a desired output. Ideally, such intervention can also guide users how to obtain a desired output in the real world.

Experiment results show that, compared with existing correlation-based CF explanation generation frameworks for fuzzy rule-based systems, the CF explanation have the potential to be more concise and avoid potentially misleading (CF) explanations. Experiments in this chapter are limited to the proof-of-concept stage. In future work, larger and more complex datasets are expected to be considered, while discussing the challenge of MB approximation and its impact on the causal quality of the rule set as a whole.

Finally, note that, this chapter considers that all input variables can be changed. However, this may not happen in the real-world. For example, if an input variable is the height of an adult, it will be unchangeable. Also, it is assumed that input variables can be modified to any extent, which is also unusually in the real-world. As alluded to in the introduction, there are at least two obvious use cases for CF explanations, i.e. AI-validation and input-intervention. Future works will explore how to generate CF explanations for different purposes and

under constraints such as the immutability of inputs.

So far, all variants of MABLAR have been introduced and analysed. To promote the dissemination and application of MABLAR, as well as fuzzy systems, an open source Python library is developed to facilitate the implementation by and dissemination to third parties of the MABLAR framework. The library will be introduce in the next chapter.

Chapter 7

mablars: a python library implementing the MABLAR framework

Previous chapters have introduced different variants of the MABLAR framework. This chapter introduces a software library, which is called *mablars*, to promote the dissemination and application of the MABLAR framework, addressing RO5. *mablars* is an open-source python library which is available at https://www.lucidresearch.org/software. This chapter demonstrates the use of *mablars* and provides additional practice-based analysis in respect to the computational complexity and run-time of the different variants and different settings, providing a reference point to potential users of the MABLAR framework in their selection of the variant most appropriate to their context. The rest of this chapter is organised as follows: Section 7.1 explains the motivation of *mablars*. Section 7.2 summarises and describes the features of *mablars*. Section 7.3 shows how to use *mablars* to implement the MABLAR framework and the run-time of the different variants and different settings. Finally, Section 7.4 summaries *mablars*.

7.1 Motivation of *mablars*

As discussed in Chapter 1 and Section 2.3, with the increasing demand for XAI, fuzzy systems are receiving more attention. Nowadays, many software tools based on different programming languages have been developed to help implement different types of fuzzy systems in real-world applications. These tools enable users to either build customised fuzzy systems using their relevant prior knowledge (such as Simpful [155], Juzzy [156] and JuzzyPy [157]) or using different data-driven approaches (such as pyFUME[158] and fuzzylite [159]). However, none of these tools are designed to help users create systems rules which reflect *causal relationships* between variables using data-driven approaches.

Crucially, as discussed in the previous chapters, rules are expected to reflect causal relationships. Fundamentally, determining causal relationships between variables from a given data set is challenging and has become a research area in of itself. Here, as introduced in Section 2.2, a number of data-driven, so-called causal-discovery methods have been proposed, such as the PC algorithm and the LiNGAM algorithm. And to help people with different background to implement data-driven causal discovery approaches, many causal discovery toolboxes have been developed, such as *causal-learn* [160], *pcalg* [161] and TETRAD [162].

The previous chapters establish the MABLAR framework and propose its different variants, which are designed to generate rules that capture causal relationships between variables from a given data set. To make the MABLAR framework accessible for replication and validation by the wider community, this chapter presents *mablars*, an open-source python library to implement different variants of the MABLAR framework. *mablars* offers an accessible platform allowing researchers and practitioners to develop their own causality-focused fuzzy systems and is expected to contribute to both the adoption and further research on this critical area of leveraging fuzzy systems for XAI.

The rest of this chapter is organised as follows: Section 7.2 summarises and describes the features of *mablars*. Section 7.3 demonstrates the use of *mablars*. Finally, Section 7.4 gives conclusions.

7.2 The python library of the Markov blanket rule generation framework

7.2.1 Overview of *mablars*

The *mablars* Python project is available at https://github.com/ztcsrookie/mablars and via https://www.lucidresearch.org/software, with source code and documentation. So far, *mablars* supports using MABLAR-ST and MABLAR-CD to generate rules for solving classification problems, and the consequents of rules obtained by *mablars* are the labels used in the data sets (an example of rules obtained by *mablars* is shown in Section 7.3). In other words, *mablars* currently only supports the rule generation for fuzzy rule-based classifiers. In future updates, *mablars* will support solving a wider range of problems, such as regression problems.

The main directory of *mablars* is shown as follows:

```
mablars/
|-- dataloaders.py
|-- mablar.py
|-- mablar_cd.py
|-- membership_functions.py
```

The library is object oriented and contains four modules which are split according to their functions for easy navigation:

- The *dataloaders* module: This module contains the classes for loading and processing data from '.csv' files.

- The *mablar* module: This module contains the classes for implementing MABLAR-ST.

- The *mablab_cd* module: This module contains the classes for implementing MABLAR-CD.

- The *membership_functions* module: This module contains utilities for implementing different types of membership functions.

The design concept of *mablars* is to help users in generating rules that reflect causal relationships between variables with natural semantics, even if they are not familiar with fuzzy systems or causal discovery. Python is selected because there has been a growing trend across research communities, particularly in the field of artificial intelligence, where Python is increasingly favoured [157, 160]. In addition, the syntax of Python is concise and intuitive, making it easy for users to understand.

Fig. 7.1 shows a basic flowchart of using *mablars* for rule generation. The demo code which follows this flowchart will be presented in Section 7.3.

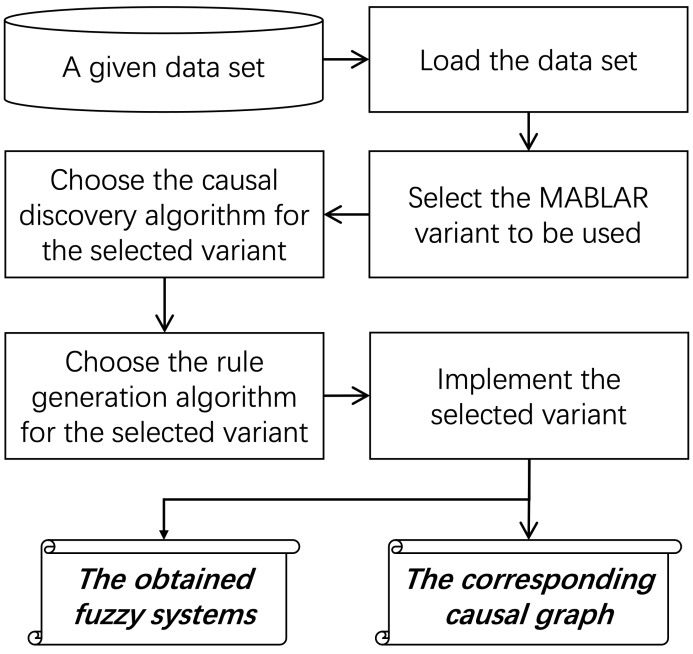


Figure 7.1: *The basic flowchart of using mablars*

The next subsection provides an overview of the features of *mablars*.

7.2.2 Features of *mablars*

mablars covers methods for users to implement data-driven causal discovery algorithms, data-driven fuzzy rule generation algorithms, as well as methods for visualising the resulting rules and causal graphs. In *mablars*, causal relationships between variables of a given data set are represented by a causal graph (which is introduced in Section 2.1).

The causal discovery of *mablars* is based on the *causal-learn* Python library [160]. Currently, *mablars* supports three causal discovery algorithms to obtain a causal graph of a given data set. They are the PC algorithm [51], the ANM algorithm [67], and the ICA-LiNGAM algorithm [55], which are introduced in Section 2.2. In some applications, the causal relationships between variables may already be known. For such cases, *mablars* also provides corresponding methods to help users manually generate the relevant causal graphs for subsequent rule generation; the details can be found in the documentation of *mablars*.

mablars provides methods for generating rules for fuzzy systems using data-driven algorithms. Currently, *mablars* supports two data-driven rule generation algorithms: 1) the WM algorithm [153] and 2) the FURIA algorithm [85]. In the future, we expect to expand the methods in *mablars* to support more rule generation algorithms.

Given a data set, after obtaining its corresponding fuzzy system using MABLAR-ST or MABLAR-CD, *mablars* provides methods to visualise the obtained causal graph and the rule base, allowing users to check whether the fuzzy system meets their expectations.

7.3 Demonstration of *mablars*

This section demonstrates the use of *mablars* step-by-step, highlighting key code sections, implementing both MABLAR-ST and MABLAR-CD based fuzzy systems. The Pima Indian Diabetes (PID) data set is selected as the ex-

ample data set, because it has a limited number of variables, which makes it easier to visualise the results. The PID data set was collected by W. Smith et al. [113] and is available at https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

7.3.1 Demonstration using the PID data set

This demo code shows how to install *mablars*, load data using *mablars*, implement MABLAR-ST or MABLAR-CD, and visualise results.

mablars can be installed by using pip:

```
(.venv) $ pip install mablars
```

The following code can be used to load data sets for implementing MABLAR-ST or MABLAR-CD:

```
from mablars.dataloaders import DataLoader


# Load the csv data from data_path
data_loader = DataLoader()
data, node_name_list = data_loader.load_csv_data(data path of
↪  the PID data set)
```

Currently, *mablars* only supports tabular data stored in CSV files. In a CSV file for implementing MABLAR-ST or MABLAR-CD, each column represents a variable, with the last column serving as the target variable or the output variable of the system. The first row of each column captures the name of the variables. The remaining rows represent the specific values of the corresponding variable for each sample, and all variable values should be numerical.

The following demo code shows how to generate rules using MABLAR-ST in *mablars*.

```
from mablars.mablar import mb
# Select MABLAR-ST and generate rules.
```

```
model = mb(node_name_list=node_name_list)
model.fit(data)
```

The following code is for implementing MABLAR-CD:

```
from mablars.mablar_cd import mbcd
# Select MABLAR-CD and generate rules.
model = mbcd(node_name_list=node_name_list)
model.fit(data)
```

As mentioned above, *mablars* provides different causal discovery algorithms and rule generation algorithms for users to implement MABLAR-ST and MABLAR-CD. This can be done by changing the parameters of the '*fit*' method. The following code shows two examples of executing MABLAR-ST and MABLAR-CD using combinations of the ICA-LiNGAM and WM, and PC and FURIA, respectively.

An example for MABLAR-ST:

```
model = mb(node_name_list=node_name_list)
model.fit(data, cd='ica',rg='wm')
```

An example for MABLAR-CD:

```
model = mbcd(node_name_list=node_name_list)
model.fit(data, cd='pc_anm',rg='furia')
```

In addition, *mablars* supports users to manually define the causal relationships between variables (For example, in the case where the users already know the causal relationships). The following code uses MABLAR-CD as an example to demonstrate how to use known causal relationships to generate rules in *mablars*.

```
# Select MABLAR-CD.
model = mbcd(causal_matrix=user_defined_causal_matrix,
↪  node_name_list=node_name_list)
```

```
# Using the FURIA algorithm for rule generation based on the
↪ pre-defined causal relationships.
model.fit(data, cd='Nothing',rg='furia')
```

In the above code, *user_defined_causal_matrix* is a $N * N$ numpy[1] matrix defined by users, where N is the number of variables in the given data set. $user_defined_causal_matrix[i, j] = 1$ represents a causal relationships between the ith variable and the jth variable, and 0 means there is no causal relationships between two variables. Then, the 'cd' parameter in the 'fit' method should be set as 'Nothing' as no causal discovery algorithm is used. The order of the variables within *user_defined_causal_matrix* should be the same as the order in the *node_name_list*.

The default setting for the '*cd*' parameter is '*pc_anm*' and the default setting for the '*rg*' parameter is '*wm*'. More details can be found in the documentation of *mablars*.

It should be noted that a number of configuration parameters in *mablars* cannot be changed at the moment without changing the source code, such as the fuzzy partitions of a variable and the types of membership functions. This simplifies the decisions users need to make when using *mablars*, making it easier for those unfamiliar with fuzzy systems or causal discovery to use the library. However, this reduces the flexibility for users who are familiar with Python when working with *mablars*. Future updates will provide those users with greater flexibility.

After generating rules using MABLAR-ST or MABLAR-CD, *mablars* provides methods to visualise the causal relationships between variables and show the obtained rule base. The following code shows how to visualise the results.

```
# Visualise the obtained causal relationships between
↪ variables
model.show_causal_graph(model.causal_dag)
```

[1] *numpy* is a Python library

```
# Visualise the obtained rule base
model.show_rule_base()
```

mablars uses causal graphs to visualise the causal relationships between variables of a given data set. Fig. 7.2 and Fig. 7.3 show the causal graph and the rule base obtained from the PID data set using MABLAR-CD with the default parameters. In Fig. 7.3, '0.0' and '1.0' are the original class labels in the PID data set and represent 'non-diabetic' and 'diabetic', respectively.

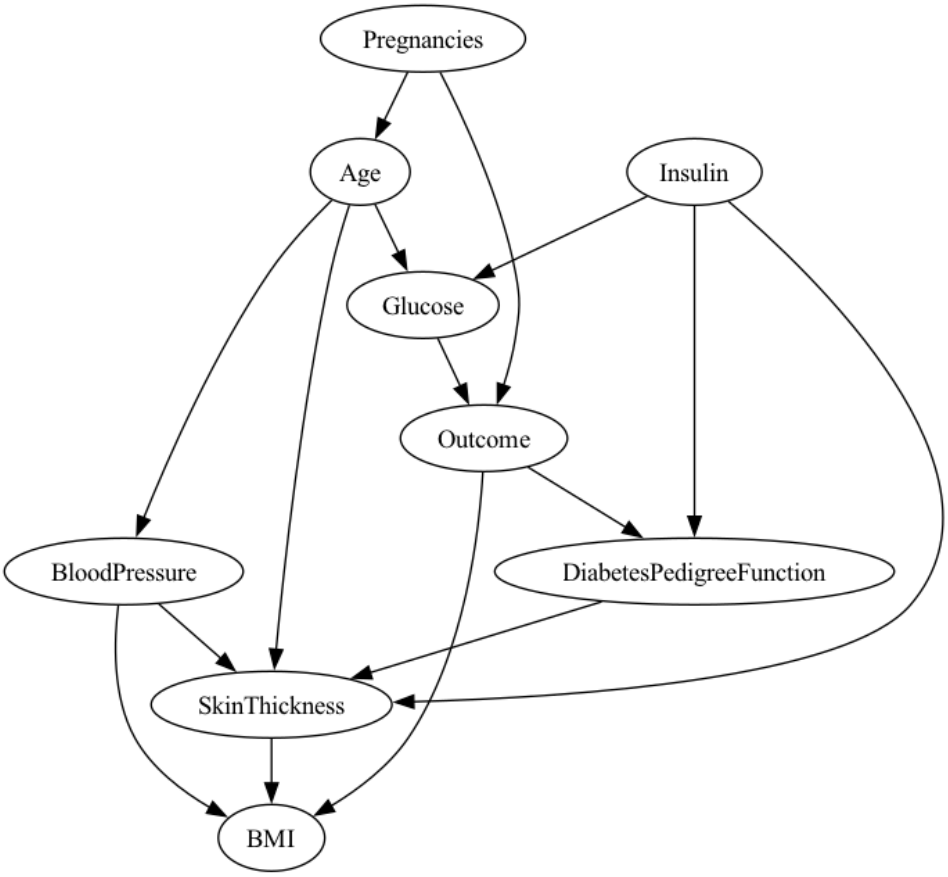


Figure 7.2: *An example of a causal graph*

The following code shows how to conduct inference using the obtained fuzzy system within *mablars*:

```
# Using the obtained fuzzy system to make prediction if the
↪  system is obtained by the WM algorithm
model.predict_wm(input)
```

```
If Pregnancies is High AND Glucose is Mid, then class is 0.0
If Pregnancies is Low AND Glucose is Low, then class is 0.0
If Pregnancies is High AND Glucose is High, then class is 1.0
If Pregnancies is Low AND Glucose is Mid, then class is 0.0
If Pregnancies is Low AND Glucose is High, then class is 1.0
If Pregnancies is Mid AND Glucose is Mid, then class is 0.0
If Pregnancies is High AND Glucose is Low, then class is 0.0
If Pregnancies is Mid AND Glucose is Low, then class is 0.0
If Pregnancies is Mid AND Glucose is High, then class is 1.0
```

Figure 7.3: *Part of the obtained rule base*

```
# Using the obtained fuzzy system to make prediction if the
↪  system is obtained by the FURIA algorithm
model.predict_furia(input)
```

7.3.2 Computational complexity in practice

While the complexity of MABLAR has been discussed in previous chapters, this section analyses the time complexity of different settings in *mablars* and provides the runtime of MABLAR-ST and MABLAR-CD under different settings using the PID dataset as an example. This information is provided as a reference to help users select the appropriate model. The experiments were executed on a 64-bit computer with 64 GB RAM and an i9-9880H 2.30GHz CPU and results were averaged over ten times.

As discussed, *mablars* currently supports two data-driven ways to achieve causal discovery:

1. Using the PC algorithm estimates the skeleton of the causal graph, and orientates the graph edges as much as possible. Then, the ANM algorithm is used to estimate the orientation of the edges that the PC algorithm could not determine the direction of.

2. Using the ICA-LiNGAM determines the directed causal graph.

For rule generation, *mablars* supports the WM algorithm and the FURIA algorithm.

Table 7.1 shows the average runtime of different settings for implementing MABLAR-ST and MABLAR-CD on the PID data set, respectively.

Table 7.1: The average running time (seconds) under different settings in the PID data set

	PC_ANM &WM	PC_ANM &FURIA	ICA-LiNGAM &WM	ICA-LiNGAM &FURIA
MB	4.9677	5.4835	0.1273	0.9181
MB	3.9297	5.2904	0.0479	0.5318

As shown in Table 7.1, the ranking from the most time consuming to the least time consuming among these four settings is as follows: PC_ANM&FURIA > PC_ANM&WM > ICA-LiNGAM&FURIA > ICA-LiNGAM&WM.

7.3.3 Impact of the causal discovery settings in practise

It should be highlighted that different settings of causal discovery may result in different approximations of the causal graphs and thus rule bases. For example, Fig. 7.4 shows the causal graph obtained by ICA-LiNGAM (i.e., the 'ICA' setting), which is different from the one obtained by the 'PC_ANM' setting shown in Fig.7.2. The differences between the graphs will result in different sets of variables used to generate rules, which in turn will lead to different rule bases.

For example, Fig. 7.5 shows the rule base obtained by MABLAR-CD using the WM algorithm based on the causal relationships shown in Fig. 7.4, which we refer to ICA-rule-base. Compared to the rule base in Fig. 7.3, which we refer to PC-rule-base, the ICA-rule-base has fewer variables in its antecedents (i.e. 1) and fewer rules. Also, the PC rule-base indicates that 'Pregnancies' and 'Glucose' are the direct causes of Diabetes, while the ICA-rule-base indicates only 'DiabetesPedigreeFunction' as a direct cause of Diabetes. These differences impact the explainability and performance of the obtained fuzzy system. Therefore, the appropriate settings should be chosen based on actual use cases.

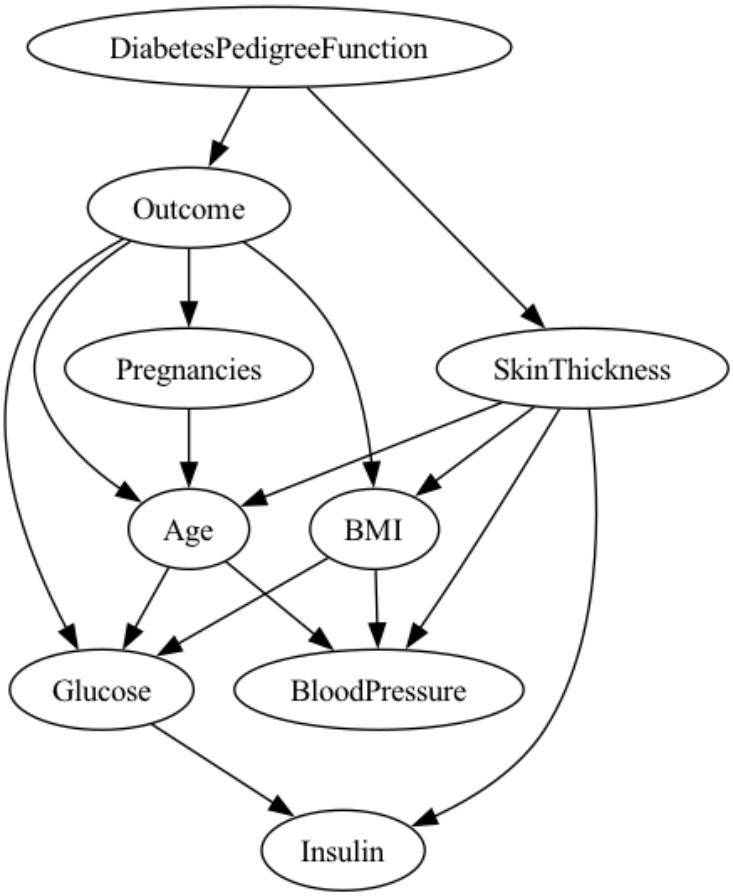


Figure 7.4: *The causal graph of the PID data set obtained by ICA-LiNGAM*

```
If DiabetesPedigreeFunction is Mid, then class is 1.0
If DiabetesPedigreeFunction is Low, then class is 0.0
If DiabetesPedigreeFunction is High, then class is 1.0
```

Figure 7.5: *The rule base obtained by MABLAR-ST using the WM algorithm based on the causal graph obtained by ICA-LiNGAM*

7.4 Summary

To make the MABLAR framework accessible for replication and validation by the wider community, this chapter introduces *mablars* which is available at https://github.com/ztcsrookie/mablars and via https://www.lucidresearch.org/software. It is an open-source Python library for automatically generating causal rules from a given data set. *mablars* provides several methods to implement causal discovery on a given data set and two rule generation algorithms. It also supports users in manually determining causal relationships between variables.

The library is actively being developed. Future updates will enable *mablars* to support more rule generation algorithms and the other two variants of the MABLAR framework, i.e., MABLAR-CW and CF-MABLAR, as well as more types of data. In addition, parameters cannot be changed at the moment without changing the source code, because *mablars* is designed with the idea of making it easily to be used by users who are not very familiar with causal discovery or fuzzy systems. For example, the natural language terms corresponding to each fuzzy set are pre-set. Future updates will give the users of *mablars* more flexibility, allowing them to more easily customise the fuzzy system to their needs. Finally, future updates will also enable *mablars* to support solving more types of problems, such as regression problems.

Chapter 8

Conclusions and future work

Fuzzy systems are often considered to have good explainability, but this is not always the case, especially when their rules are generated by data-driven approaches [163]. As discussed in Section 3.1, existing data-driven rule generation approaches have the risk of generating spurious rules, because they usually use data sets directly, exploiting correlations between variables to achieve good performance on a given data set.

Spurious rules reflect (spurious) correlations but not actual causal relationships between the variables, increasing the complexity of the resulting rule-based systems and negatively impacting performance in particular in terms of generalisation-capability and explainability of rule-based systems. As discussed in Section 1.3, rules of fuzzy systems are expected to reflect causal relationships between variables. However, it is still an open problem to automatically generate rules which reflect causal relationships between variables from a given data set.

This thesis assumes that where data sets exist, it is at least in principle possible and worthwhile to design fuzzy systems which seek to identify and model relationships between the data sets' variables which are of a causal nature. Based on this assumption, this thesis establishes a data-driven causal rule generation framework, i.e., the MABLAR framework, to generate causal rules for fuzzy systems.

The MABLAR framework leverages data-driven causal discovery algorithms to extract causal information from a given data set. The extracted causal information is used to generate rules which reflect causal relationships between variables.

As discussed in Section 2.1.2, different facets of causal relationships can be derived from a causal graph of a given data set. This thesis discusses the different roles that different facets of causal relationships play in the context of fuzzy systems, and proposes different variants of the MABLAR framework which are designed to generate rules that capture different facets of causal relationships.

This thesis shows that fuzzy systems obtained by the MABLAR framework have important benefits. More specifically, such systems (i.e., the systems obtained by the MABLAR framework) benefit from improved performance and robustness; facilitate explainability and even contribute towards one of the core ambitions for AI: the ability to communicate important relationships amongst a system's real-world variables to the human users of AI.

The rest of this chapter shows the detailed contributions of this thesis. More specifically, Section 8.1 summaries the main contributions of this thesis, highlighting how the aim and objectives have been addressed throughout the thesis. Then, Section 8.2 shows the limitations and future research directions. Finally, Section 8.3 shows the list of publications arising from this thesis.

8.1 Main contributions

8.1.1 MABLAR — a data-driven causal rule generation framework

Chapter 3 establishes the MABLAR framework. The MABLAR framework establishes how to source causal information from a given data set and leverage the obtained causal information to generate rules.

As discussed in Section 3.1, many existing data-driven rule generation approaches use data sets directly, which has the risk of generating spurious rules. Compared with rule generation approaches which uses data sets directly, the MABLAR framework uses the causal information for a causal graph of a given data set to refine the original data set in respect to the given outcomes of interest (as shown in Fig. 3.1). By doing this, the MABLAR framework reduces the risk of generating spurious rules.

The MABLAR framework extracts causal relationships among variables in a given data set by generating a causal graph using data-driven causal discovery algorithms. As discussed in Section 2.1.2, three facets of causal relationships can be derived from a causal graph: 1) *local causal relationships*, 2) *direct cause-effect relationships* and 3) *causal mechanisms*. Consequently, three variants of the MABLAR framework have been proposed: MABLAR-ST (designed for local causal relationships), MABLAR-CD (designed for direct cause-effect relationships), and MABLAR-CW (designed for causal mechanisms).

In addition, a meta-variant of the MABLAR framework, i.e., CF-MABLAR, has been proposed for generating counterfactual rules. Rules obtained by CF-MABLAR provide counterfactual explanations which can complement the explanations provided by the rules obtained by the MABLAR framework, making them more meaningful, because, as discussed in Section 6.1, human explanations are often counterfactual [147].

8.1.2 MABLAR-ST — Implementing MABLAR using the Markov blanket information

After establishing the MABLAR framework, Chapter 3 provides the first variant of the MABLAR framework: MABLAR-ST. MABLAR-ST focuses on the *local causal facet* of causal relationships derived from a causal graph of a given data set. Consequently, rules obtained by MABLAR-ST reflect the relationships between the system output variable and the input variables that are

(potentially) causally related to it.

As shown in Fig. 3.3, MABLAR-ST uses the causal information about the dependencies between variables information to refine the original data set into the MB subset of the system output variable, i.e., the target variable. Then, only variables within the MB subset of the target variable are used for rule generation. Thus, MABLAR-ST ensures that all rule antecedents are based on variables which are causally related to the target variable, which avoids generating *spurious rules*. Through this mechanism, MABLAR-ST reduces the complexity of fuzzy systems and improves the explainability of the obtained fuzzy system, while maintaining performance.

The experiment results obtained in Chapter 3 show that, compared to the correlation-based rule generation algorithms used in the experiments, the rules generated by MABLAR-ST have less (or equal) variables in their antecedents and the obtained fuzzy system contains fewer rules.

8.1.3 MABLAR-CD — Implementing MABLAR using the causal direction information

Rules obtained by MABLAR-ST focus on local causal structure of the output variable, which may conflate both the causes and effects of the output variable. Such rules may not be suitable for the cases where rules are expected to only represent the relationships between the output variable and the input variables which are its direct causes, i.e., *the 'direct cause-effect' relationships* (as pointed out in Section 4.1).

To address this issue, Chapter 4 provides MABLAR-CD, which is another variant of the MABLAR framework. MABLAR-CD focuses on the *direct cause-effect relationships* derived from a causal graph of a given data set. Thus, compare to MABLAR-ST, MABLAR-CD further narrows the MB subset to those which are direct causes by using the causal direction information obtained from a causal graph of the given data set.

In MABLAR-CD, as shown in Fig.4.2, only variables within the direct causes subset are used for rule generation, which ensures that the antecedent part of rules obtained by MABLAR-CD only contains the direct cause variables of the system output variable. Thus, MABLAR-CD helps to generate rules which reflect relationships between the system output variable and its direct cause(s).

Experiment results obtained in Chapter 4 show that the rules obtained by MABLAR-CD benefit fuzzy systems in A) providing domain-based explanations which highlight the 'direct cause-effect' relationship between variables in the antecedent and the consequent of each rule and B) improving the model-based explainability due to the reduction of both the number of rules and the complexity of the rules vis-a-vis classical data-driven rule generation methods.

8.1.4 MABLAR-CW — Implementing MABLAR using the causal weight and causal path information

Both MABLAR-ST and MABLAR-CD are designed to focus on modelling one causal mechanism of the system output variable. However, in the real-world, an effect is generally related to multiple causal mechanisms. For a given sample, as discussed in Section 5.1, it is important to distinguish the *locally relevant causal mechanism* of each specific effect from its *possible causal mechanisms*. Chapter 5 proposes MABLAR-CW, which is another variant of MABLAR, to address this issue.

MABLAR-CW uses the causal path information to construct different subsets of the original data set, thereby identifying possible causal mechanisms of the system output variable. As shown in Fig. 5.2, a fuzzy system obtained by MABLAR-CW is constructed by a set of fuzzy sub-systems, and rules of each fuzzy sub-system reflect one possible causal mechanism between the input variables and the target variable derived from a causal graph of a given data set.

For a given sample (i.e., a specific set of inputs), the fuzzy system obtained by MABLAR-CW uses the causal weight information between variables which are causally related and the rule firing strength information to distinguish the locally relevant causal mechanism of its effect from the possible causal mechanisms captured by the system, thereby avoiding both overly simplistic single-cause and exhaustive local explanations.

Experiment results obtained in Chapter 5 show that fuzzy systems obtained by MABLAR-CW provide local explanations without unnecessary components by identifying the locally relevant causal mechanism of the given sample, while maintaining comparable performance compared to classical rule generation approaches.

8.1.5 CF-MABLAR — Implementing MABLAR for CF rule generation

MABLAR-ST, MABLAR-CD and MABLAR-CW are designed to generate rules which provide factual explanations based on their captured causal relationships. Research has shown that human explanations are often counterfactual (CF), i.e. rather than explaining why a given output was reached, they show why other potential outputs were not [23]. Chapter 6 proposes a mete-variant of the MABLAR framework, i.e. CF-MABLAR, to complement an explanation provided by the rules obtained by a variant obtained by the MABLAR framework.

CF-MABLAR generates rules using a variant of the MABLAR framework (such as MABLAR-ST) to approximate the causal links between inputs and output(s) of fuzzy systems, and it generates CF explanations leveraging the approximated causal relationships. The CF explanation provided by CF-MABLAR can tell users how to intervene, i.e., how to change given input features of a sample to obtain a desired output.

Experiment results obtained in Chapter 6 show that CF-MABLAR can obtain

the desired output by the interventions provided by its CF explanations. The results also show that, compared with existing correlation-based CF explanation generation frameworks for fuzzy systems, the CF explanations obtained by MABLAR-CW are more concise and avoid potentially misleading (CF) explanations.

8.1.6 *mablars* — Implementing MABLAR in the wider Community

To promote the dissemination and application of the MABLAR framework, as well as fuzzy systems, Chapter 7 develops an open source Python library is developed to facilitate the implementation of MABLAR.

The design concept of *mablars* is to help users in implementing the MABLAR framework, even if they are not familiar with fuzzy systems or causal discovery. Python is selected because there has been a growing trend in certain communities, particularly in the field of AI, where Python is increasingly favoured.

mablars is actively being developed and has already supported using MABLAR-ST and MABLAR-CD to solve classification problems. Several methods are provided by *mablars* to implement data-driven causal discovery and rule generation.

Chapter 7 demonstrates the use of *mablars* and analyses the time consuming under different settings as a reference to help users select the appropriate algorithms for implement the MABLAR framework.

8.2 Limitations and future works

The causal information used by the MABLAR framework comes from the causal graph of the given data set. Consequently, the quality of the causal graph obtained from the given data set affects the quality of the rules generated by the MABLAR framework.

In this thesis, the MABLAR framework uses established data-driven causal discovery approaches to generate a causal graph. As discussed in Section 2.2, these causal discovery approaches assume that a given data set satisfies certain assumptions. However, in real-world applications, the data set may not necessarily satisfy the assumptions of the adopted algorithm. As a result, the obtained causal graph may not be perfect, i.e., it may not be the ground truth causal graph of the given data set.

Although experiment results obtained in this thesis show that even imperfect causal graphs, contribute to some extent in generating rules that reflect the causal relationships between the variables, future works will explore further approaches which focus on improving the quality of the rules when facing an imperfect causal graph.

Chapter 5 shows that when the causal graph of a given data set presents a large number of possible causal mechanisms, MABLAR-CW risks generating a fuzzy system that contains a large number of rules, because MABLAR-CW should use a set of flat fuzzy sub-systems, i.e., a fuzzy system consists of a single layer [89], to model all identified causal mechanisms. Thus, MABLAR-CW is suggested to be used when the task is only focusing on the the local explanations of each specific effect.

It is worth noting that a causal graph has the potential to be represented by one hierarchical fuzzy system rather than a set of fuzzy sub-systems, which may reduce the number of rules used to represent different causal mechanisms. Thus, future work will explore how to build one hierarchical fuzzy system to model potential causal mechanisms of a causal graph. In addition, future work will develop corresponding prediction mechanisms based on the captured possible causal mechanisms within the constructed hierarchical fuzzy system.

In addition, MABLAR-CW adopts a 'winner-takes-all' strategy for prediction to maintain the simplicity of the provided explanations. However, this approach may result in the neglect of a significant amount of information which could be useful for prediction, potentially reducing the performance of the

system (e.g., classification accuracy). Future research will explore alternative prediction mechanisms that combine a set of best-fitting rules with causal mechanisms and conduct corresponding experiments for validation and comparison.

In Chapter 6, CF-MABLAR considers that all input variables can be changed. However, this may not happen in the real-world. For example, if an input variable is the age of a person, it will be unchangeable. Also, CF-MABLAR assumes that input variables can be modified to any extent, which is also unusually in the real-world. Future works will explore how to generate CF explanations under constraints such as the immutability of inputs.

mablars – the developed Python library – introduced in Chapter 7 provides essential methods to implement the MABLAR framework. However, the functionality of *mablars* can be further expanded. As mentioned in Chapter 7, *mablars* is actively being developed. In the future, *mablars* will support more rule generation algorithms.

In addition, many parameters have been fixed and cannot be modified without changing the source code, because *mablars* is designed with the idea of making it easily to be used by users who are not very familiar with causal discovery or fuzzy systems. For example, the number of fuzzy sets of each fuzzy partition for each variable and the natural language terms corresponding to each fuzzy set are preset. In future updates, *mablars* will give the users more flexibility, allowing them to better customise the fuzzy system to their needs.

Finally, on a wider level, the real-world data sets used in this thesis are widely used as benchmarks for rule generation, however, these data sets currently do not exist a fully recognized ground-truth causal graph. Some causal relationships between variables in the data sets used in this thesis are supported by relevant literature and have been used in this thesis to validate the rules obtained by the MABLAR framework (see Section 3.4.5, Section 4.3.4, Section 5.3.3, and Section 6.3.2).

Future research aims to collaborate with domain experts to better validate the

causal relationships captured by the rules obtained by the MABLAR framework. It will also explore ways to integrate experts' prior knowledge with the causal relationships discovered by AI models from data, thereby enhancing human insights into the real world.

8.3 List of publications arising from this thesis

The following list of works is completed during the PhD studies (sorted by published date):

- Te Zhang, Christian Wagner, "A python library for data-driven causal fuzzy classification rule generation - *mablars*", 2025 SSCI (submitted), Trondheim, Norway.

  The content of this paper is covered in Chapter 7.

- Te Zhang, Christian Wagner and Jonathan M. Garibaldi, "Explain the world – Using causality to facilitate better rules for fuzzy systems," IEEE Transactions on Fuzzy Systems (Early access), 2024.

  The content of this paper is covered in Chapter 3 and Chapter 4.

- Te Zhang, Christian Wagner, "Generating locally relevant explanations using causal rule discovery," 2024 IEEE WCCI, Yokohama, Japan, 2024

  The content of this paper is covered in Chapter 5.

- Te Zhang, Jingda Ying, Christian Wagner and Jonathan M. Garibaldi, "Towards causal fuzzy system rules using causal direction," 2023 FUZZ-IEEE, Songdo Incheon, Korea, 2023, pp. 1-6.

  The content of this paper is covered in Chapter 4.

- Te Zhang, Christian Wagner and Jonathan M. Garibaldi, "Counterfactual rule generation for fuzzy rule-based classification systems," 2022 WCCI, Padua, Italy, 2022, pp. 1-8.

The content of this paper is covered in Chapter 6.

- Te Zhang, Christian Wagner, "Learning Causal Fuzzy Logic Rules by Leveraging Markov Blanket," 2021 IEEE SMC, Melbourne, Australia, 2021, pp. 2794-2799.

  The content of this paper is covered in Chapter 3.

Bibliography

[1] R. R. Hoffman and G. Klein, "Explaining explanation, part 1: theoretical foundations," *IEEE Intelligent Systems*, vol. 32, no. 3, pp. 68–73, 2017.

[2] J. Pearl, *Causality*. Cambridge university press, 2009.

[3] S. Kleinberg, *Causality, probability, and time*. Cambridge University Press, 2013.

[4] J. Y. Halpern, *Actual Causality*, 1st ed. The MIT Press, 2016.

[5] T. Honderich, *The Oxford companion to philosophy*. OUP Oxford, 2005.

[6] "Causality in wikipedia," https://en.wikipedia.org/wiki/Causality, accessed: 2024-6-4.

[7] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu, "Causal interpretability for machine learning - problems, methods and evaluation," *ACM SIGKDD Explorations Newsletter*, vol. 22, no. 1, p. 18–33, 2020.

[8] P. Cui and S. Athey, "Stable learning establishes some common ground between causal inference and machine learning," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 110–115, 2022.

[9] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[11] O. Bastani, C. Kim, and H. Bastani, "Interpretability via model extraction," *arXiv preprint arXiv:1706.09773*, 2017.

[12] G. Vandewiele, O. Janssens, F. Ongenae, F. De Turck, and S. Van Hoecke, "Genesim: genetic extraction of a single, interpretable model," *arXiv preprint arXiv:1611.05722*, 2016.

[13] J. M. Mendel and P. P. Bonissone, "Critical thinking about explainable AI (XAI) for rule-based fuzzy systems," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 12, pp. 3579–3593, 2021.

[14] A. Moral, C. Castiello, L. Magdalena, and C. Mencar, *Explainable fuzzy systems*. Springer, 2021.

[15] Z. Zhou, Y. Cao, C. Hu, S. Tang, C. Zhang, and J. Wang, "The interpretability of rule-based modeling approach and its development," *Acta Automatica Sinica*, vol. 47, no. 6, pp. 1201–1216, 2021.

[16] Y. You, J. Sun, Y. Guo, Y. Tan, and J. Jiang, "Interpretability and accuracy trade-off in the modeling of belief rule-based systems," *Knowledge-Based Systems*, vol. 236, p. 107491, 2022.

[17] A. S. Mamaghani and W. Pedrycz, "Structural optimization of fuzzy rule-based models: Towards efficient complexity management," *Expert Systems with Applications*, vol. 152, p. 113362, 2020.

[18] M. Negnevitsky, *Artificial intelligence: a guide to intelligent systems*. Pearson education, 2005.

[19] X. Zhu, D. Wang, W. Pedrycz, and Z. Li, "Fuzzy rule-based local surrogate models for black-box model explanation," *IEEE Transactions on Fuzzy Systems*, vol. 31, no. 6, pp. 2056–2064, 2023.

[20] S.-M. Zhou and J. Q. Gan, "Low-level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modelling," *Fuzzy Sets and Systems*, vol. 159, no. 23, pp. 3091–3131, 2008.

[21] A. Rawal, A. Raglin, D. B. Rawat, B. M. Sadler, and J. McCoy, "Causality for trustworthy artificial intelligence: Status, challenges and perspectives," *ACM Computing Surveys*, 2024.

[22] J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys*, vol. 3, no. none, pp. 96 – 146, 2009.

[23] M. Riveiro and S. Thill, ""that's (not) the output i expected!" on the role of end user expectations in creating explanations of ai systems," *Artificial Intelligence*, vol. 298, p. 103507, 2021.

[24] Aristotle, *Physics, (translated by Waterfield, Robin and Bostock, David)*. Oxford University Press, USA, 1999.

[25] ——, *The metaphysics, (translated by Lawson-Tancred, Hugh)*. Penguin UK, 1998.

[26] D. Hume, *A treatise of human nature*. Clarendon Press, 1896.

[27] J. L. Mackie, "Causes and conditions," *American philosophical quarterly*, vol. 2, no. 4, pp. 245–264, 1965.

[28] D. Lewis, "Causation," *The journal of philosophy*, vol. 70, no. 17, pp. 556–567, 1973.

[29] P. Suppes, *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Pub. Co., 1970.

[30] H. Reichenbach, *The direction of time*. Univ of California Press, 1991, vol. 65.

[31] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

[32] C. Granger, "Causality in economics," *Thinking About Causes: From Greek Philosophy to Modern Physics*, pp. 284–296, 2007.

[33] S. L. Bressler and A. K. Seth, "Wiener–granger causality: a well established methodology," *Neuroimage*, vol. 58, no. 2, pp. 323–329, 2011.

[34] M. Maziarz, "A review of the granger-causality fallacy," *The journal of philosophical economics: Reflections on economic and social issues*, vol. 8, no. 2, pp. 86–105, 2015.

[35] D. Galles and J. Pearl, "Axioms of causal relevance," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 9–43, 1997.

[36] J. Y. Halpern and J. Pearl, "Causes and explanations: a structural-model approach part 1: causes," in *The conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI, 2001, pp. 194–202.

[37] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[38] J. Pearl, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995.

[39] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Computing Surveys (CSUR)*, vol. 53, no. 4, pp. 1–37, 2020.

[40] E. C. Garrido-Merchán, C. Puente, A. Sobrino, and J. A. Olivas, "Uncertainty weighted causal graphs," *arXiv:2002.00429*, 2020.

[41] C. Puente, M. López, J. Rodrigo, and J. Olivas, "Weighted graphs to model causality," in *Proceedings on the International Conference on Artificial Intelligence (ICAI)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2017, pp. 297–301.

[42] A. Sobrino, C. Puente, and J. A. Olivas, "Extracting answers from causal mechanisms in a medical document," *Neurocomputing*, vol. 135, pp. 53–60, 2014.

[43] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation." *Journal of Machine Learning Research*, vol. 11, no. 7, pp. 171–234, 2010.

[44] J.-P. Pellet and A. Elisseeff, "Using Markov blankets for causal structure learning," *Journal of Machine Learning Research*, vol. 9, no. 7, pp. 1295–1342, 2008.

[45] K. Yu, L. Liu, and J. Li, "Learning Markov blankets from multiple interventional data sets," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 6, pp. 2005–2019, 2020.

[46] J. Pearl, "Morgan kaufmann series in representation and reasoning," *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, 1988.

[47] W. Spohn, "Direct and indirect causes," *Topoi*, vol. 9, no. 2, pp. 125–145, 1990.

[48] C. Ruichu, C. Wei, Z. Kun, and H. Zhifeng, "A survey on non-temporal series observational data based causal discovery," *Chinese Journal of Computers*, vol. 40, no. 6, pp. 1470–1490, 2017.

[49] R. Cai, Z. Zhang, and Z. Hao, "Sada: A general framework to support robust causation discovery," in *International conference on machine learning*. PMLR, 2013, pp. 208–216.

[50] H. A. Simon, "Spurious correlation: A causal interpretation," *Journal of the American statistical Association*, vol. 49, no. 267, pp. 467–479, 1954.

[51] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search*. MIT press, 2000.

[52] P. Spirtes and K. Zhang, "Causal discovery and inference: concepts and recent methodological advances," in *Applied informatics*, vol. 3. Springer, 2016, pp. 1–28.

[53] N. Weinberger, "Faithfulness, coordination and causal coincidences," *Erkenntnis*, vol. 83, no. 2, pp. 113–133, 2018.

[54] D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin, "Towards a learning theory of cause-effect inference," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1452–1461.

[55] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, "A linear non-gaussian acyclic model for causal discovery." *Journal of Machine Learning Research*, vol. 7, no. 10, 2006.

[56] T. Verma and J. Pearl, "Equivalence and synthesis of causal models," in *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, USA, 1990, p. 255–270.

[57] L. Kunitomo-Jacquin, A. Lomet, and J.-P. Poli, "Causal discovery for fuzzy rule learning," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2022, pp. 1–8.

[58] M. Kalisch and P. Bühlman, "Estimating high-dimensional directed acyclic graphs with the pc-algorithm." *Journal of Machine Learning Research*, vol. 8, no. 3, 2007.

[59] D. Colombo, M. H. Maathuis *et al.*, "Order-independent constraint-based causal structure learning." *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3741–3782, 2014.

[60] Y. He, J. Jia, and B. Yu, "Counting and exploring sizes of markov equivalence classes of directed acyclic graphs," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 2589–2609, 2015.

[61] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in genetics*, vol. 10, p. 524, 2019.

[62] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of machine learning research*, vol. 3, no. Nov, pp. 507–554, 2002.

[63] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Machine learning*, vol. 65, no. 1, pp. 31–78, 2006.

[64] I. Tsamardinos, C. F. Aliferis, and A. 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*, 2003, pp. 673–678.

[65] D. Malinsky and D. Danks, "Causal discovery algorithms: A practical guide," *Philosophy Compass*, vol. 13, no. 1, p. e12470, 2018.

[66] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen, "Directlingam: A direct method for learning a linear non-gaussian structural equation model," *Journal of Machine Learning Research*, vol. 12, pp. 1225–1248, 2011.

[67] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," *Advances in neural information processing systems*, vol. 21, 2008.

[68] T. Richardson and P. Spirtes, "Ancestral graph markov models," *The Annals of Statistics*, vol. 30, no. 4, pp. 962–1030, 2002.

[69] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson, "Learning high-dimensional directed acyclic graphs with latent and selection variables," *The Annals of Statistics*, vol. 40, no. 1, pp. 294–321, 2012.

[70] E. Kummerfeld, J. Ramsey, R. Yang, P. Spirtes, and R. Scheines, "Causal clustering for 2-factor measurement models," in *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg, 2014, pp. 34–49.

[71] U. Jaimini and A. Sheth, "Causalkg: Causal knowledge graph explainability using interventional and counterfactual reasoning," *IEEE Internet Computing*, vol. 26, no. 1, pp. 43–50, 2022.

[72] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "Xai—explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, 2019.

[73] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[74] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[75] E. H. Mamdani, "Application of fuzzy logic to approximate reasoning using linguistic synthesis," *IEEE transactions on computers*, vol. 26, no. 12, pp. 1182–1191, 1977.

[76] M. J. Gacto, R. Alcalá, and F. Herrera, "Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures," *Information Sciences*, vol. 181, no. 20, pp. 4340–4360, 2011.

[77] M. Ziatdinov, C. T. Nelson, X. Zhang, R. K. Vasudevan, E. Eliseev, A. N. Morozovska, I. Takeuchi, and S. V. Kalinin, "Causal analysis of competing atomistic mechanisms in ferroelectric materials from high-resolution scanning transmission electron microscopy data," *npj Computational Materials*, vol. 6, no. 1, p. 127, 2020.

[78] F. Oviedo, J. L. Ferres, T. Buonassisi, and K. T. Butler, "Interpretable and explainable machine learning for materials science and chemistry," *Accounts of Materials Research*, vol. 3, no. 6, pp. 597–607, 2022.

[79] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE transactions on systems, man, and cybernetics*, no. 1, pp. 116–132, 1985.

[80] J. M. Alonso, C. Castiello, and C. Mencar, "Interpretability of fuzzy systems: Current research trends and prospects," *Springer handbook of computational intelligence*, pp. 219–237, 2015.

[81] J. Jassbi, P. J. Serra, R. A. Ribeiro, and A. Donati, "A comparison of mandani and sugeno inference systems for a space fault detection application," in *2006 World Automation Congress*. IEEE, 2006, pp. 1–8.

[82] H. Jones, B. Charnomordic, D. Dubois, and S. Guillaume, "Practical inference with systems of gradual implicative rules," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 1, pp. 61–78, 2008.

[83] D. Dubois, H. Prade, and L. Ughetto, "A new perspective on reasoning with fuzzy rules," *International Journal of Intelligent Systems*, vol. 18, no. 5, pp. 541–567, 2003.

[84] L.-X. Wang and J. M. Mendel, "Generating fuzzy rules by learning from examples," *IEEE Transactions on systems, man, and cybernetics*, vol. 22, no. 6, pp. 1414–1427, 1992.

[85] J. Hühn and E. Hüllermeier, "Furia: an algorithm for unordered fuzzy rule induction," *Data Mining and Knowledge Discovery*, vol. 19, no. 3, pp. 293–319, 2009.

[86] T. R. Razak, J. M. Garibaldi, C. Wagner, A. Pourabdollah, and D. Soria, "Interpretability and complexity of design in the creation of fuzzy logic systems — a user study," in *IEEE Symposium Series on Computational Intelligence (IEEE-SSCI)*. IEEE, 2018, pp. 420–426.

[87] M. Pota, M. Esposito, and G. De Pietro, "Interpretability indexes for fuzzy classification in cognitive systems," in *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2016, pp. 24–31.

[88] J. Gou, F. Hou, W. Chen, C. Wang, and W. Luo, "Improving wang–mendel method performance in fuzzy rules generation using the fuzzy c-means clustering algorithm," *Neurocomputing*, vol. 151, pp. 1293–1304, 2015.

[89] T. R. Razak, J. M. Garibaldi, C. Wagner, A. Pourabdollah, and D. Soria, "Toward a framework for capturing interpretability of hierarchical fuzzy systems—a participatory design approach," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 5, pp. 1160–1172, 2020.

[90] G. Raju, J. Zhou, and R. A. Kisner, "Hierarchical fuzzy control," *International journal of control*, vol. 54, no. 5, pp. 1201–1216, 1991.

[91] L.-X. Wang, "The wm method completed: a flexible fuzzy system approach to data mining," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 6, pp. 768–782, 2003.

[92] T. Zhang, Y. Jingda, C. Wagner, and J. M. Garibaldi, "Towards causal fuzzy system rules using causal direction," in *202e IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2023.

[93] D. Alvarez-Estevez and V. Moret-Bonillo, "Revisiting the w ang–m endel algorithm for fuzzy classification," *Expert Systems*, vol. 35, no. 4, p. e12268, 2018.

[94] L. Jara, A. González, and R. Pérez, "Qchi: A faster classification algorithm based on wang–mendel algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 32, no. 4, pp. 2261–2271, 2024.

[95] H. Ishibuchi, K. Nozaki, and H. Tanaka, "Distributed representation of fuzzy rules and its application to pattern classification," *Fuzzy sets and systems*, vol. 52, no. 1, pp. 21–32, 1992.

[96] M. J. Gacto, M. Galende, R. Alcalá, and F. Herrera, "METSK-HDe: A multiobjective evolutionary algorithm to learn accurate tsk-fuzzy systems in high-dimensional and large-scale regression problems," *Information Sciences*, vol. 276, pp. 63–79, 2014.

[97] H. Ishibuchi and Y. Nojima, "Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning," *International Journal of Approximate Reasoning*, vol. 44, no. 1, pp. 4–31, 2007.

[98] O. Cordón and F. Herrera, "A proposal for improving the accuracy of linguistic modeling," *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 3, pp. 335–344, 2000.

[99] J. Alcala-Fdez, R. Alcala, and F. Herrera, "A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 5, pp. 857–872, 2011.

[100] B. Kavšek, N. Lavrač, and V. Jovanoski, "Apriori-sd: Adapting association rule learning to subgroup discovery," in *Advances in Intelligent Data Analysis V: 5th International Symposium on Intelligent Data Analysis, IDA 2003, Berlin, Germany, August 28-30, 2003. Proceedings 5*. Springer, 2003, pp. 230–241.

[101] L.-C. Dutu, G. Mauris, and P. Bolon, "A linear-complexity rule base generation method for fuzzy systems," in *2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT-15)*. Atlantis Press, 2015, pp. 520–527.

[102] Z. Chi, J. Wu, and H. Yan, "Handwritten numeral recognition using self-organizing maps and fuzzy rules," *Pattern Recognition*, vol. 28, no. 1, pp. 59–66, 1995.

[103] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[104] J. M. Alonso and L. Magdalena, "Hilk++: an interpretability-guided fuzzy modeling methodology for learning readable and comprehensible fuzzy rule-based classifiers," *Soft Computing*, vol. 15, pp. 1959–1980, 2011.

[105] C. Higgins and R. Goodman, "Fuzzy rule-based networks for control," *IEEE Transactions on Fuzzy Systems*, vol. 2, no. 1, pp. 82–88, 1994.

[106] A. Palacios, L. Sanchez, I. Couso, and S. Destercke, "An extension of the furia classification algorithm to low quality data through fuzzy rankings and its application to the early diagnosis of dyslexia," *Neurocomputing*, vol. 176, pp. 60–71, 2016.

[107] S. Porebski, "Evaluation of fuzzy membership functions for linguistic rule-based classifier focused on explainability, interpretability and reliability," *Expert Systems with Applications*, vol. 199, p. 117116, 2022.

[108] B. Schölkopf, "Causality for machine learning," in *Probabilistic and Causal Inference: The Works of Judea Pearl*, 2022, pp. 765–804.

[109] R. R. Hoffman and G. Klein, "Explaining explanation, part 1: theoretical foundations," *IEEE Intelligent Systems*, vol. 32, no. 3, pp. 68–73, 2017.

[110] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 5, pp. 1–46, 2021.

[111] T. Zhang and C. Wagner, "Learning causal fuzzy logic rules by leveraging markov blankets," in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2021, pp. 2794–2799.

[112] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima indians diabetes mellitus classification based on machine learning (ml) algorithms," *Neural Computing and Applications*, vol. 35, no. 22, pp. 16 157–16 173, 2023.

[113] J. W. Smith, J. E. Everhart, W. Dickson, W. C. Knowler, and R. S. Johannes, "Using the adap learning algorithm to forecast the onset of diabetes mellitus," in *Proceedings of the annual symposium on computer application in medical care*. American Medical Informatics Association, 1988, pp. 261–265.

[114] "Pima indians diabetes database," https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database, 2016, accessed: 2022-8-21.

[115] D. Cheng, J. Li, L. Liu, J. Liu, and T. D. Le, "Data-driven causal effect estimation based on graphical causal modelling: A survey," *ACM Computing Surveys*, vol. 56, no. 5, pp. 1–37, 2024.

[116] Y. Wu, Y. Ding, Y. Tanaka, and W. Zhang, "Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention," *International journal of medical sciences*, vol. 11, no. 11, pp. 1185–1200, 2014.

[117] C. D. Society, "Guidelines for the prevention and control of type 2 diabetes in china (2017 edition)," *Chinese Journal of Practical Internal Medicine*, vol. 38, no. 4, pp. 292–344, 2018.

[118] B. Fletcher, M. Gulanick, and C. Lamendola, "Risk factors for type 2 diabetes mellitus," *Journal of Cardiovascular Nursing*, vol. 16, no. 2, pp. 17–23, 2002.

[119] "Diabetes risk factors," https://www.cdc.gov/diabetes/basics/risk-factors.html, 2022, accessed: 2022-9-18.

[120] W. C. Knowler, P. H. Bennett, R. F. Hamman, and M. Miller, "Diabetes incidence and prevalence in pima indians: a 19-fold greater incidence than in rochester, minnesota," *American journal of epidemiology*, vol. 108, no. 6, pp. 497–505, 1978.

[121] W. H. Organization, *Diabetes Mellitus: Report of a WHO Study Group*. World Health Organization, 1985, vol. 727.

[122] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[123] L.-X. Wang, "The wm method completed: a flexible fuzzy system approach to data mining," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 6, pp. 768–782, 2003.

[124] T. R. Razak, J. M. Garibaldi, C. Wagner, A. Pourabdollah, and D. Soria, "Interpretability and complexity of design in the creation of fuzzy logic systems — a user study," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2018, pp. 420–426.

[125] "Python weka wrapper3," https://fracpete.github.io/python-weka-wrapper3/, 2014, accessed: 2023-6-12.

[126] Y. Zheng, B. Huang, W. Chen, J. Ramsey, M. Gong, R. Cai, S. Shimizu, P. Spirtes, and K. Zhang, "Causal-learn: Causal discovery in python," *Journal of Machine Learning Research*, vol. 25, no. 60, pp. 1–8, 2024.

[127] "causal-learn: Causal discovery in python," https://github.com/py-why/causal-learn, 2021, accessed: 2023-6-12.

[128] T.-Z. Wang, S.-J. Huang, and Z.-H. Zhou, "Towards identifying causal relation between instances and labels," in *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 2019, pp. 289–297.

[129] D. D. Lewis, "A sequential algorithm for training text classifiers: Corrigendum and additional data," *SIGIR Forum*, vol. 29, no. 2, pp. 13–19, 1995.

[130] D. D. Lewis, R. E. Schapire, J. P. Callan, and R. Papka, "Training algorithms for linear text classifiers," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 1996, p. 298–306.

[131] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[132] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the american statistical association*, vol. 32, no. 200, pp. 675–701, 1937.

[133] ——, "A comparison of alternative tests of significance for the problem of m rankings," *The annals of mathematical statistics*, vol. 11, no. 1, pp. 86–92, 1940.

[134] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, no. 1, p. 6086, 2024.

[135] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.

[136] "Pima indians diabetes background knowledge," https://github.com/cmu-phil/example-causal-datasets/tree/main/real/pima-diabetes/ground.truth, 2022, accessed: 2023-5-21.

[137] M. Młyńczak, "Data-driven causal path discovery without prior knowledge-a benchmark study," *arXiv preprint arXiv:1807.02348*, 2018.

[138] M. Lachman, "Adult development, psychology of," in *International Encyclopedia of the Social & Behavioral Sciences*, N. J. Smelser and P. B. Baltes, Eds. Oxford: Pergamon, 2001, pp. 135–139.

[139] W. E. Committee *et al.*, "Physical status: the use and interpretation of anthropometry," *World Health Organ Tech Rep Ser.*, vol. 854, pp. 312–344, 1995.

[140] J. G. Richens, C. M. Lee, and S. Johri, "Improving the accuracy of medical diagnosis with causal machine learning," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.

[141] G. Klein, "Explaining explanation, part 3: The causal landscape," *IEEE Intelligent Systems*, vol. 33, no. 2, pp. 83–88, 2018.

[142] H. Meng, C. Wagner, and I. Triguero, "Segal time series classification — stable explanations using a generative model and an adaptive weighting method for lime," *Neural Networks*, vol. 176, p. 106345, 2024.

[143] "Kaggle data sets," https://www.kaggle.com/datasets, accessed: 2023-11-25.

[144] X. Qian, Y. Xu, and Z. Hu, *Flora of China: Irisaceae*. Beijing Science press, 1985.

[145] "Mammographic mass data set," https://www.kaggle.com/datasets/overratedgman/mammographic-mass-data-set, 2016, accessed: 2023-11-25.

[146] K. A. Bertrand, R. M. Tamimi, C. G. Scott, M. R. Jensen, V. S. Pankratz, D. Visscher, A. Norman, F. Couch, J. Shepherd, B. Fan *et al.*, "Mammographic density and risk of breast cancer by age and tumor characteristics," *Breast Cancer Research*, vol. 15, no. 6, pp. 1–13, 2013.

[147] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[148] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña, "Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2020, pp. 1–8.

[149] I. Stepin, A. Catala, M. Pereira-Fariña, and J. M. Alonso, "Factual and counterfactual explanation of fuzzy information granules," in *Interpretable Artificial Intelligence: A Perspective of Granular Computing*. Springer, 2021, pp. 153–185.

[150] A. A. Freitas, "Comprehensible classification models: a position paper," *ACM SIGKDD explorations newsletter*, vol. 15, no. 1, pp. 1–10, 2014.

[151] A. Garcia-Garcia, M. Z. Reformat, and A. Mendez-Vazquez, "Similarity-based method for reduction of fuzzy rules," in *2016 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS)*, 2016, pp. 1–6.

[152] P. Jaccard, "Nouvelles recherches sur la distribution florale," *Bull. Soc. Vaud. Sci. Nat.*, vol. 44, pp. 223–270, 1908.

[153] L.-X. Wang and J. M. Mendel, "Generating fuzzy rules by learning from examples," *IEEE Transactions on systems, man, and cybernetics*, vol. 22, no. 6, pp. 1414–1427, 1992.

[154] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: A review," *arXiv preprint arXiv:2010.10596*, 2020.

[155] S. Spolaor, C. Fuchs, P. Cazzaniga, U. Kaymak, D. Besozzi, and M. S. Nobile, "Simpful: a user-friendly python library for fuzzy logic," *International Journal of Computational Intelligence Systems*, vol. 13, no. 1, pp. 1687–1698, 2020.

[156] C. Wagner, "Juzzy - a java based toolkit for type-2 fuzzy logic," in *2013 IEEE Symposium on Advances in Type-2 Fuzzy Logic Systems (T2FUZZ)*, 2013, pp. 45–52.

[157] M. S. Ahmad and C. Wagner, "Juzzypy — a python library to create type—1, interval type-2 and general type-2 fuzzy logic systems," in *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2022, pp. 735–742.

[158] C. Fuchs, S. Spolaor, M. S. Nobile, and U. Kaymak, "pyfume: a python package for fuzzy model estimation," in *IEEE International Conference on Fuzzy Systems*, 2020, pp. 1–8.

[159] J. Rada-Vilela, "fuzzylite: a fuzzy logic control library," 2017. [Online]. Available: http://www.fuzzylite.com/

[160] Y. Zheng, B. Huang, W. Chen, J. Ramsey, M. Gong, R. Cai, S. Shimizu, P. Spirtes, and K. Zhang, "Causal-learn: Causal discovery in python," *Journal of Machine Learning Research*, vol. 25, no. 60, pp. 1–8, 2024.

[161] M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann, "Causal inference using graphical models with the r package pcalg," *Journal of statistical software*, vol. 47, pp. 1–26, 2012.

[162] J. D. Ramsey, K. Zhang, M. Glymour, R. S. Romero, B. Huang, I. Ebert-Uphoff, S. Samarasinghe, E. A. Barnes, and C. Glymour, "Tetrad—a toolbox for causal discovery," in *8th international workshop on climate informatics*, 2018, pp. 1–4.

[163] Yaochu Jin, "Fuzzy modeling of high-dimensional systems: complexity reduction and interpretability improvement," *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 2, pp. 212–221, 2000.

[164] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

Appendix A

A supplementary explanation of the experimental settings in Section 4.3

For the causal discovery, the PC algorithm is implemented using the *causal-learn.search.ConstraintBased.PC.pc* function from the causal-learn Python package [126]. All parameters except the *uc_rule* parameter are set to their default values. The default *uc_rule* parameter is 0, however, the value of the *uc_rule* parameter is set to 1 so that the PC algorithm can orient as many edges as possible. For the illustrative example, i.e., the Pima Indian Diabetes (PID) data set, the *background_knowledge* parameter of the *causal-learn.search.ConstraintBased.PC* function is further adjusted using the information provided in [136] and [137] to assist the PC algorithm in discovering causal relationships between variables. The ANM algorithm is implemented using *causallearn.search.FCMBased.ANM.ANM* in the the causal-learn Python package. All parameters are set as default.
For rule generation, the *weka.classifiers.Classifier(classname=weka.classifiers.rules.FURIA)* from the python-weka-wrapper3 Python package is adopted to implement the FURIA algorithm. All parameters of FURIA are set to their default values.
The WM algorithm is implemented on the Matlab 2020b platform. The linguistic terms of the fuzzy partitions of variables are 'Low', 'Medium' and 'High', unless stated otherwise in the main paper. Trapezoidal and triangle membership functions are adopted for the WM algorithm as they facilitate explainability [13]. The parameters of fuzzy sets of each variable are estimated by the fuzzy C-means (FCM) algorithm which is implemented using the *fcm* function from the *Fuzzy Logic Toolbox* in Matlab 2020b. The iterations of the *fcm* function is set to 3000. All the remaining parameters are set to their default.
The random forest model is implemented using the *sklearn.ensemble.RandomForestClassifier* function in the scikit-learn Python package [164]. All parameters are set to their default values.

Appendix B

The execution time in Section 4.3.3

B.1 The execution time of causal discovery in Section 4.3.3

Section 4.3 adopts the PC algorithm [51] and the ANM algorithm [67] for causal discovery as two popular discovery algorithms with broad availability of open source code libraries, facilitating replication of the research. Specifically, in Section 4.3, the causal-learn Python package [126] is used to implement the ANM and the PC algorithms. The PC algorithm is adopted to determine the causal graph of each data set and the ANM algorithm is used to infer causal directions which cannot be determined by the PC algorithm. Table B.1 shows the execution time of causal discovery on each data set [1].
Table B.2 shows the execution time of different approaches on each data set. Note that times shown for MB and MBCD approaches do not include the time for causal discovery, the latter is shown in Table B.1.

B.2 The execution time (seconds) of fuzzy system rule generation approaches in Section 4.3.3

While not critical for the applications considered here, the results shown in Table 7.1 shows that the MB-WM and MBCD-WM have generally lower execution time. However, this trend is not observed in MB-FURIA and MBCD-FURIA.

[1] The execution time of each data set reported in Table B.1 is the sum of the execution times of the PC algorithm and the ANM algorithm.

Table B.1: The execution time (seconds) of causal discovery on each data set

Data sets	Execution time
Authorship	62.67
Breast	0.38
Dry_bean	25349.87
Ecoli	3.86
Glass	0.41
Haberman	0.31
HTRU2	20745.98
Iris	0.62
mammographic	21.97
Page_blocks	84.85
Penditgits	2150.06
Pima_diabetes	5.63
Sonar	9.77
Vehicle	7.47
Vowel	19.09
Waveform_5000	5122.41
Wine	0.46

Table B.2: The execution time (seconds) of fuzzy system rule generation approaches on each data set.

	WM	MB-WM	MBCD-WM	FURIA	MB-FURIA	MBCD-FURIA
Authorship	16.26	10.95	0.60	0.29	0.24	0.13
Breast	1.60	0.95	0.24	0.04	0.03	0.04
Dry_bean	126.21	4.49		6.91	1.79	
Ecoli	2.54	0.10		0.02	0.02	
Glass	0.45	0.32	0.05	0.02	0.02	0.02
Haberman	0.15	0.10		0.02	0.01	
HTRU2	26.37	22.24		1.62	1.60	
Iris	0.16	0.11		0.01	0.01	
mammographic	0.88	1.27	0.17	0.02	0.02	0.06
Page_blocks	12.34	7.44		0.38	0.29	
Penditgits	301.82	108.18		5.18	6.91	
Pima_diabetes	2.87	2.36	0.23	0.04	0.04	0.06
Sonar	4.15	0.06		0.02	0.01	
Vehicle	6.64	4.11		0.12	0.09	
Vowel	8.82	3.39		0.19	0.14	
Waveform_5000	424.17	83.75		3.64	1.61	
Wine	0.74	0.56	0.10	0.01	0.01	0.02

Appendix C

The complete rule bases obtained by WM, MB-WM and MBCD-WM in Section 4.3.4

C.1 The complete rule base obtained by WM

Table C.1: The rule base of WM for the PID data set

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPF	Age	Outcome
R1	Medium	Medium	Medium	High	Low	Overweight	Medium	Young old	Yes
R2	Low	Low	Medium	Medium	Low	Healthy	Low	Middle adulthood	No
R3	High	High	Medium	Low	Low	Healthy	Medium	Middle adulthood	Yes
R4	Low	Low	Medium	Medium	Medium	Healthy	Low	Young adulthood	No
R5	Low	Medium	Medium	High	Medium	Obese	High	Middle adulthood	Yes
R6	Medium	Medium	High	Low	Low	Healthy	Low	Middle adulthood	No
R7	Medium	Low	Medium	High	Medium	Overweight	Low	Young adulthood	No
R8	High	Medium	Low	Low	Low	Overweight	Low	Middle adulthood	No
R9	Low	High	Medium	High	High	Overweight	Low	Old old	Yes
R10	High	Medium	High	Low	Low	Under Weight	Low	Old old	Yes
R11	Medium	Medium	High	Low	Low	Overweight	Low	Middle adulthood	No
R12	High	High	High	Low	Low	Overweight	Medium	Middle adulthood	Yes
R13	High	Medium	High	Low	Low	Healthy	High	Old old	No
R14	Low	High	Medium	Medium	High	Overweight	Low	Old old	Yes
R15	Medium	High	Medium	Medium	Medium	Healthy	Medium	Old old	Yes
R16	High	Low	Low	Low	Low	Overweight	Medium	Middle adulthood	Yes
R17	Low	Medium	High	High	Medium	Obese	Medium	Middle adulthood	Yes
R18	High	Low	High	Low	Low	Overweight	Low	Middle adulthood	No
R19	Low	Low	Low	High	Low	Obese	Low	Middle adulthood	No
R20	Low	Medium	Medium	Medium	Medium	Overweight	Medium	Middle adulthood	Yes
R21	Medium	Medium	High	High	Medium	Obese	Medium	Young adulthood	No
R22	High	Low	High	Low	Low	Overweight	Low	Young old	No
R23	High	High	High	Low	Low	Obese	Low	Young old	Yes
R24	High	Medium	High	High	Low	Healthy	Low	Middle adulthood	Yes
R25	High	Medium	High	High	Medium	Overweight	Low	Old old	Yes
R26	High	Medium	Medium	Medium	Medium	Overweight	Low	Young old	Yes
R27	High	Medium	High	Low	Low	Obese	Low	Young old	Yes
R28	Low	Low	Medium	Medium	Medium	Healthy	Medium	Young adulthood	No
R29	High	Medium	High	Medium	Medium	Healthy	Low	Old old	No
R30	Medium	Medium	High	Low	Low	Overweight	Low	Young old	No
R31	Medium	Medium	High	Medium	Low	Overweight	Medium	Old old	No
R32	Medium	High	High	High	Medium	Overweight	Medium	Middle adulthood	Yes
R33	Medium	Low	Medium	Low	Low	Healthy	Low	Young adulthood	No
R34	Medium	Low	High	Low	Low	Healthy	Low	Middle adulthood	No
R35	High	Medium	High	High	Low	Healthy	Medium	Young old	No
R36	Medium	Low	Medium	High	Medium	Healthy	Medium	Middle adulthood	No
R37	High	Medium	High	Low	Low	Overweight	Low	Middle adulthood	No
R38	High	Low	High	High	Low	Overweight	Medium	Young old	Yes
R39	Low	Low	Medium	High	Low	Overweight	Medium	Young adulthood	Yes
R40	Medium	Medium	Medium	High	Medium	Overweight	High	Old old	Yes
R41	Medium	High	Medium	Medium	Low	Overweight	Low	Young adulthood	No
R42	High	Medium	High	Low	Low	Obese	Medium	Middle adulthood	No
R43	High	Low	High	Medium	Low	Healthy	Low	Young old	No
R44	High	High	High	Medium	Medium	Obese	Medium	Old old	Yes
R45	High	High	Medium	Low	Low	Healthy	Low	Young old	No
R46	Low	High	Medium	High	Low	Obese	High	Young adulthood	Yes
R47	Low	Medium	Medium	Low	Low	Overweight	Medium	Middle adulthood	No
R48	Low	Low	Medium	Medium	Low	Healthy	Medium	Young adulthood	No
R49	High	Low	Medium	High	Low	Obese	Low	Middle adulthood	Yes
R50	High	Low	Low	Low	Low	Under Weight	Low	Young adulthood	No
R51	Low	Low	High	Low	Low	Healthy	Medium	Young adulthood	No
R52	Medium	Low	Medium	Medium	Low	Healthy	Low	Middle adulthood	No
R53	High	High	High	High	Medium	Overweight	Low	Old old	Yes
R54	High	High	Medium	High	High	Overweight	Medium	Young old	No
R55	Low	Low	Medium	Low	Low	Healthy	Low	Young adulthood	No
R56	High	High	Medium	High	High	Overweight	Low	Young old	Yes

R57	Low	Low	High	High	Medium	Obese	Medium	Middle adulthood	No
R58	Low	Medium	High	Low	Low	Obese	High	Young old	No
R59	Low	Low	Medium	High	Medium	Obese	Low	Young adulthood	No
R60	Low	Low	Low	Low	Low	Under Weight	Low	Young adulthood	No
R61	High	Medium	Medium	Low	Low	Overweight	Low	Young old	No
R62	Medium	Low	Medium	Low	Low	Healthy	Medium	Middle adulthood	No
R63	Low	Medium	Medium	High	Medium	Healthy	Medium	Young adulthood	No
R64	Medium	Low	High	Medium	Low	Healthy	Low	Middle adulthood	No
R65	Low	Medium	High	Medium	Low	Overweight	Medium	Young old	Yes
R66	Low	Medium	High	Low	Low	Obese	Medium	Old old	No
R67	Low	Low	Medium	Medium	Low	Healthy	Low	Young adulthood	No
R68	Medium	Medium	High	Medium	Medium	Healthy	Low	Young adulthood	No
R69	Low	Low	Medium	Medium	Medium	Overweight	Medium	Middle adulthood	Yes
R70	Medium	Medium	Medium	High	Medium	Healthy	Low	Young adulthood	No
R71	High	Medium	High	Low	Low	Obese	Medium	Young old	Yes
R72	Medium	Medium	High	Medium	Medium	Overweight	Low	Young adulthood	No
R73	Low	Low	High	Medium	Low	Overweight	Low	Young adulthood	No
R74	Medium	Low	Medium	High	Low	Overweight	Low	Young adulthood	No
R75	Low	Medium	Low	Low	Low	Obese	Low	Young adulthood	Yes
R76	Low	Medium	Medium	Medium	Low	Healthy	Low	Young adulthood	No
R77	Medium	Medium	Medium	Medium	Low	Healthy	Low	Young adulthood	No
R78	High	Low	High	Medium	Low	Healthy	Medium	Middle adulthood	No
R79	Medium	Medium	High	Low	Low	Obese	Low	Middle adulthood	Yes
R80	Low	Medium	High	Medium	Medium	Overweight	Medium	Young adulthood	No
R81	High	Low	Medium	High	Low	Overweight	Low	Young old	No
R82	Low	Low	Medium	Medium	Low	Obese	Low	Young adulthood	No
R83	High	Medium	Medium	High	Medium	Overweight	Low	Young old	No
R84	Medium	Medium	High	Medium	Medium	Overweight	Low	Middle adulthood	No
R85	High	Low	High	High	Low	Obese	Low	Young old	Yes
R86	Medium	Medium	Medium	Low	Low	Healthy	Low	Old old	Yes
R87	Low	Medium	High	Medium	Low	Healthy	Medium	Young adulthood	No
R88	Medium	Medium	Medium	Medium	Medium	Overweight	Low	Young old	No
R89	Low	Low	Medium	Medium	Low	Overweight	Low	Young adulthood	No
R90	Medium	Low	Medium	Medium	Low	Healthy	Low	Young adulthood	No
R91	Low	Medium	High	High	Medium	Obese	Low	Middle adulthood	Yes
R92	Low	High	Medium	Low	Low	Obese	High	Middle adulthood	Yes
R93	Low	High	Medium	Low	Low	Healthy	Low	Young adulthood	No
R94	Low	Medium	High	Low	Low	Healthy	Low	Young adulthood	No
R95	Low	Low	Medium	Low	Low	Obese	Medium	Young adulthood	No
R96	Low	Medium	Medium	Medium	Medium	Healthy	Medium	Young adulthood	No
R97	Low	Low	High	Low	Low	Healthy	Low	Young adulthood	No
R98	Medium	Medium	Medium	Medium	Medium	Overweight	Low	Middle adulthood	No
R99	Medium	High	Medium	High	Medium	Overweight	Low	Young adulthood	Yes
R100	High	High	Medium	Medium	High	Overweight	Medium	Young old	Yes
R101	Low	Low	High	High	Low	Overweight	Low	Young adulthood	No
R102	Medium	Low	Medium	Low	Low	Overweight	Low	Young adulthood	No
R103	High	High	Medium	High	Medium	Overweight	Medium	Young old	Yes
R104	Medium	Medium	High	Low	Low	Overweight	Medium	Old old	Yes
R105	Medium	Low	Medium	Low	Low	Overweight	Medium	Young adulthood	No
R106	Medium	Low	High	Medium	Low	Healthy	Low	Young adulthood	No
R107	Low	High	High	High	Medium	Obese	Medium	Young adulthood	Yes
R108	Medium	Medium	Medium	High	Low	Overweight	Low	Young adulthood	No
R109	Low	Low	High	Medium	Medium	Overweight	Low	Young adulthood	No
R110	Medium	Medium	High	Low	Low	Healthy	Low	Old old	No
R111	Low	Medium	High	Low	Low	Overweight	Low	Young adulthood	No
R112	Low	Low	Low	High	Medium	Obese	Medium	Young adulthood	Yes
R113	Medium	Medium	Medium	Medium	Medium	Obese	Low	Middle adulthood	No
R114	Low	Medium	Medium	High	Medium	Overweight	Low	Young adulthood	No
R115	Low	Medium	High	Medium	Medium	Overweight	Low	Young old	Yes
R116	Low	Low	High	Low	Low	Healthy	Medium	Old old	Yes
R117	Medium	High	Medium	Medium	Medium	Overweight	Low	Middle adulthood	Yes
R118	High	Medium	Medium	Low	Low	Overweight	High	Middle adulthood	Yes
R119	Medium	High	Medium	High	Medium	Overweight	Low	Middle adulthood	Yes
R120	High	Low	High	High	Low	Overweight	Low	Young old	No
R121	Low	Medium	Medium	Medium	Medium	Overweight	Low	Middle adulthood	No
R122	Low	Low	Medium	Medium	Low	Overweight	Medium	Young adulthood	No
R123	Low	Medium	High	Low	Low	Overweight	Medium	Middle adulthood	Yes
R124	Medium	Low	Medium	Medium	High	Overweight	Low	Middle adulthood	No
R125	Medium	Low	High	Medium	Low	Obese	Low	Young old	No
R126	High	Low	Medium	Low	Low	Overweight	Low	Young old	No
R127	Low	Low	High	Medium	Low	Under Weight	Medium	Young adulthood	No
R128	Low	Low	Medium	High	Medium	Overweight	High	Middle adulthood	No
R129	Medium	Medium	High	Low	Low	Overweight	Low	Old old	No
R130	Low	Medium	High	High	Medium	Overweight	Low	Young adulthood	No
R131	Medium	Medium	Medium	Low	Low	Healthy	Low	Middle adulthood	No
R132	High	High	High	Medium	Medium	Overweight	High	Young old	Yes
R133	Low	High	High	High	High	Obese	Medium	Young adulthood	No
R134	High	High	High	High	Low	Obese	Low	Middle adulthood	Yes
R135	Low	Low	High	Medium	Low	Healthy	Low	Young adulthood	No
R136	High	High	Medium	High	Medium	Obese	Medium	Young old	Yes
R137	Medium	High	High	High	Low	Overweight	Low	Middle adulthood	No
R138	High	Low	High	High	Medium	Overweight	Low	Young old	No
R139	Low	Medium	High	High	Medium	Obese	Low	Young adulthood	Yes
R140	Medium	Low	High	Medium	Medium	Overweight	Medium	Young old	Yes
R141	Medium	Medium	Medium	Medium	Low	Overweight	Low	Young adulthood	No
R142	Medium	Medium	Medium	Low	Low	Overweight	Medium	Middle adulthood	No
R143	Medium	Medium	Medium	Low	Low	Overweight	Low	Middle adulthood	Yes
R144	Medium	Medium	High	Medium	Low	Healthy	Medium	Middle adulthood	No
R145	Medium	Low	High	Low	Low	Overweight	Low	Middle adulthood	Yes
R146	Medium	Medium	Medium	Medium	Medium	Overweight	Medium	Middle adulthood	Yes
R147	Low	Low	Low	Medium	Low	Healthy	Medium	Young adulthood	No
R148	Low	Low	Medium	High	Low	Obese	Medium	Young adulthood	No
R149	Low	Low	Medium	Medium	Low	Overweight	Low	Middle adulthood	No
R150	High	High	Medium	High	Medium	Overweight	Medium	Middle adulthood	Yes
R151	Medium	Low	High	Low	Low	Overweight	Low	Young old	No
R152	Medium	Medium	High	Low	Low	Obese	Low	Young old	No
R153	Medium	Medium	High	Low	Low	Obese	Medium	Middle adulthood	Yes
R154	Low	Medium	Medium	Medium	Medium	Overweight	Medium	Young adulthood	No
R155	Medium	Medium	High	Low	Low	Healthy	Low	Young old	No
R156	High	High	Medium	Medium	Low	Overweight	Medium	Young old	Yes

R157	High	High	Medium	High	High	Overweight	Medium	Old old	Yes
R158	Low	Medium	High	High	Low	Overweight	High	Middle adulthood	Yes
R159	High	Medium	High	High	Medium	Healthy	Medium	Middle adulthood	Yes
R160	Medium	Medium	High	High	Medium	Overweight	Low	Young adulthood	Yes
R161	Medium	Medium	Medium	Low	Low	Healthy	Low	Young adulthood	No
R162	High	High	Medium	Low	Low	Overweight	Low	Middle adulthood	No
R163	High	Medium	Low	Low	Low	Obese	Medium	Young old	Yes
R164	High	Low	Medium	Medium	Low	Healthy	Low	Young old	No
R165	Medium	High	High	High	Medium	Obese	Low	Middle adulthood	Yes
R166	Medium	Low	Medium	Medium	Low	Healthy	Medium	Young adulthood	Yes
R167	Medium	Medium	Medium	High	Medium	Overweight	Medium	Young adulthood	Yes
R168	Medium	Medium	Medium	Medium	High	Overweight	Low	Middle adulthood	Yes
R169	Low	Medium	High	Medium	Low	Overweight	Medium	Young adulthood	No
R170	Low	Medium	High	Low	Low	Obese	Low	Middle adulthood	No
R171	Low	Low	Medium	Medium	Low	Healthy	Medium	Middle adulthood	No
R172	Medium	Low	Medium	High	Medium	Overweight	Low	Old old	No
R173	High	High	High	Medium	Medium	Overweight	Medium	Old old	Yes
R174	Medium	High	High	Low	Low	Overweight	Low	Old old	Yes
R175	Low	Low	Medium	Medium	Medium	Overweight	Low	Young adulthood	No
R176	High	High	High	High	Low	Overweight	Low	Young old	Yes
R177	Low	Medium	High	High	Low	Obese	Low	Young adulthood	No
R178	High	High	High	High	Low	Overweight	Low	Old old	No
R179	Low	Medium	Medium	Medium	Medium	Obese	Low	Young adulthood	Yes
R180	High	Medium	High	High	Medium	Overweight	Low	Middle adulthood	Yes
R181	Medium	Low	High	Medium	Low	Healthy	High	Middle adulthood	Yes
R182	Medium	Medium	Medium	Low	Low	Overweight	Low	Young old	Yes
R183	Low	High	Medium	Medium	High	Overweight	High	Young adulthood	Yes
R184	Low	High	High	Low	Low	Overweight	Medium	Old old	Yes
R185	High	Medium	Low	Low	Low	Healthy	Low	Middle adulthood	No
R186	High	Medium	Medium	High	Medium	Healthy	Medium	Old old	No
R187	Low	Low	High	Low	Low	Overweight	Low	Young adulthood	No
R188	Medium	High	Medium	High	Low	Overweight	Medium	Young adulthood	Yes
R189	Medium	High	Medium	High	High	Overweight	High	Middle adulthood	No
R190	Medium	Medium	High	Low	Low	Obese	Medium	Young adulthood	Yes
R191	Medium	Medium	High	High	High	Obese	Low	Young old	Yes
R192	Low	Low	High	Medium	Low	Healthy	Medium	Young adulthood	No
R193	Medium	Low	Medium	Medium	Low	Overweight	Low	Young adulthood	No
R194	Medium	High	Medium	Low	Low	Obese	Medium	Young adulthood	Yes
R195	Low	High	High	Medium	Low	Obese	Medium	Young adulthood	Yes
R196	High	High	High	Medium	Low	Overweight	Medium	Middle adulthood	Yes
R197	Medium	Medium	Medium	Medium	Medium	Healthy	High	Middle adulthood	Yes
R198	Low	Medium	High	High	Medium	Overweight	Low	Middle adulthood	No
R199	High	High	High	Medium	Low	Overweight	High	Young old	Yes
R200	Low	High	High	High	High	Obese	Low	Young adulthood	Yes
R201	High	Medium	Medium	High	High	Overweight	Low	Middle adulthood	No
R202	Low	Medium	High	Medium	Low	Overweight	Low	Young adulthood	No
R203	Low	Low	Medium	High	Low	Overweight	Low	Young adulthood	No
R204	High	Low	Medium	Low	Medium	Healthy	Medium	Young old	Yes
R205	Low	Medium	Medium	High	Low	Overweight	Medium	Young adulthood	Yes
R206	Medium	Medium	Medium	High	Low	Overweight	Medium	Middle adulthood	No
R207	Low	High	Medium	Medium	High	Healthy	Medium	Young adulthood	No
R208	High	High	High	Medium	Medium	Overweight	High	Old old	Yes
R209	Medium	Medium	Low	Low	Low	Overweight	Medium	Young adulthood	Yes
R210	Medium	Low	Medium	High	Low	Overweight	Medium	Young adulthood	No
R211	Medium	Medium	High	Medium	Low	Overweight	Low	Old old	Yes
R212	Medium	Low	High	Medium	Low	Overweight	Medium	Young old	No
R213	Low	Medium	Low	Low	Low	Overweight	Medium	Young adulthood	Yes
R214	Low	Medium	Medium	High	Low	Obese	High	Young adulthood	No
R215	Low	Medium	Low	Low	Low	Healthy	Low	Middle adulthood	Yes
R216	High	Low	High	High	Low	Obese	High	Young old	Yes
R217	Low	Low	Medium	High	Low	Healthy	Low	Young adulthood	No
R218	High	Low	Medium	Low	Low	Overweight	Low	Old old	No
R219	High	Low	Medium	Medium	Low	Healthy	Low	Middle adulthood	Yes
R220	Medium	Medium	High	Low	Low	Healthy	Medium	Old old	No
R221	Low	Low	Medium	Low	Medium	Healthy	Medium	Young adulthood	No
R222	Low	Medium	Medium	Low	Low	Overweight	Low	Middle adulthood	Yes
R223	High	Medium	High	Medium	Medium	Overweight	Low	Young old	No
R224	High	Medium	High	Medium	Medium	Overweight	Low	Middle adulthood	No
R225	High	High	High	Low	Low	Overweight	Low	Young old	Yes
R226	Low	Low	High	Low	Low	Healthy	Low	Old old	Yes
R227	High	Medium	High	Medium	Medium	Healthy	Medium	Old old	No
R228	Medium	High	High	High	High	Obese	Medium	Middle adulthood	No
R229	Medium	Low	Medium	High	Low	Overweight	Low	Middle adulthood	No
R230	Low	Medium	High	High	Medium	Obese	High	Middle adulthood	Yes
R231	Low	Medium	Medium	High	Medium	Obese	Medium	Young adulthood	Yes
R232	Low	High	Medium	Low	Low	Healthy	Low	Old old	No
R233	Medium	High	Medium	High	Medium	Overweight	Medium	Middle adulthood	No
R234	Low	Medium	Medium	High	High	Healthy	Low	Middle adulthood	Yes
R235	High	Low	High	Medium	Medium	Overweight	Low	Young old	Yes
R236	High	Medium	Medium	Low	Low	Healthy	Medium	Old old	No
R237	Low	High	Low	Low	Low	Overweight	Medium	Middle adulthood	Yes
R238	Medium	Low	High	High	Low	Overweight	Low	Middle adulthood	No
R239	Medium	High	High	Low	Low	Healthy	Low	Middle adulthood	No
R240	High	High	Medium	Medium	Medium	Healthy	Low	Young old	Yes
R241	Low	Medium	Medium	Medium	Medium	Healthy	Low	Young adulthood	No
R242	Low	Medium	Medium	Medium	Medium	Overweight	High	Young adulthood	Yes
R243	Medium	Low	Medium	Medium	Low	Healthy	Low	Young old	No
R244	Low	Low	Medium	High	Medium	Obese	Medium	Young adulthood	No
R245	Low	High	High	Medium	Medium	Healthy	Low	Young adulthood	Yes
R246	Medium	Medium	Medium	Low	Low	Overweight	Medium	Young adulthood	No
R247	High	Medium	High	High	Low	Overweight	High	Young old	Yes
R248	Low	Medium	Medium	Medium	Medium	Overweight	Low	Young adulthood	No
R249	Medium	High	High	Low	Low	Overweight	Low	Middle adulthood	Yes
R250	Medium	Medium	Medium	High	Medium	Overweight	Low	Middle adulthood	No
R251	Medium	High	High	Low	Low	Healthy	Low	Old old	Yes
R252	Medium	Medium	Medium	Medium	Medium	Healthy	Medium	Middle adulthood	No
R253	Medium	Medium	High	Medium	Low	Overweight	Low	Young adulthood	Yes
R254	Low	Medium	Medium	Medium	Low	Healthy	Low	Middle adulthood	Yes
R255	High	High	High	High	Low	Healthy	Medium	Young old	Yes
R256	Low	Medium	High	High	Low	Overweight	Low	Young adulthood	No

R257	Low	High	Medium	Medium	Medium	Healthy	Low	Young adulthood	No
R258	Low	Medium	Medium	High	Medium	Overweight	Medium	Middle adulthood	Yes
R259	Low	Low	High	High	Medium	Obese	Low	Young adulthood	Yes
R260	High	Medium	Medium	Medium	Low	Healthy	High	Young old	No
R261	Low	High	Low	Low	Low	Obese	Low	Young old	Yes
R262	High	Low	High	Low	Low	Healthy	Low	Young old	No
R263	Low	High	High	High	Medium	Obese	Low	Young adulthood	No
R264	Low	Medium	Low	Low	Low	Overweight	Medium	Young old	No
R265	High	High	High	High	Medium	Overweight	Medium	Middle adulthood	Yes
R266	Low	Low	High	Medium	Low	Healthy	Medium	Middle adulthood	No
R267	High	Low	Medium	Low	Low	Overweight	Medium	Old old	No
R268	High	Medium	High	High	Medium	Obese	Low	Young old	No
R269	Low	Medium	Medium	Medium	Low	Healthy	Medium	Young adulthood	No
R270	Medium	Medium	Low	Low	Low	Healthy	Low	Young adulthood	No
R271	Medium	Low	High	High	Low	Obese	Low	Middle adulthood	No
R272	Medium	Low	High	Low	Low	Obese	Low	Middle adulthood	No
R273	Medium	Low	High	Medium	Low	Overweight	Low	Young old	No
R274	Medium	Low	High	Low	Low	Obese	Medium	Young adulthood	No
R275	High	Medium	Low	Medium	Low	Obese	Medium	Young old	Yes
R276	Low	High	High	High	Medium	Overweight	Medium	Middle adulthood	Yes
R277	Medium	High	Medium	High	High	Overweight	Medium	Middle adulthood	Yes
R278	Medium	High	Medium	Low	Low	Overweight	Low	Old old	No
R279	Medium	Low	High	High	Low	Obese	Low	Old old	No
R280	Medium	Medium	High	Low	Low	Obese	Medium	Old old	Yes
R281	Medium	Low	Medium	Medium	Low	Overweight	Medium	Middle adulthood	No
R282	Medium	High	High	High	High	Obese	High	Young adulthood	Yes
R283	Low	Medium	Medium	Medium	Medium	Under Weight	High	Young adulthood	No
R284	Low	Low	Medium	High	Medium	Overweight	Low	Young adulthood	No
R285	High	Medium	High	High	High	Obese	Medium	Old old	Yes
R286	Medium	High	High	Low	Low	Obese	Low	Middle adulthood	Yes
R287	Low	Low	High	High	Low	Obese	Medium	Middle adulthood	No
R288	Low	Medium	Medium	Low	Medium	Healthy	Medium	Young adulthood	No
R289	Low	Low	Medium	Medium	Low	Healthy	High	Young adulthood	No
R290	Medium	Medium	High	Medium	Low	Overweight	Medium	Middle adulthood	Yes
R291	Medium	Medium	High	Medium	Medium	Overweight	Low	Old old	Yes
R292	Low	Low	Medium	Medium	Medium	Overweight	Low	Young old	No
R293	Medium	High	High	Low	Low	Obese	Low	Young adulthood	Yes
R294	Low	Medium	Medium	Medium	High	Healthy	Low	Young adulthood	No
R295	Medium	Medium	Medium	Medium	Medium	Healthy	Low	Middle adulthood	No
R296	Medium	High	High	Low	Low	Overweight	Medium	Middle adulthood	Yes
R297	Low	Medium	Medium	Medium	Medium	Healthy	High	Young adulthood	No
R298	Medium	Low	Medium	High	Medium	Healthy	Medium	Young old	No
R299	Low	Medium	Medium	High	Low	Overweight	Low	Young adulthood	Yes
R300	Medium	High	Medium	High	Low	Overweight	Low	Young adulthood	Yes
R301	Medium	Low	Medium	Low	Low	Overweight	Low	Middle adulthood	Yes
R302	Medium	Medium	High	High	Medium	Overweight	Low	Middle adulthood	Yes
R303	High	Low	High	Medium	Low	Overweight	Low	Young old	No
R304	Medium	High	Medium	Low	Low	Overweight	Low	Young old	Yes
R305	Medium	Medium	Medium	Low	Low	Healthy	Low	Young old	Yes
R306	High	High	High	Low	Low	Healthy	High	Young old	Yes
R307	Low	High	Medium	High	High	Obese	Medium	Middle adulthood	Yes
R308	Medium	Low	High	High	Low	Overweight	Medium	Middle adulthood	No
R309	Low	Medium	High	Medium	High	Obese	High	Young adulthood	No
R310	Low	Medium	High	Medium	Low	Healthy	Low	Young adulthood	No
R311	Low	Medium	Medium	High	Medium	Overweight	Medium	Young adulthood	Yes
R312	Medium	High	High	High	High	Overweight	Low	Young adulthood	Yes
R313	Medium	Medium	High	High	Low	Obese	Medium	Middle adulthood	Yes
R314	Low	Low	Medium	Low	Low	Healthy	Medium	Young adulthood	No
R315	Low	Medium	High	High	Medium	Obese	Medium	Young adulthood	No
R316	Low	Medium	Medium	Medium	Low	Overweight	Low	Young adulthood	No
R317	High	High	High	High	Medium	Obese	Medium	Middle adulthood	Yes
R318	Medium	High	High	High	Medium	Overweight	Low	Middle adulthood	Yes
R319	Low	High	Medium	Medium	Medium	Overweight	Low	Young old	Yes
R320	Low	Low	High	Medium	Medium	Overweight	Low	Young old	Yes
R321	Low	Low	Low	Low	Low	Healthy	Low	Young adulthood	No
R322	Low	Low	High	Low	Low	Overweight	Medium	Young adulthood	No
R323	Low	Medium	High	Low	Low	Healthy	Low	Middle adulthood	No
R324	Low	Low	Medium	Low	Low	Healthy	High	Middle adulthood	No
R325	Low	Medium	Low	Low	Low	Obese	Low	Middle adulthood	Yes
R326	High	Medium	High	High	Low	Overweight	Low	Young old	No
R327	Medium	Low	High	Low	Low	Overweight	Medium	Middle adulthood	No
R328	Low	High	High	Medium	Low	Overweight	Low	Young old	Yes
R329	Medium	Medium	Medium	Medium	Medium	Overweight	Low	Young adulthood	No
R330	High	Low	Medium	Low	Low	Overweight	Medium	Middle adulthood	Yes
R331	Medium	Medium	Medium	Medium	Low	Overweight	Low	Middle adulthood	Yes
R332	Low	High	High	High	Low	Obese	High	Young adulthood	Yes
R333	Low	Low	High	High	Medium	Overweight	Low	Young adulthood	No
R334	Low	Medium	Medium	Low	Low	Healthy	Medium	Young adulthood	Yes
R335	Low	Medium	Low	Low	Low	Healthy	Medium	Old old	No
R336	Low	Low	Medium	Medium	Medium	Overweight	Medium	Young adulthood	No
R337	High	High	Medium	Medium	Low	Overweight	Low	Young old	Yes
R338	Low	Medium	Medium	Low	Low	Healthy	Medium	Old old	No
R339	High	Medium	High	High	Medium	Overweight	Medium	Old old	Yes
R340	High	Medium	High	High	Low	Healthy	Low	Old old	No
R341	High	Medium	Medium	Medium	Low	Healthy	Medium	Young old	No
R342	High	Low	Medium	High	Low	Overweight	Medium	Young old	No
R343	High	Medium	High	Low	Low	Healthy	Medium	Middle adulthood	No
R344	Low	Low	Medium	High	Medium	Overweight	Medium	Young adulthood	Yes
R345	High	Medium	Low	Low	Low	Overweight	Low	Young old	Yes
R346	Medium	High	High	High	Medium	Obese	Medium	Young adulthood	No
R347	Low	Medium	High	High	Low	Obese	Medium	Middle adulthood	No
R348	Low	Medium	Medium	Medium	Low	Obese	Low	Young adulthood	No
R349	High	Medium	High	Low	Low	Overweight	Low	Young old	No
R350	Low	Medium	High	Medium	Low	Healthy	Low	Old old	No
R351	Low	Low	High	High	Medium	Overweight	Medium	Middle adulthood	Yes
R352	High	Medium	High	Medium	Medium	Healthy	Low	Middle adulthood	No
R353	High	Medium	High	High	Low	Healthy	Low	Young old	No
R354	Medium	Medium	High	High	Low	Healthy	Low	Old old	No
R355	Medium	High	Medium	Medium	High	Overweight	Low	Middle adulthood	Yes
R356	Low	Medium	High	High	Low	Overweight	Low	Middle adulthood	No

R357	Low	Medium	Low	Low	Low	Obese	Medium	Middle adulthood	Yes
R358	Low	Medium	Medium	High	Medium	Obese	Low	Young adulthood	Yes
R359	Low	Medium	Medium	High	High	Obese	Medium	Young adulthood	No
R360	Low	High	High	High	Medium	Obese	High	Old old	No
R361	High	High	High	Low	Low	Healthy	Medium	Old old	No
R362	Low	Low	High	Medium	Low	Overweight	Low	Young old	No
R363	Medium	Medium	Medium	Medium	Medium	Healthy	High	Young old	Yes
R364	Medium	Low	Low	Low	Low	Under Weight	Low	Young adulthood	No
R365	High	High	Medium	High	Medium	Healthy	Low	Old old	Yes
R366	Medium	High	High	High	Medium	Healthy	Medium	Young old	No
R367	Medium	Low	Medium	High	Low	Obese	Medium	Young old	Yes
R368	High	Low	Medium	Medium	Low	Overweight	Medium	Young old	No
R369	Medium	Low	High	High	Low	Overweight	Low	Young old	No
R370	Low	High	High	Medium	Medium	Overweight	Low	Middle adulthood	Yes
R371	High	Medium	High	Low	Low	Healthy	Low	Old old	No
R372	High	Low	Medium	Low	Low	Healthy	Low	Old old	No
R373	Medium	Low	Medium	Medium	Medium	Healthy	Low	Young adulthood	No
R374	High	Medium	High	Low	Low	Overweight	Low	Old old	No
R375	Medium	Medium	High	Low	High	Healthy	Medium	Old old	No
R376	Medium	Medium	Low	Low	Low	Under Weight	Low	Young adulthood	No
R377	High	Medium	Medium	Low	Low	Overweight	Medium	Young old	Yes
R378	Medium	Medium	Medium	Low	Low	Overweight	Low	Young adulthood	No
R379	Low	Medium	Medium	Low	Low	Healthy	Medium	Middle adulthood	No
R380	Low	Low	High	Low	Low	Obese	Medium	Young adulthood	No
R381	Low	Low	Medium	High	Low	Obese	Medium	Middle adulthood	No
R382	Medium	Low	Low	Low	Low	Overweight	Medium	Middle adulthood	No
R383	Low	Low	Medium	Medium	Low	Overweight	High	Young adulthood	No
R384	Medium	Medium	Low	Low	Low	Overweight	Low	Young adulthood	Yes
R385	Low	Low	High	Low	Low	Overweight	Low	Young old	No
R386	Low	Low	Medium	Low	Low	Healthy	Medium	Old old	No
R387	Low	Medium	High	High	Medium	Overweight	Medium	Young adulthood	No
R388	Medium	Medium	High	High	Medium	Overweight	Medium	Middle adulthood	Yes
R389	High	Low	High	High	Medium	Obese	Medium	Young old	Yes
R390	Medium	Medium	Medium	Medium	Medium	Overweight	Medium	Young adulthood	Yes
R391	High	Low	High	High	Low	Overweight	Medium	Old old	Yes
R392	Medium	Low	High	Medium	Low	Obese	Low	Young adulthood	No
R393	Low	Low	High	Medium	Low	Overweight	Low	Middle adulthood	No
R394	High	High	High	High	Medium	Overweight	Low	Middle adulthood	Yes
R395	Medium	High	High	Medium	Medium	Obese	High	Old old	Yes
R396	Low	High	High	High	Low	Overweight	Low	Young old	No
R397	Medium	High	High	High	Low	Healthy	Medium	Middle adulthood	No
R398	Medium	Low	Medium	Medium	Medium	Overweight	Medium	Young adulthood	No
R399	Low	Low	Medium	Medium	Medium	Overweight	Low	Middle adulthood	No
R400	High	Medium	Medium	High	Medium	Healthy	Low	Middle adulthood	No
R401	Low	Low	Medium	High	Low	Overweight	Low	Middle adulthood	No
R402	High	Low	Medium	High	Low	Obese	Low	Young old	No
R403	Low	High	Medium	High	Medium	Obese	Medium	Middle adulthood	Yes
R404	Medium	Low	Medium	Medium	Low	Healthy	Medium	Middle adulthood	No
R405	Medium	Low	Medium	High	Medium	Overweight	Low	Young old	No
R406	Medium	Low	Medium	Low	Low	Overweight	Low	Young old	No
R407	Low	Medium	High	Medium	High	Overweight	Medium	Young adulthood	No
R408	Medium	Low	Medium	Medium	Medium	Healthy	Medium	Middle adulthood	No
R409	Low	Medium	High	Low	Low	Obese	Medium	Young adulthood	Yes
R410	High	Medium	Medium	Low	Low	Healthy	Low	Middle adulthood	No
R411	Low	High	Medium	High	Low	Overweight	Medium	Old old	Yes
R412	Low	High	High	High	Low	Obese	Low	Young adulthood	Yes
R413	High	Medium	High	Medium	Low	Healthy	Low	Old old	No
R414	High	Low	High	Low	Low	Obese	Low	Young old	No
R415	High	Medium	High	Medium	High	Healthy	Medium	Old old	Yes
R416	Medium	Low	Medium	Low	Low	Healthy	Low	Middle adulthood	No
R417	Medium	High	High	Medium	Medium	Overweight	High	Old old	Yes
R418	High	Medium	High	High	Low	Obese	Medium	Young old	Yes
R419	Low	Low	Medium	Medium	Medium	Healthy	High	Young adulthood	No
R420	Medium	Medium	Medium	High	Medium	Overweight	Medium	Middle adulthood	No
R421	Low	High	High	Medium	Medium	Overweight	Medium	Young adulthood	Yes
R422	Low	Low	High	Low	Low	Obese	Low	Young old	No
R423	Low	High	High	Low	Low	Overweight	Low	Young old	Yes
R424	Medium	Low	Low	Low	Low	Healthy	Low	Middle adulthood	No
R425	Low	Medium	High	High	Low	Healthy	Low	Middle adulthood	No
R426	Medium	High	Low	Low	Low	Healthy	Low	Middle adulthood	Yes
R427	Low	High	High	High	Medium	Obese	High	Young adulthood	Yes
R428	Medium	Low	Medium	Medium	Medium	Overweight	Low	Young adulthood	No
R429	Medium	High	Medium	Medium	Medium	Overweight	Medium	Middle adulthood	Yes
R430	High	High	High	High	High	Overweight	Medium	Young old	Yes
R431	Medium	Low	High	Medium	Low	Overweight	Medium	Young adulthood	No
R432	High	Medium	High	Medium	Medium	Overweight	Medium	Young old	Yes
R433	High	Medium	High	Medium	Low	Healthy	High	Young old	Yes
R434	Low	Medium	Low	Low	Low	Overweight	Low	Young adulthood	Yes
R435	Low	Low	High	Medium	Low	Healthy	High	Middle adulthood	No
R436	Medium	High	High	Low	Low	Obese	High	Young old	No
R437	Low	Low	Medium	Medium	Medium	Obese	Low	Young adulthood	No
R438	Low	Low	Medium	Low	Low	Overweight	Low	Young adulthood	No
R439	Low	Medium	Medium	Low	Low	Healthy	Low	Young adulthood	No
R440	High	Medium	Medium	Low	Low	Healthy	Medium	Middle adulthood	Yes
R441	Low	Medium	High	Medium	Medium	Healthy	Low	Young adulthood	No
R442	High	Low	Medium	Low	Low	Healthy	Low	Middle adulthood	No
R443	Medium	Low	High	Low	Low	Healthy	Low	Young old	No
R444	Low	Low	High	Medium	Low	Overweight	Medium	Young adulthood	No
R445	High	Low	High	High	Medium	Obese	Medium	Middle adulthood	Yes
R446	Low	Low	High	Medium	Low	Healthy	Low	Middle adulthood	No
R447	Low	Low	High	Medium	Medium	Healthy	Medium	Young adulthood	No
R448	Medium	Low	Low	Low	Low	Healthy	Medium	Middle adulthood	No
R449	Medium	Low	Medium	Medium	Medium	Healthy	Medium	Young adulthood	No
R450	Low	High	High	High	High	Obese	Low	Middle adulthood	No
R451	Low	High	High	Medium	Medium	Healthy	Low	Middle adulthood	Yes
R452	Low	High	Medium	High	Medium	Overweight	Low	Young adulthood	Yes
R453	High	Medium	High	High	Medium	Healthy	Low	Young old	Yes
R454	Medium	Medium	High	High	Low	Overweight	Low	Middle adulthood	No
R455	Low	High	Medium	Medium	High	Obese	Low	Young adulthood	Yes
R456	Low	Low	Medium	High	Medium	Healthy	Low	Young adulthood	No

R457	Low	Medium	High	High	Medium	Obese	High	Young old	No
R458	High	Medium	High	Low	Low	Obese	Low	Old old	No
R459	Medium	Low	High	High	Low	Overweight	High	Young adulthood	Yes
R460	High	High	High	Low	Low	Healthy	Low	Old old	Yes
R461	High	High	High	High	Medium	Overweight	Low	Young old	Yes
R462	High	Medium	High	High	Medium	Overweight	Medium	Young old	Yes
R463	Medium	Medium	Medium	High	Low	Overweight	Low	Young old	Yes
R464	High	Medium	Medium	Medium	Low	Healthy	Low	Young old	Yes
R465	High	High	High	Medium	Medium	Overweight	Low	Young old	No
R466	Medium	High	Medium	Medium	Medium	Overweight	Medium	Young old	No
R467	High	Low	High	Low	Low	Overweight	Medium	Old old	No
R468	Medium	High	Medium	Low	Low	Overweight	Low	Middle adulthood	Yes
R469	Medium	Medium	High	Low	Low	Overweight	Medium	Young adulthood	Yes
R470	Medium	Medium	High	Low	Low	Under Weight	Medium	Old old	No
R471	Low	Medium	High	Medium	Medium	Healthy	Medium	Young adulthood	No
R472	Low	Medium	High	High	Medium	Obese	Low	Young old	Yes
R473	High	Low	High	Low	Low	Healthy	Medium	Middle adulthood	No
R474	High	Medium	Medium	High	Medium	Obese	Low	Young old	Yes
R475	High	Medium	High	Medium	High	Overweight	Low	Young old	Yes
R476	Medium	High	High	Medium	Medium	Overweight	Low	Middle adulthood	Yes
R477	Medium	Medium	High	Low	Medium	Overweight	Medium	Middle adulthood	No
R478	Medium	Medium	Medium	Low	Low	Obese	Medium	Young adulthood	No
R479	Medium	Medium	High	High	Low	Healthy	Medium	Young old	Yes
R480	Low	High	High	Medium	Low	Overweight	Medium	Old old	Yes
R481	Low	Medium	Low	Low	Low	Obese	Low	Young old	No
R482	High	Medium	Low	Low	Low	Under Weight	Low	Middle adulthood	Yes
R483	Low	Medium	Medium	Medium	High	Overweight	Low	Young adulthood	No
R484	Medium	High	Medium	Medium	High	Overweight	Low	Young adulthood	No
R485	Medium	Medium	High	Medium	Low	Overweight	Low	Young old	No
R486	High	Medium	Medium	High	Low	Obese	Low	Young old	Yes
R487	High	High	Medium	High	High	Overweight	Medium	Middle adulthood	Yes
R488	High	Low	Medium	Medium	Low	Healthy	Medium	Old old	No
R489	Medium	Low	High	Medium	Low	Overweight	Low	Old old	Yes
R490	Low	High	Medium	Medium	Medium	Healthy	Low	Young old	Yes
R491	Medium	Medium	High	Medium	Medium	Obese	Low	Young old	No
R492	Low	Medium	High	Low	Low	Overweight	Low	Young old	No
R493	Medium	Medium	High	High	Low	Obese	Low	Young old	No
R494	Low	High	High	Low	Low	Healthy	Low	Young adulthood	No
R495	Medium	Medium	High	Medium	Low	Healthy	Low	Middle adulthood	Yes
R496	High	Medium	High	Low	Low	Healthy	Low	Young adulthood	Yes
R497	High	Medium	High	High	Medium	Obese	Medium	Young old	Yes
R498	High	Medium	High	Low	Low	Overweight	Medium	Young old	Yes
R499	High	High	High	High	Medium	Obese	High	Young old	No
R500	High	Low	High	High	Medium	Overweight	Medium	Young old	No
R501	Low	Low	High	High	Low	Obese	High	Middle adulthood	No
R502	Medium	High	Medium	Low	Low	Healthy	Low	Young old	Yes
R503	Medium	Medium	Medium	Low	Low	Overweight	High	Young adulthood	Yes
R504	Low	Medium	High	High	Low	Obese	Low	Middle adulthood	No
R505	Low	Medium	High	High	Medium	Overweight	High	Middle adulthood	Yes
R506	Low	Medium	Medium	Low	Low	Overweight	Low	Old old	Yes
R507	High	High	High	High	Low	Obese	Low	Young old	Yes
R508	High	Low	High	High	Medium	Overweight	Low	Old old	No
R509	Low	Medium	Medium	Low	Low	Overweight	Low	Young old	Yes

C.2 The complete rules obtained by MB-WM

Table C.2: The rule base of MB-WM for the PID data set

	Pregnancies	Glucose	BloodPressure	SkinThickness	BMI	DPF	Outcome
R1	Medium	Medium	Medium	High	Overweight	Medium	No
R2	Low	Low	Medium	Medium	Healthy	Low	No
R3	High	High	Medium	Low	Healthy	Medium	Yes
R4	Low	Medium	Medium	High	Obese	High	Yes
R5	Medium	Medium	High	Low	Healthy	Low	No
R6	Medium	Low	Medium	High	Overweight	Low	No
R7	High	Medium	Low	Low	Overweight	Low	No
R8	Low	High	Medium	High	Overweight	Low	Yes
R9	High	Medium	High	Low	Under Weight	Low	Yes
R10	Medium	Medium	High	Low	Overweight	Low	No
R11	High	High	High	Low	Overweight	Medium	Yes
R12	High	Medium	High	Low	Healthy	High	No
R13	Low	High	Medium	Medium	Overweight	Low	Yes
R14	Medium	High	Medium	Medium	Healthy	Medium	Yes
R15	High	Low	Low	Low	Overweight	Medium	Yes
R16	Low	Medium	High	High	Obese	Medium	Yes
R17	High	Low	High	Low	Overweight	Low	No
R18	Low	Low	Low	High	Obese	Low	No
R19	Low	Medium	Medium	Medium	Overweight	Medium	No
R20	Medium	Medium	High	High	Obese	Medium	No
R21	High	High	High	Low	Obese	Low	Yes
R22	High	Medium	High	High	Healthy	Low	Yes
R23	High	Medium	High	High	Overweight	Low	Yes
R24	High	Medium	Medium	Medium	Overweight	Low	Yes
R25	High	Medium	High	Low	Obese	Low	No
R26	Low	Low	Medium	Medium	Healthy	Medium	No
R27	High	Medium	High	Medium	Healthy	Low	No
R28	Medium	Medium	High	Medium	Overweight	Medium	Yes
R29	Medium	High	High	High	Overweight	Medium	Yes
R30	Medium	Low	Medium	Low	Healthy	Low	No
R31	Medium	Low	High	Low	Healthy	Low	No
R32	High	Medium	High	High	Healthy	Medium	Yes
R33	Medium	Low	Medium	High	Healthy	Medium	No
R34	High	Medium	High	Low	Overweight	Low	No
R35	High	Low	High	High	Overweight	Medium	Yes
R36	Low	Low	Medium	High	Overweight	Medium	No
R37	Medium	Medium	Medium	High	Overweight	High	Yes
R38	Medium	High	Medium	Medium	Overweight	Low	No
R39	High	Medium	High	Low	Obese	Medium	Yes
R40	High	Low	High	Medium	Healthy	Low	No
R41	High	High	High	Medium	Obese	Medium	Yes
R42	High	High	Medium	Low	Healthy	Low	No
R43	Low	High	Medium	High	Obese	High	Yes
R44	Low	Medium	Medium	Low	Overweight	Medium	No
R45	High	Low	Medium	High	Obese	Low	No
R46	High	Low	Low	Low	Under Weight	Low	No
R47	Low	Low	High	Low	Healthy	Medium	Yes
R48	Medium	Low	Medium	Medium	Healthy	Low	No
R49	High	High	High	High	Overweight	Low	Yes
R50	High	High	Medium	High	Overweight	Medium	Yes
R51	Low	Low	Medium	Low	Healthy	Low	No
R52	High	High	Medium	High	Overweight	Low	Yes
R53	Low	Low	High	High	Obese	Medium	No
R54	Low	Medium	High	Low	Obese	High	No
R55	Low	Low	Medium	High	Obese	Low	No
R56	Low	Low	Low	Low	Under Weight	Low	No
R57	High	Medium	Medium	Low	Overweight	Low	No
R58	Medium	Low	Medium	Low	Healthy	Medium	No
R59	Low	Medium	Medium	High	Healthy	Medium	No
R60	Medium	Low	High	Medium	Healthy	Low	No
R61	Low	Medium	High	Medium	Overweight	Medium	No
R62	Low	Medium	High	Low	Obese	Medium	Yes
R63	Medium	Medium	High	Medium	Healthy	Low	Yes
R64	Low	Low	Medium	Medium	Overweight	Medium	No
R65	Medium	Medium	Medium	High	Healthy	Low	No
R66	Medium	Medium	High	Medium	Overweight	Low	No
R67	Low	Low	High	Medium	Overweight	Low	Yes
R68	Low	Medium	Low	Low	Obese	Low	Yes
R69	Low	Medium	Medium	Medium	Healthy	Low	No
R70	Medium	Medium	Medium	Medium	Healthy	Low	No
R71	High	Low	High	Medium	Healthy	Medium	No
R72	Medium	Medium	High	Low	Obese	Low	Yes
R73	High	Low	Medium	High	Overweight	Low	No
R74	Low	Low	Medium	Medium	Obese	Low	No
R75	High	Medium	Medium	High	Overweight	Low	No
R76	High	Low	High	High	Obese	Low	No
R77	Medium	Medium	Medium	Low	Healthy	Low	No
R78	Low	Medium	High	Medium	Healthy	Medium	No
R79	Medium	Medium	Medium	Medium	Overweight	Low	No
R80	Low	Low	Medium	Medium	Overweight	Low	No
R81	Low	Medium	High	High	Obese	Low	Yes
R82	Low	High	Medium	Low	Obese	High	Yes
R83	Low	High	Medium	Low	Healthy	Low	No
R84	Low	Medium	High	Low	Healthy	Low	No
R85	Low	Low	Medium	Low	Obese	Medium	No
R86	Low	Medium	Medium	Medium	Healthy	Medium	No
R87	Low	Low	High	Low	Healthy	Low	No
R88	Medium	High	Medium	High	Overweight	Low	Yes
R89	High	High	Medium	Medium	Overweight	Medium	Yes
R90	Low	Low	High	High	Overweight	Low	No
R91	Medium	Low	Medium	Low	Overweight	Low	Yes
R92	Medium	Medium	High	Low	Overweight	Medium	Yes

R93	Medium	Low	Medium	Low	Overweight	Medium	No
R94	Low	High	High	High	Obese	Medium	Yes
R95	Medium	Medium	Medium	High	Overweight	Low	Yes
R96	Low	Medium	High	Low	Overweight	Low	No
R97	Low	Low	Low	High	Obese	Medium	Yes
R98	Medium	Medium	Medium	Medium	Obese	Low	No
R99	Low	Medium	Medium	High	Overweight	Low	Yes
R100	Low	Medium	High	Medium	Overweight	Low	Yes
R101	High	Medium	Medium	Low	Overweight	High	Yes
R102	High	Low	High	High	Overweight	Low	No
R103	Low	Medium	Medium	Medium	Overweight	Low	No
R104	Low	Medium	High	Low	Overweight	Medium	Yes
R105	Medium	Low	Medium	Medium	Overweight	Low	No
R106	Medium	Low	High	Medium	Obese	Low	No
R107	High	Low	Medium	Low	Overweight	Low	No
R108	Low	Low	High	Medium	Under Weight	Medium	No
R109	Low	Low	Medium	High	Overweight	High	No
R110	Low	Medium	High	High	Overweight	Low	No
R111	High	High	High	Medium	Overweight	High	Yes
R112	High	High	High	High	Obese	Low	Yes
R113	Low	Low	High	Medium	Healthy	Low	No
R114	High	High	Medium	High	Obese	Medium	Yes
R115	Medium	High	High	High	Overweight	Low	Yes
R116	Medium	Low	High	Medium	Overweight	Medium	No
R117	Medium	Medium	Medium	Low	Overweight	Medium	No
R118	Medium	Medium	Medium	Low	Overweight	Low	Yes
R119	Medium	Medium	High	Medium	Healthy	Medium	No
R120	Medium	Low	High	Low	Overweight	Low	Yes
R121	Medium	Medium	Medium	Medium	Overweight	Medium	Yes
R122	Low	Low	Low	Medium	Healthy	Medium	No
R123	Low	Low	Medium	High	Obese	Medium	No
R124	Medium	Medium	High	Low	Obese	Medium	Yes
R125	Low	Medium	High	High	Overweight	High	Yes
R126	Medium	Medium	High	High	Overweight	Low	Yes
R127	High	High	Medium	Low	Overweight	Low	No
R128	High	Medium	Low	Low	Obese	Medium	Yes
R129	High	Low	Medium	Medium	Healthy	Low	No
R130	Medium	High	High	High	Obese	Low	Yes
R131	Medium	Low	Medium	Medium	Healthy	Medium	No
R132	Low	Medium	High	Low	Obese	Low	No
R133	High	High	High	Medium	Overweight	Medium	Yes
R134	Medium	High	High	Low	Overweight	Low	Yes
R135	Low	Medium	Medium	Medium	Obese	Low	Yes
R136	Medium	Low	High	Medium	Healthy	High	Yes
R137	Low	High	Medium	Medium	Overweight	High	Yes
R138	Low	High	High	Low	Overweight	Medium	Yes
R139	High	Medium	Low	Low	Healthy	Low	No
R140	High	Medium	Medium	High	Healthy	Medium	No
R141	Low	Low	High	Low	Overweight	Low	No
R142	Medium	High	Medium	High	Overweight	Medium	Yes
R143	Medium	High	Medium	High	Overweight	High	No
R144	Medium	Medium	High	High	Obese	Low	Yes
R145	Low	Low	High	Medium	Healthy	Medium	No
R146	Medium	High	Medium	Low	Obese	Medium	Yes
R147	Low	High	High	Medium	Obese	Medium	Yes
R148	Medium	Medium	Medium	Medium	Healthy	High	Yes
R149	Low	High	High	High	Obese	Low	Yes
R150	Low	Low	Medium	High	Overweight	Low	No
R151	High	Low	Medium	Low	Healthy	Medium	Yes
R152	Low	Medium	Medium	High	Overweight	Medium	Yes
R153	Low	High	Medium	Medium	Healthy	Medium	No
R154	Medium	Medium	Low	Low	Overweight	Medium	Yes
R155	Medium	Low	Medium	High	Overweight	Medium	No
R156	Low	Medium	Low	Low	Overweight	Medium	No
R157	Low	Medium	Low	Low	Healthy	Low	Yes
R158	High	Low	High	High	Obese	High	Yes
R159	Low	Low	Medium	High	Healthy	Low	No
R160	Medium	Medium	High	Low	Healthy	Medium	No
R161	Low	Low	Medium	Low	Healthy	Medium	No
R162	Low	Medium	Medium	Low	Overweight	Low	Yes
R163	High	Medium	High	Medium	Overweight	Low	No
R164	High	High	High	Low	Overweight	Low	Yes
R165	High	Medium	High	Medium	Healthy	Medium	Yes
R166	Medium	High	High	High	Obese	Medium	No
R167	Low	Medium	High	High	Obese	High	Yes
R168	Low	Medium	Medium	High	Obese	Medium	Yes
R169	Low	Medium	Medium	High	Healthy	Low	Yes
R170	High	Low	High	Medium	Overweight	Low	No
R171	High	Medium	Medium	Low	Healthy	Medium	Yes
R172	Low	High	Low	Low	Overweight	Medium	Yes
R173	Medium	Low	High	High	Overweight	Low	No
R174	Medium	High	High	Low	Healthy	Low	Yes
R175	High	High	Medium	Medium	Healthy	Low	Yes
R176	Low	Medium	Medium	Medium	Overweight	High	Yes
R177	Low	High	High	Medium	Healthy	Low	Yes
R178	High	Medium	High	High	Overweight	High	Yes
R179	Medium	Medium	Medium	Medium	Healthy	Medium	No
R180	High	High	High	High	Healthy	Medium	Yes
R181	Low	High	Medium	Medium	Healthy	Low	No
R182	Low	Low	High	High	Obese	Low	Yes
R183	High	Medium	Medium	Medium	Healthy	High	No
R184	Low	High	Low	Low	Obese	Low	Yes
R185	High	Low	High	Low	Healthy	Low	No
R186	High	High	High	High	Overweight	Medium	Yes
R187	High	Low	Medium	Low	Overweight	Medium	No
R188	High	Medium	High	High	Obese	Low	No
R189	Medium	Medium	Low	Low	Healthy	Low	No
R190	Medium	Low	High	High	Obese	Low	No
R191	Medium	Low	High	Low	Obese	Low	No
R192	Medium	Low	High	Medium	Overweight	Low	No

R193	Medium	Low	High	Low	Obese	Medium	No
R194	High	Medium	Low	Medium	Obese	Medium	Yes
R195	Low	High	High	High	Overweight	Medium	Yes
R196	Medium	High	Medium	Low	Overweight	Low	Yes
R197	Medium	Low	Medium	Medium	Overweight	Medium	No
R198	Medium	High	High	High	Obese	High	Yes
R199	Low	Medium	Medium	Medium	Under Weight	High	No
R200	High	Medium	High	High	Obese	Medium	Yes
R201	Medium	High	High	Low	Obese	Low	Yes
R202	Low	Medium	Medium	Low	Healthy	Medium	No
R203	Low	Low	Medium	Medium	Healthy	High	No
R204	Medium	High	High	Low	Overweight	Medium	Yes
R205	Low	Medium	Medium	Medium	Healthy	High	No
R206	High	High	High	Low	Healthy	High	Yes
R207	Low	High	Medium	High	Obese	Medium	Yes
R208	Medium	Low	High	High	Overweight	Medium	No
R209	Low	Medium	High	Medium	Obese	High	No
R210	Low	Medium	High	Medium	Healthy	Low	No
R211	High	High	High	High	Obese	Medium	Yes
R212	Low	Low	Low	Low	Healthy	Low	No
R213	Low	Low	High	Low	Overweight	Medium	No
R214	Low	Low	Medium	Low	Healthy	High	No
R215	Medium	Low	High	Low	Overweight	Medium	No
R216	Low	High	High	Medium	Overweight	Low	Yes
R217	Low	High	High	High	Obese	High	Yes
R218	Low	Medium	Low	Low	Healthy	Medium	No
R219	High	High	Medium	Medium	Overweight	Low	Yes
R220	High	Medium	High	High	Overweight	Medium	Yes
R221	High	Medium	Medium	Medium	Healthy	Medium	No
R222	High	Low	Medium	High	Overweight	Medium	No
R223	High	Medium	High	Low	Healthy	Medium	No
R224	Low	Low	High	High	Overweight	Medium	Yes
R225	Medium	Medium	High	High	Healthy	Low	No
R226	Low	Medium	Low	Low	Obese	Medium	Yes
R227	Low	Medium	Medium	High	Obese	Low	Yes
R228	High	High	High	Low	Healthy	Medium	No
R229	Medium	Low	Low	Low	Under Weight	Low	No
R230	High	High	Medium	High	Healthy	Low	Yes
R231	Medium	High	High	High	Healthy	Medium	No
R232	Medium	Low	Medium	High	Obese	Medium	Yes
R233	High	Low	Medium	Medium	Overweight	Medium	No
R234	High	Medium	High	Low	Healthy	Low	Yes
R235	High	Low	Medium	Low	Healthy	Low	No
R236	Medium	Medium	Low	Low	Under Weight	Low	No
R237	High	Medium	Medium	Low	Overweight	Medium	Yes
R238	Low	Low	High	Low	Obese	Medium	No
R239	Medium	Low	Low	Low	Overweight	Medium	No
R240	Low	Low	Medium	Medium	Overweight	High	No
R241	Medium	Medium	Low	Low	Overweight	Low	Yes
R242	Low	Medium	High	High	Overweight	Medium	No
R243	Medium	Medium	High	High	Overweight	Medium	Yes
R244	High	Low	High	High	Obese	Medium	Yes
R245	Medium	High	High	Medium	Obese	High	Yes
R246	Low	High	High	High	Overweight	Low	No
R247	High	Medium	Medium	High	Healthy	Low	No
R248	High	Medium	Medium	Low	Healthy	Low	No
R249	Low	High	Medium	High	Overweight	Medium	Yes
R250	High	Low	High	Low	Obese	Low	No
R251	Medium	High	High	Medium	Overweight	High	Yes
R252	Low	High	High	Medium	Overweight	Medium	Yes
R253	Low	Low	High	Low	Obese	Low	No
R254	Low	High	High	Low	Overweight	Low	Yes
R255	Medium	Low	Low	Low	Healthy	Low	No
R256	Low	Medium	High	High	Healthy	Low	No
R257	Medium	High	Low	Low	Healthy	Low	Yes
R258	Medium	High	Medium	Medium	Overweight	Medium	Yes
R259	High	Medium	High	Medium	Overweight	Medium	Yes
R260	High	Medium	High	Medium	Healthy	High	Yes
R261	Low	Medium	Low	Low	Overweight	Low	Yes
R262	Low	Low	High	Medium	Healthy	High	No
R263	Medium	High	High	Low	Obese	High	No
R264	Low	Low	Medium	Low	Overweight	Low	No
R265	Low	Medium	Medium	Low	Healthy	Low	No
R266	Low	Low	High	Medium	Overweight	Medium	No
R267	Medium	Low	Low	Low	Healthy	Medium	No
R268	Low	High	Medium	Medium	Obese	Low	Yes
R269	Medium	Low	High	High	Overweight	High	Yes
R270	High	High	High	Low	Healthy	Low	Yes
R271	High	Medium	Medium	Medium	Healthy	Low	Yes
R272	High	High	High	Medium	Overweight	Low	No
R273	High	Low	High	Low	Overweight	Medium	No
R274	Medium	Medium	High	Low	Under Weight	Medium	No
R275	High	Low	High	Low	Healthy	Medium	No
R276	High	Medium	Medium	High	Obese	Low	Yes
R277	Medium	High	High	Medium	Overweight	Low	Yes
R278	Medium	Medium	Medium	Low	Obese	Medium	No
R279	Medium	Medium	High	High	Healthy	Medium	Yes
R280	High	Medium	Low	Low	Under Weight	Low	Yes
R281	High	Low	Medium	Medium	Healthy	Medium	No
R282	Medium	Medium	High	Medium	Obese	Low	No
R283	Low	High	High	Low	Healthy	Low	No
R284	High	Medium	High	Low	Overweight	Medium	Yes
R285	High	High	High	High	Obese	High	No
R286	Low	Low	High	High	Obese	High	No
R287	Medium	High	Medium	Low	Healthy	Low	Yes
R288	Medium	Medium	Medium	Low	Overweight	High	Yes

C.3 The complete rules obtained by MBCD-WM

Table C.3: The rule base of the PID data set of MB-CDWM

	Pregnancies	Glucose	DPF	Outcome
R1	High	High	Medium	Yes
R2	Low	Medium	High	Yes
R3	High	Medium	Low	Yes
R4	Low	High	Low	Yes
R5	Medium	High	Medium	Yes
R6	Medium	High	Low	Yes
R7	Low	High	Medium	Yes
R8	High	Medium	Medium	Yes
R9	High	Low	Medium	Yes
R10	Medium	Medium	High	Yes
R11	Low	High	High	Yes
R12	High	High	High	Yes
R13	Medium	Low	High	Yes
R14	High	Low	High	Yes
R15	Medium	Medium	Medium	No
R16	Low	Low	Low	No
R17	Medium	Medium	Low	No
R18	High	Medium	High	No
R19	Medium	Low	Medium	No
R20	Low	Medium	Medium	No
R21	Medium	Low	Low	No
R22	High	Low	Low	No
R23	Low	Low	Medium	No
R24	High	High	Low	No
R25	Low	Medium	Low	No
R26	Low	Low	High	No
R27	Medium	High	High	No

Appendix D

The complete rule bases obtained by FURIA, MB-FURIA and MBCD-FURIA in Section 4.3.4

D.1 The complete rule base obtained by FURIA

Table D.1: The rule base of FURIA for the PID data set

	IF	THEN
R1	Glusoce is [0.7739, 0.7789, inf, inf]	Outcome is Yes. (CF = 0.8)
R2	Age is [0.1500, 0.1667, inf, inf] AND Glucose is [0.6131, 0.6181, inf, inf] AND BMI is [0.4471, 0.4486, inf, inf]	Outcome is Yes. (CF = 0.75)
R3	Glucose is [-inf, -inf, 0.6181, 0.6432] AND Age is [-inf, -inf, 0.1167, 0.1333] AND BMI is [-inf, -inf, 0.4605, 0.4620]	Outcome is No. (CF = 0.98)
R4	Glucose is [-inf, -inf, 0.5025, 0.5075]	Outcome is No. (CF = 0.9)
R5	Glucose is [0.6281, 0.6332, 0.7186, 0.7337] AND BMI is [-inf, -inf, 0.4292, 0.4307]	Outcome is No. (CF = 0.87)
R6	Glucose is [-inf, -inf, 0.7186, 0.7638] AND BMI is [-inf, -inf, 0.4069, 0.4083]	Outcome is No. (CF = 0.95)
R7	Glucose is [-inf, -inf, 0.5528, 0.5578] AND DPF is [-inf, -inf, 0.0739, 0.0751]	Outcome is No. (CF = 0.92)
R8	Glucose is [-inf, -inf, 0.6482, 0.6583] AND Age is [-inf, -inf, 0.1500, 0.1667]	Outcome is No. (CF = 0.88)
R9	Glucose is [-inf, -inf, 0.7739, 0.7789] AND Age is [-inf, -inf, 0.1500, 0.1667] AND Blood Pressure is [0, 0.6066, inf, inf]	Outcome=No. (CF = 0.86)

D.2 The complete rule base obtained by MB-FURIA

Table D.2: The rule base of MB-FURIA for the PID data set

	IF	THEN
R1	Glucose is [0.798995, 0.809045, inf, inf]	Outcome is Yes. (CF = 0.81)
R2	Glucose is [0.638191, 0.643216, inf, inf] AND BMI is [0.445604, 0.451565, inf, inf]	Outcome is Yes. (CF = 0.72)
R3	Glucose is [-inf, -inf, 0.61809, 0.623116]	Outcome is No. (CF = 0.82)
R4	BMI is [-inf, -inf, 0.447094, 0.448584] and Glucose is [0.628141, 0.633166, 0.728643, 0.733668]	Outcome is No. (CF = 0.84)
R5	Glucose is [-inf, -inf, 0.798995, 0.809045] and DPF is [-inf, -inf, 0.107173, 0.109308]	Outcome is No. (CF = 0.79)

D.3 The complete rule base obtained by MBCD-FURIA

Table D.3: The rule base of MBCD-FURIA for the PID data set

	IF	THEN
R1	Glucose is [0.7739, 0.7789, inf, inf]	Outcome is Yes. (CF = 0.81)
R2	Glucose is [0.6382, 0.6432, inf, inf] AND BMI is [0.4456, 0.4516, inf, inf]	Outcome is Yes. (CF = 0.72)
R3	Glucose is [-inf, -inf, 0.6181, 0.6231]	Outcome is No. (CF = 0.82)