Expert Systems With Applications 136 (2019) 187–200

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Causality measures and analysis: A rough set framework

Ning Yao[a,b], Duoqian Miao[a,b,*], Witold Pedrycz[a,c,*], Hongyun Zhang[a,b], Zhifei Zhang[a,b]

[a] Department of Computer Science and Technology, Tongji University, Shanghai 201804, PR China
[b] Key Laboratory of Embedded System and Service Computing, Ministry of Education, Shanghai 201804, PR China
[c] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2G7, Canada

ARTICLE INFO

Article history:
Received 21 August 2018
Revised 1 June 2019
Accepted 2 June 2019
Available online 3 June 2019

Keywords:
Rough set
Lower approximation
Causal effect
Intervention
Counterfactual

ABSTRACT

Data and rules power expert systems and intelligent systems. Rules, as a form of knowledge representation, can be acquired by experts or learned from data. The accuracy and precision of knowledge largely determines the success of the systems, which awakens the concern for causality. The ability to elicit cause–effect rules directly from data is key and difficult to any expert systems and intelligent systems. Rough set theory has succeeded in automatically transforming data into knowledge, where data are often presented as an attribute-value table. However, the existing tools in this theory are currently incapable of interpreting counterfactuals and interventions involved in causal analysis. This paper offers an attempt to characterize the cause–effect relationships between attributes in attribute-value tables with intent to overcome existing limitations. First, we establish the main conditions that attributes need to satisfy in order to estimate the causal effects between them, by employing the back-door criterion and the adjustment formula for a directed acyclic graph. In particular, based on the notion of lower approximation, we extend the back-door criterion to an original data table without any graphical structures. We then identify the effects of the interventions and the counterfactual interpretation of causation between attributes in such tables. Through illustrative studies completed for some attribute-value tables, we show the procedure for identifying the causation between attributes and examine whether the dependency of the attributes can describe causality between them.

1. Introduction

The theory of rough sets aims to tackle imperfect knowledge by making minimal model assumption (Düntsch & Gediga, 2001; Pawlak, 1991). Rough set-based data analysis starts with an attribute-value table or an information system and utilizes only knowledge derived from data. Rough set theory not only can convert data into (available) knowledge but can deal with knowledge representation, attribute reduction and reasoning about knowledge (e.g., Ciucci, Chiaselotti, Gentile, & Infusino, 2016; Lingras & Haider, 2015; Qian, Miao, Zhang, & Yue, 2014; Wang, Zhao, Zhao, & Han, 2003; Wu, Qian, Li, & Gu, 2017; Yao & Azam, 2015; Yao, 2011; Zhang & Miao, 2016; Zhou & Miao, 2011; Ziarko, 2008). Rough sets consider that vagueness is associated with a boundary region of a set, and objects characterized by the same information are indiscernible (similar) in view of the available information about them (Pawlak & Skowron, 2007). Indiscernibility describes our lack of knowledge about objects, which means that elements of a set are indiscernible by employing their features for any attributes in a feature set, and an indiscernibility relation usually stands for an equivalence relation (Pawlak, 2004; Peters & Skowron, 2006). Equivalence classes implied by the indiscernibility relation represent knowledge that can be characterized by employing this relation. Based on the interior and closure of equivalence classes, any vague concept can be replaced by a pair of precise concepts called the lower and the upper approximation of the vague concept (Pawlak & Skowron, 2007). The notion of the lower approximation not only can be used to characterize the dependencies of knowledge but can be used for rule-based reasoning under uncertainty (Pawlak, 1991; Yao, Miao, Zhang, & Lang, 2016). Dependency, defined by lower approximation, often captures the state that another attribute can be derived from a given attribute and then discover *if-then* rules, trying to describe causal relationships hidden in data. The degree of dependency can be viewed as a counterpart of correlation coefficient used in statistics (Pawlak, 2004). Consider that causality in causal analysis measures the capacity of one variable to entail another variable by formalizing interventions and

* Corresponding authors.
E-mail addresses: ningyao.cn@gmail.com (N. Yao), dqmiao@tongji.edu.cn (D. Miao), wpedrycz@ualberta.ca (W. Pedrycz), zhanghongyun@tongji.edu.cn (H. Zhang), zhifeizhang@tongji.edu.cn (Z. Zhang).

https://doi.org/10.1016/j.eswa.2019.06.004
0957-4174/

counterfactuals. However, dependency and other existing tools, defined in rough set theory, cannot formalize and interpret interventions and counterfactuals. Causality can provide more valuable information for judgments in daily life than dependency.

The mathematical formation of causality was introduced by Pearl (2009). With the aid of graphical models and structural equations, the concepts involved in the expression of causal relations have been formalized (Maathuis & Colombo, 2015; Pearl, 2009; 2010; Spirtes, Glymour, & Scheines, 1993). For example, associate the physical intervention mathematically with the "*do*" operator and the corresponding manipulation of the figure by deleting the incoming arrows of the causal variable and keeping the rest unchanged; encode the counterfactuals into "probability of necessity" (how necessary the cause is for the formation of the effect) and "probability of sufficiency" (how sufficient a cause is for the production of the effect) for causation; present the back-door criterion and the front-door criterion to identify the effects of interventions from non-experimental data without bias (Pearl, 2009; Tian & Pearl, 2000). In terms of causality, causal relationships describe objective physical constraints in our world and remain invariant to changes (i.e., external interventions) in the mechanism that governs the causal variables, instead of estimating the results of passive observations which reflect what we know or believe about the world (e.g., probabilistic relationships given in the standard language of statistics) (Pearl, 2009). The nonparametric structural equation modeling (Bareinboim & Pearl, 2016; Bareinboim & Tian, 2015; Pearl, 2009; 2010; Tian & Pearl, 2000), established by Pearl and his colleagues, has become a part of the mainstream of causal analysis realized so far, embracing the potential-outcome framework of Neyman and Rubin in its algebraic component and Wright's method of path diagrams in its graphical component (Neyman, 1923; Pearl, 2010; 2015; Rubin, 1974; Wright, 1921). The structural causal model framework undeniably provides a friendly mathematical machinery for a general cause–effect analysis, and as a result enables causality to be computationally manageable rather than left to be determined by intuition and good judgment.

For expert systems and machine intelligence, knowledge represented in the form of *if-then* rules is an essential ingredient and knowledge acquisition has always been a challenging issue. Counterfactuals and interventions are viewed as the building blocks of causal knowledge and are essential to scientific thinking as well as legal and moral reasoning. Rough sets excel at an automated change of data into knowledge using only the notions of finite set, equivalence relation and cardinality, without any preliminary or additional information about the data. Unfortunately, the ability to answer questions about interventions and counterfactuals, for *if-then* rules or relationships discovered in data by rough sets, is currently rarely studied and still under development. This paper explores the causation of attributes within the rough set framework using the ideas of the nonparametric structural equation modeling. Consider that rough set theory starts with an attribute-value data table without any data structures, but structural causal models assume that the causal structure represented by a directed acyclic graph is known. Methodologically, by applying the back-door criterion to the attribute-value tables, we try to identify the effects of interventions and the probabilities of counterfactuals between attributes in the given data tables based on the following insights. First, the back-door criterion involves graphical structures and statistical probabilities. These two factors can be obtained directly from data using rough set tools, such as certainty factor and dependency of knowledge. Second, the covariates for adjustment from the graphical structures embedded in the back-door criterion refer to a set of variables in most cases. In rough set theory, the combination of different attributes can be viewed as a new single attribute variable, which means we can construct the structures, embedded in the back-door criterion with the covariates as a single variable, by rough set tools when dealing with an attribute-value table. By doing so, we can get the unbiased effect estimate of interventions, through extending the conditions for back-door adjustment by the lower approximations of the attributes and addressing the problem of under which conditions the interventional distribution is estimable for such table. Moreover, regarding the adjustment formula used in the back-door criterion, we yield the conditions for estimating the effects of interventions with confounding bias when the back-door criterion does not hold. The diagrams (a) and (b) below show the basic principle of structural causal modeling and the original idea of this study, respectively.

$$\textit{Graph} \longrightarrow \textit{Graphical criteria} \longrightarrow \textit{Variables for adjustment} \xrightarrow{\textit{Data}} \textit{Causal effects}$$

(a) The principle of structural causal model

$$\left.\begin{array}{l}\textit{Attribute–value table}\\ \textit{Graphical criteria}\end{array}\right\} \xrightarrow{\textit{Rough sets}} \textit{Graph strucures} \longrightarrow \textit{Variables for adjustment} \xrightarrow{\textit{Data}} \textit{Causal effects}$$

(b) The combination of rough sets and structural causal model

From an application point of view, the only prerequisite of our approach is that the data within a domain of interest can be collected in the form of attribute-value tables from non-experimental studies, simultaneously without any missing values and with all the attributes in the table having nominal values. Our theoretical results will then be feasible to discover the cause–effect relationships in data, i.e., interventional interpretation and counterfactual analysis. As to the data tables collected from experimental studies, the data in these tables illustrate the effects of interventions and can be directly used to estimate the counterfactuals of interest. The systems built based on the proposed results can enable to automatically capture the causal knowledge hidden in the attribute-value data tables and then improve our daily lives in many ways that causality acts on, such as medical diagnosis, crime analysis, decision making, psychotherapy, reasoning about the facts or new things, and understanding the complex phenomena.

The paper is organized as follows. Section 2 offers an overview of related work. In Section 3, the background to the related theory of rough sets and structural causal models is briefly outlined. Section 4 focuses on the relations of the lower approximation in the rough set framework as well as graphical criteria in the structural causal model framework, and then establishes the criteria or the scheme of discovering causality between attributes under the intervention for attribute-value tables with no graphical structures given. Based on the causal effects of the intervention identified in the above section, Section 5 concerns the identification of causation defined by the counterfactuals, examines the interpretation of causation based on the specific data tables under some constraints, and illustrates the use of the established results as well as the discovered patterns in data. The summary and some directions of our future work are covered in Section 6.

2. Related work

Causality notion in the rough set framework as well as the application of structural causal model in rough set-based data analysis remains rarely studied despite its importance.

In rough set theory, only a few works have addressed the attribute dependency based on observations (e.g., Pawlak, 1991; Raza & Qamar, 2018; Yamaguchi, 2009; Ziarko, 2007; 2008) with the purpose of attribute reduction and rule extraction. Attribute dependency measures how another knowledge can be induced from a given knowledge. Here some representative measures of attribute dependency are presented. Pawlak (1991) proposed the dependency and partial dependency of attributes using the notion of lower approximation of a set and an indiscernibility relation, however, Pawlak's definition of attribute dependency excludes any misclassification errors. To allow for some degree of misclassification, the probabilistic generalization of the Pawlak's measure was introduced as approximate dependency of attributes based on the notion of β-lower approximation of a set in Ziarko (1993) and on the notions of u-lower approximation and l-negative region of a set in Ziarko (2007), where β, u, l represent precision control parameters, together with probabilistic dependency measure between attributes based on the notion of the expected gain function in Ziarko (2007, 2008). These models require parameters and a selected target set, but how to define the parameters and how to determine a proper target set are not clear. Yamaguchi (2009) considered data efficiency in the computation of the dependency degree and developed a new dependency measure based on decision-relative discernibility matrices instead of the indiscernibility relation to overcome the inadequacy of the existing measures in computing dependency degrees. Raza and Qamar (2018) introduced a heuristics dependency measure by summation of all consistent records for each decision class to enhance the computational efficiency, which can be viewed as another generalization of Pawlak's attribute dependency. Tran, Arch-Int, and Arch-Int (2018) developed differential dependency based on differential-relation-based rough sets from the perspectives of relational databases and information systems, concluding that differential dependency in relational databases corresponds to attribute dependency on the differential decision systems. Despite all these existing models, the representation and the handling of interventions and counterfactuals between attributes have not been involved or closer to being solved in rough set theory.

In Pearl's structural theory of causation, the notion of the "*do*" operator representing interventions and the back-door criterion play a crucial role. Benferhat and Smaoui (2007) provided the counterpart of the "*do*" operator in possibility theory framework and Benferhat (2010) described interventions in possibilistic networks as a belief revision process. Boukhris, Elouedi, and Benferhat (2013) generalized the "*do*" operator to the belief function framework to deal with imperfect situations of ignorance. As Pearl's back-door criterion assumes a directed acyclic graph is given, Maathuis and Colombo (2015) generalized this back-door criterion to more general types of graphs such as the completed partially directed acyclic graph, i.e., Markov equivalence classes of the directed acyclic graph, the maximal ancestral graph on the observed variables, and the partial ancestral graph, i.e., Markov equivalence classes of the maximal ancestral graph. Concerning the estimation of causal effects, the existing work requires some assumptions (Maathuis & Colombo, 2015) on causal structures, causal relationships between the variables, or a set of variables that can be used for covariate adjustment. In the current paper, we start from a given attribute-value table with nominal-valued data, making no such assumptions, and for the real-valued data, some form of discretization is needed first. This is also the advantage of rough set theory.

3. Preliminaries

In this section, we present the basic notions and theory of structural causal modeling and rough sets (for more details the reader is referred to Pawlak, 1991; 2004; Pawlak & Skowron, 2007; Pearl, 2009; 2010; Peters & Skowron, 2006; Spirtes et al., 1993; Tian & Pearl, 2000).

3.1. Rough set theory

In rough set theory, a concept is represented as a set of objects relative to their attribute values, and the equivalence relation, as the simplest form of the indiscernibility relation, is widely used. By $|\cdot|$ denote the cardinality of a set, i.e., the number of elements of a set.

Definition 1 (Equivalence class (Pawlak, 1991))**.** Let $U \neq \emptyset$ be a finite set called the universe and whose elements are called objects. Let R be an equivalence relation over U characterized by the attributes of the objects. We denote by $[x]_R = \{y \in U : xRy\}$ the equivalence class of an object $x \in U$ with respect to R and by U/R the set partition on U induced by R, i.e., the family of all equivalence classes of R. The equivalence class denotes a concept in R, while the set partition U/R means knowledge associated with R, in short knowledge R.

Based on the attributes of the objects and the inclusion relation, one can give an approximate description of any subset of objects.

Definition 2 (Lower and upper approximation (Pawlak, 1991; 2004))**.** Let X be a subset of U, i.e., $X \subseteq U$ and let $x \in U$. The characterization of the set X with respect to R can be given by R-lower and R-upper approximations of X:

$$\underline{R}(X) = \left\{x \in U : [x]_R \subseteq X\right\}, \ \overline{R}(X) = \left\{x \in U : [x]_R \cap X \neq \emptyset\right\}.$$

The lower approximation of the set X is the set of all objects that are surely included in X according to R. The upper approximation of X is the set of objects that are possibly included in X due to R. The difference between the upper and the lower approximation of X, called the boundary region of X, is the set of objects which cannot be included uniquely in X or its complement due to R. A set $X \subseteq U$ is called exact with respect to R if the boundary region of X is empty, rough with respect to R otherwise.

Suppose we are given a pair $S = (U, \boldsymbol{A})$, where U is the universe of objects and $\boldsymbol{A}$ denotes a set of attributes, call this pair an information system or an attribute-value table, i.e., a data table, columns of which are labeled by attributes, rows by objects of interest and entries of the table are attribute values; if we distinguish two classes of attributes, called condition attributes C and decision attributes D, denote the system by $S = (U, C, D)$, then call this system a decision table, each row of which determines a decision rule (Pawlak, 2004). With every decision rule a certainty factor of the rule is associated:

Definition 3 (Certainty factor (Pawlak, 2004))**.** Let $x \in U$ and $[x]_C$ denote the set of all objects satisfying C in the system S, i.e., the equivalence class of C determined by element x. For every x, the sequence $[x]_C$, $[x]_D$ will be called a decision rule induced by x in S and denoted by $[x]_C \rightarrow [x]_D$ or in short $C \rightarrow_x D$. The certainty factor of the decision rule $C \rightarrow_x D$, denoted by $cer_x(C, D)$, is

$$cer_x(C, D) = \frac{|\ [x]_C \cap [x]_D\ |}{|\ [x]_C\ |}.$$

In general, we omit subscript x for simplicity if x is understood. The notion of the certainty factor can also be interpreted as a conditional probability that an object x satisfies the decision, provided it satisfies the condition of the rule, which means one can estimate the conditional probability straight from the data according to $P(D|C) = \frac{P(C,D)}{P(C)} = \frac{|[x]_C \cap [x]_D|}{|[x]_C|} = cer(C, D)$. The unconditional probability that objects of U satisfy C is defined as $P(C) = \frac{|[x]_C|}{|U|}$.

The dependency between attributes can be understood by the lower approximation in the following way:

Definition 4 (Dependency of knowledge (Pawlak, 1991)). Let **P** and **Q** be two different equivalence relations or two different sets of equivalence relations over U. Knowledge **Q** means the partition of U determined by **Q**. Concepts of knowledge **Q** refer to equivalence classes of **Q**.

1. Knowledge **Q** is derivable from knowledge **P**, if all equivalence classes of **Q** can be defined in terms of some equivalence classes of knowledge **P**, and then we will say that **Q** depends on **P**, written $\mathbf{P} \rightarrow \mathbf{Q}$, read as "if **P** then **Q**".
2. Knowledge **Q** depends in a degree k $(0 \le k \le 1)$ on knowledge **P** *if and only if*

$$k = \gamma_{\mathbf{P}}(\mathbf{Q}) = \frac{|\underline{\mathbf{P}}(\mathbf{Q}) = \{x \in U : \ [x]_{\mathbf{P}} \subseteq [x]_{\mathbf{Q}}\}|}{|U|}.$$

If $k = 1$, we say that **Q** depends totally on **P**, i.e., $\mathbf{P} \rightarrow \mathbf{Q}$; if $0 < k < 1$, we say that **Q** depends partially on **P**, i.e., $\mathbf{P} \rightarrow_k \mathbf{Q}$; and if $k = 0$, we say that **Q** is totally independent from **P**. Knowledge **P** and **Q** are independent *if and only if* **Q** is totally independent from **P** and **P** is totally independent from **Q**. The coefficient k expresses the ratio of all elements of the universe, which can be properly classified to concepts of the partition $U/\mathbf{Q}$, employing knowledge **P**.

3.2. Causality based on the graphical-counterfactual symbiosis

Two graphical criteria are introduced at the beginning, together with some basic graphical notation and terminology.

Definition 5 (Causal graph (Bareinboim & Pearl, 2016; Pearl, 2009)). A graph consists of a set of nodes, corresponding to variables, and a set of edges or links that connect some pairs of nodes, denoting a certain relationship that holds in pairs of nodes. Each edge in a graph can be either directed, i.e., marked by a single arrowhead on the edge, or undirected, i.e., unmarked links. A *path* in a graph is a sequence of consecutive edges in the graph regardless of direction. A *directed path* is a sequence of edges in the graph such that every edge is an arrow that points from the first to the second node of the pair. If all edges in a graph are directed, we then have a directed graph. Directed graphs may include directed cycles (e.g., $W \rightarrow Y$, $Y \rightarrow W$), representing mutual causation or feedback processes, but not self-loops (e.g., $W \rightarrow W$). A graph that contains no directed cycles is called *acyclic*. A graph that is both directed and acyclic is called a *directed acyclic graph* (DAG).

A causal model M is a triple $<U, V, F>$, where: U is a set of background variables, also called exogenous, that are determined by factors outside the model; V is a set $\{V_1, V_2, \ldots, V_n\}$ of variables, called endogenous, that are determined by variables in the model, that is, variables in $U \cup V$; and a set F of functions that determine or simulate how values are assigned to each variable $V_i \in V$. The diagram that captures the relationships among the variables in a causal model is called the *causal graph* G of M. The "child–parent or ancestor–descendant" relations are commonly used in a causal graph. Suppose W and Y are two different nodes in a causal graph. The arrow in $W \rightarrow Y$ designates W as a parent of Y and Y as a child of W, which merely indicates the possibility of causal connection between W and Y; a missing arrow from Y to W represents the assumption that Y has no effect on W once we intervene and hold the parents of W fixed; every missing bidirected link between W and Y represents the assumption that there are no common causes for W and Y, except those shown in the graph. Every causal model M can be associated with a directed graph.

The first criterion is known as a "separation" condition, a purely graphical characterization of conditional independence that the distribution compatible with a DAG must satisfy. When a causal graph is associated with a probability distribution, the Faithfulness/Stability condition will be imposed on the distribution, formally claiming that the Markov Condition applied to the graph characterizes all of the conditional independence relations that hold in the distribution (Spirtes et al., 1993).

Definition 6 (d-separation (Pearl, 2009)). Let W, Y and Z be three disjoint sets of variables in a directed acyclic graph G, and let p be any path between a variable in W and a variable in Y. A set Z is said to block a path p *if and only if* there is a node m on p satisfying one of the following two conditions:

1. p contains a chain $\rightarrow m \rightarrow$ or a fork $\leftarrow m \rightarrow$ such that the middle node m is in Z, or
2. p contains an inverted fork or collider $\rightarrow m \leftarrow$ such that the middle node m is not in Z and such that no descendant of m is in Z.

A set Z is said to d-separate W from Y in G, written $(W \perp\!\!\!\perp Y|Z)_G$, *if and only if* Z blocks every path from a node in W to a node in Y.

The "d" denotes *directional* and "blocking" means stopping the flow of information or of dependency between the variables that are connected by such paths, as defined in Definition 6. The unconditional independence, also called marginal independence, can be denoted by $W \perp\!\!\!\perp Y|\emptyset$ or $W \perp\!\!\!\perp Y$.

The second criterion concerns selecting a set of factors, called a "sufficient set" or "admissible set", for adjustment in a graphical fashion.

Definition 7 (Back-door criterion (Pearl, 2009)). Let W and Y be the two disjoint sets of nodes in a directed acyclic graph G, and $W_i \in W$ and $W_j \in Y$. A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (W_i, W_j), in other words, Z is admissible or sufficient for adjustment, if :

1. no node in Z is a descendant of W_i ; and
2. Z blocks every path between W_i and W_j that contains an arrow into W_i.

Z is said to satisfy the back-door criterion relative to (W, Y) if it satisfies the criterion relative to any pair (W_i, W_j).

Lemma 1 (Back-Door Adjustment (Pearl, 2009)). *If a set of variables Z satisfies the back-door criterion relative to (W, Y), then the causal effect of W on Y is identifiable and is given by the formula*

$$P(Y = y|do(W = w)) = \sum_z P(Y = y|W = w, Z = z)P(Z = z), \quad (1)$$

where W, Y and Z are the three disjoint sets of variables and w, y, z respectively denote their values; $do(W = w)$, or $do(w)$ for short, denotes an external intervention, i.e., setting $W = w$; $P(Y = y|do(W = w))$ stands for the probability of achieving a yield level of $Y = y$, given that the treatment is set to level $W = w$ by external intervention.

Eq. (1) represents the standard adjustment for Z when W is conditionally ignorable given Z and the ignorability conditions are reduced to the back-door criterion of Definition 7 (Pearl, 2009).

Definition 8 (Stable Unbiasedness (Pearl, 2009)). Let A be a set of assumptions or restrictions on the data-generating process, and let C_A be a class of causal models satisfying A. The effect estimate of W on Y is said to be stably unbiased given A if $P(y|do(w)) = P(y|w)$

holds in every model M in C_A. Correspondingly, we say that the pair (W, Y) is stably unconfounded given A.

Definition 9 (Structurally Stable No-Confounding (Pearl, 2009))**.** Let A_D be the set of assumptions embedded in a causal diagram D. We say that W and Y are stably unconfounded given A_D if $P(y|do(w)) = P(y|w)$ holds in every parameterization of D. By "parameterization" we mean an assignment of functions to the links of the diagram and prior probabilities to the background variables in the diagram.

Lemma 2 (Common-Cause Principle (Pearl, 2009))**.** *Let A_D be the set of assumptions embedded in an acyclic causal diagram D. Variables W and Y are stably unconfounded given A_D if and only if W and Y have no common ancestor in D.*

Lemma 3 (Criterion for Stable No-Confounding (Pearl, 2009))**.** *Let A_Z denote the assumptions that (i) the data are generated by some (unspecified) acyclic model M and (ii) Z is a variable in M that is unaffected by W but may possibly affect Y. If both of the associational criteria $P(w|z) = P(w)$ and $P(y|z, w) = P(y|w)$ are violated, then (W, Y) are not stably unconfounded given A_Z.*

Whenever the diagram D is acyclic, the back-door criterion provides a necessary and sufficient test for stable no-confounding, given A_D (Pearl, 2009). In the simple case of no adjustment for covariates, the back-door criterion reduces to the nonexistence of a common ancestor, observed or latent, of W and Y (Pearl, 2009).

To assess the likelihood that one event W was the cause of another Y, the counterfactual definition of causation is given by the notions of "necessary causation" and "sufficient causation" (Pearl, 2009). The necessary component of causation is predominant in legal settings and in ordinary discourse; when the necessary component is either dormant or ensured, the sufficiency component can be uncovered and shows a definite influence on causal thoughts (Pearl, 2009).

Definition 10 (Pearl, 2009; Tian & Pearl, 2000)**.** Let W and Y be two binary variables in a causal model. Let w and y stand respectively for the propositions $W =$ true and $Y =$ true, and let w' and y' denote their complements. $Y_{w'} = y'$ means that y would not have occurred in the absence of w, denoted as $y'_{w'}$ for short. y_w is the short for $Y_w = y$ representing that y would have occurred by setting w.

The probability that Y would be y' in the absence of event w, given that w and y did in fact occur, denoted as probability of necessity (PN), is defined as the expression $PN(w, y) = P(Y_{w'} = y'|W = w, Y = y)$.

The probability that setting w would produce y in a situation where w and y are in fact absent, denoted as probability of sufficiency (PS), is defined as $PS(w, y) = P(Y_w = y|W = w', Y = y')$.

Lemma 4 (Tian & Pearl, 2000)**.** *Whenever the causal effects $P(y|do(w'))$, $P(y|do(w))$ and $P(y'|do(w'))$ are identifiable, PN and PS are bounded with the following tight upper and lower bounds:*

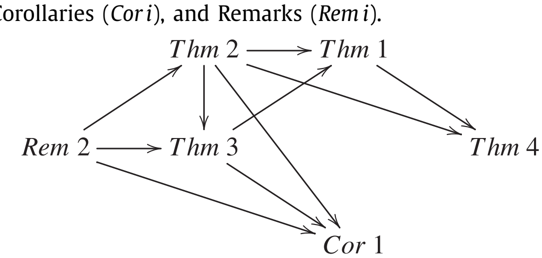
$$\max\left\{0, \frac{P(y) - P(y|do(w'))}{P(w, y)}\right\} \le PN \le \min\left\{1, \frac{P(y'|do(w')) - P(w', y')}{P(w, y)}\right\}, \tag{2}$$

$$\max\left\{0, \frac{P(y|do(w)) - P(y)}{P(w', y')}\right\} \le PS \le \min\left\{1, \frac{P(y|do(w)) - P(w, y)}{P(w', y')}\right\}. \tag{3}$$

Remark 1. Note that the structural theory starts with the causal graph, whereas there is no (available) graphical structure of the variables in the original data tables in the context of rough set theory. The graphical criteria, i.e., d-separation, back-door criterion and stable no-confounding criterion, find the variables for adjustment directly from the known (directed) acyclic graph while rough set theory gets the variable for adjustment having the graphical structures that satisfy these criteria, directly from data. These structures need to be first constructed through the dependency defined by the lower approximations.

4. The identification of effects under the interventions for attributes based on rough sets

In this section, we mainly focus on learning cause–effect relationships by means of attribute-value tables and rough set theory. Consider that statistical relationships between two variables alone cannot reveal causal information, may be reversed by increasing additional factors, and cannot predict the response of systems to hypothetical interventions (e.g., actions or decisions). To estimate the causal effect $P(y|do(w))$ of the attribute variables W and Y, namely the effect of action $do(W = w)$ on Y, and to decide which variable is appropriate for adjustment, the basic formal notions in our analysis rest on the lower approximation and the criterion concerning the adjustment for appropriate variables as defined in the above section. The key to the analysis is the exposition of how the lower approximation can guarantee the validity of the back-door adjustment formula in the given data table, considering that probabilities involved in the adjustment can be computed using rough set theory as soon as the attribute-value table is available. The diagram below shows the dependency between the Theorems (*Thm i*), Corollaries (*Cor i*), and Remarks (*Rem i*).



An organization of the section

- Thm 1: conditions of identifying the unbiased causal effects of interventions via the adjustment formula by employing lower approximations for variables in attribute-value tables.
- Thm 2: the relation between joint probabilities and lower approximations involved in Thm 1.
- Thm 3: the relation between conditional probabilities and lower approximations involved in Thm 1.
- Rem 2: the details on the relation between the numbers of attribute values and the simultaneous existence of empty lower approximations;
- Cor 1: the extension of Thm 2 and 3 with two attribute values to the case with more than two values.
- Thm 4: conditions of identifying causal effects of interventions with confounding bias via the adjustment formula by

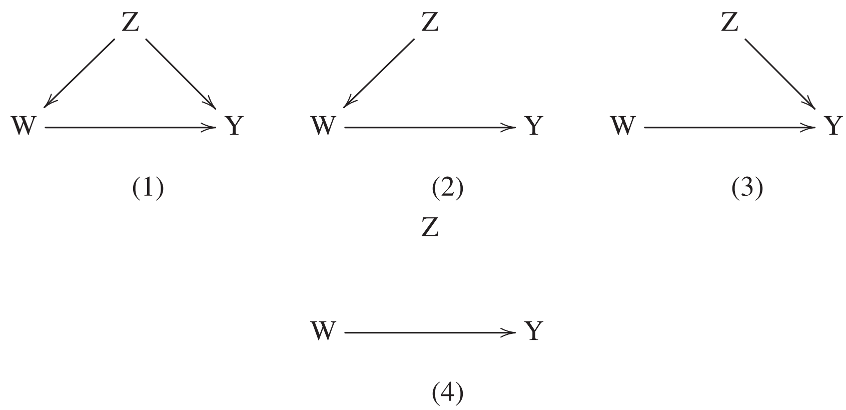


Fig. 1. The acyclic graph structures from W to Y with no directed paths from W to Z.

employing lower approximations for variables in attribute-value tables.

Theorem 1. *Let $(U, \mathbf{A})$ be an information system, U and $\mathbf{A}$ are finite, nonempty sets, called the universe and the set of attributes, respectively. Let W, Y and Z be any three different attributes from $\mathbf{A}$, each can be a single attribute or a combination of attributes, and let their attribute values be binary, denoted as $\{w, w'\}$, $\{y, y'\}$ and $\{z, z'\}$. The causal effect $P(y|do(w))$ can be identified without confounding bias via the adjustment formula*

$$P(Y=y|do(W=w)) = \sum_z P(Y=y|W=w, Z=z)P(Z=z)$$

and likewise $P(y|do(w'))$, $P(y'|do(w))$ and $P(y'|do(w'))$ whenever there exists a variable Z which satisfies the following conditions:

(a) *There exists “$\underline{Z}(W)=\emptyset=\underline{W}(Z)$, $\underline{W}(Y)\cap\underline{Y}(Z)=\emptyset$, $\underline{W}(Y)\cap\underline{Y}(W)=\emptyset$, $\underline{Z}(Y)\cap\underline{Y}(Z)=\emptyset$ and $\underline{Z}(Y)\cap\underline{Y}(W)=\emptyset$” or “$\underline{Z}(W)\neq\emptyset$, $\underline{W}(Z)=\emptyset$, $\underline{W}(Y)\cap\underline{Y}(Z)=\emptyset$, $\underline{W}(Y)\cap\underline{Y}(W)=\emptyset$, $\underline{Z}(Y)\cap\underline{Y}(Z)=\emptyset$ and $\underline{Z}(Y)\cap\underline{Y}(W)=\emptyset$”;*
(b) *There exists either $P(W)=P(W|Z)$ or $P(Y|W,Z)=P(Y|W)$;*
(c) *The joint probability function $P(W,Z)=\{P(w,z), P(w',z), P(w,z'), P(w',z')\}$ can be estimated from the sample data and none of the elements equal zero.*

Proof. The proof is based on the view that, when a variable Z satisfies the conditions (a)–(c), Z meets the back-door criterion in Definition 7 with a constraint that the covariates are substituted by a single variable. In this case, according to Definition 7 and Lemma 2, the back-door criterion reduces to the nonexistence of a common ancestor.

According to the assumptions embedded in Definitions 6 and 7 as well as Lemmas 2 and 3, the graph among Z, W and Y needs to be acyclic, that is, the graph contains no directed circles. Meanwhile, Z is unaffected by W, i.e., the graph must contain no directed paths from W to Z (here it means $W \nrightarrow Z$ and $W \rightarrow Y \nrightarrow Z$).

In rough set theory, the links that connect some pairs of attribute variables can be represented by the dependency between two variables through the notion of the lower approximation relative to the attributes in Definition 4. An arrow pointing to W from Z, i.e., $Z \rightarrow W$ is obtained by $\underline{Z}(W)\neq\emptyset$, which means W can be derived from Z; a missing arrow from W to Z (i.e., $W \nrightarrow Z$) is derived from $\underline{W}(Z)=\emptyset$ which means Z cannot be derived from W; a missing bidirected arrow between W and Z is derived from $\underline{W}(Z)=\emptyset$ and $\underline{Z}(W)=\emptyset$ which means attributes W and Z are independent, namely none of the elements of the universe can be classified to the equivalence classes of Z by employing W and none of the elements of the universe can be classified to the equivalence classes of W by employing Z. The absence of directed circle $W \rightarrow Y \rightarrow W$ is obtained by $\underline{W}(Y)\cap\underline{Y}(W)=\emptyset$. The arrows in $W \rightarrow Z \rightarrow Y$ can be obtained by $\underline{W}(Z)\cap\underline{Z}(Y)\neq\emptyset$. The arrows in $W \rightarrow Y \nrightarrow Z$ can be obtained by $\underline{W}(Y)\cap\underline{Y}(Z)=\emptyset$. The arrows in $Z \rightarrow Y \nrightarrow W$ can be obtained by $\underline{Z}(Y)\cap\underline{Y}(W)=\emptyset$.

When Z, W and Y fulfil the condition (a), expressed in graphic terms as the following graphs (1)–(4) from W to Y in Fig. 1, it is remarkable that the graphs relative to the ordered triple (Z, W, Y) satisfy the assumptions in Definitions 6–7 and Lemmas 2 and 3. Further when Z from graphs (1)–(4) satisfies the condition (b), namely Z is not associated with W or Z is not associated with Y, conditional on W, it is easy to verify according to Lemmas 2 and 3 that W and Y have no common ancestor and Z satisfies the conditions of d-separation and back-door criterion, from the probabilistic viewpoint. The independence is required due to the difference between “$\subseteq$” operator in the definitions of the lower approximation and “$\cap$” operator in the definitions of certainty factor given by rough set theory, that is, $\underline{Z}(W)=\emptyset=\underline{W}(Z)$ in graph (3)–(4) does not necessarily lead to $P(W)=P(W|Z)$; $\underline{Z}(W)\neq\emptyset$ as well as $\underline{W}(Y)\cap\underline{Y}(Z)=\emptyset$ in graphs (1) and (2) does not necessarily result in $P(Y|W,Z)=P(Y|W)$. The condition (c) for Z is based on the Kolmogorov definition of conditional probability, i.e., $P(A|B)=\frac{P(A,B)}{P(B)}$ $(P(B)>0)$, which is also embedded in Lemma 1. Therefore, the variable Z satisfies the back-door criterion and thus $P(y|do(w))$ can be given by the adjustment formula without bias. □

The intuitive motivation behind Theorem 1 is to build the graphical structures directly from data, embedded in the back-door criterion and the stable no-confounding criterion, through the dependencies between variables defined by the lower approximation, provided the data set is known and can be converted into the form of decision table or the attribute-value table. Since the back-door criterion can guarantee the elimination of all confounding bias relative to the causal effects and acts directly on a given causal graph, the variable Z obtained from data, rather than a causal graph, must not only satisfy the structures hidden in back-door criterion but assure the no-confounding of W and Y in Definition 9 and Lemmas 2 and 3. This is also the reason for the existence of condition (b). Theorem 1 permits us to measure causal effect of W on Y by doing some physical interventions (e.g., setting $W=w$) and can be used as a basis for the interpretation and identification of causation. Another point in this theorem that deserves more attention is the relationship between dependencies (more precisely, lower approximations) and probabilities. The following conclusion can help us, a certain extent, understand this issue.

Theorem 2. *There exists a joint probability function*

$$P_{W,Z} = \{P(w,z), P(w',z), P(w,z'), P(w',z')\}$$

such that none of the elements equal zero if and only if $\underline{Z}(W)=\emptyset$ *and* $\underline{W}(Z)=\emptyset$.

Proof. Let $\underline{Z}(W)=\emptyset$. According to the definition of *Z*-lower approximation of *W*, namely $\underline{Z}(W)=\{x\in U:[x]_Z\subseteq[x]_W\}$, we can get $[x]_Z\nsubseteq[x]_W$. More specifically,

$$\{x\in U:[x]_z\subseteq[x]_w\}=\emptyset,\{x\in U:[x]_z\subseteq[x]_{w'}\}=\emptyset,$$

$$\{x\in U:[x]_{z'}\subseteq[x]_w\}=\emptyset,\{x\in U:[x]_{z'}\subseteq[x]_{w'}\}=\emptyset.$$

Similarly, if $\underline{W}(Z)=\emptyset$, then we have $[x]_W\nsubseteq[x]_Z$, that is,

$$\{x\in U:[x]_w\subseteq[x]_z\}=\emptyset,\{x\in U:[x]_w\subseteq[x]_{z'}\}=\emptyset,$$

$$\{x\in U:[x]_{w'}\subseteq[x]_z\}=\emptyset,\{x\in U:[x]_{w'}\subseteq[x]_{z'}\}=\emptyset.$$

Consider the fact that $[x]_w+[x]_{w'}=[x]_z+[x]_{z'}$, both U/W and U/Z are the partitions of the universe *U*. Thus,

$$\begin{aligned}&\emptyset\neq[x]_z\cap[x]_w\subsetneqq\min\{[x]_w,[x]_z\},\\&\emptyset\neq[x]_z\cap[x]_{w'}\subsetneqq\min\{[x]_{w'},[x]_z\},\\&\emptyset\neq[x]_{z'}\cap[x]_w\subsetneqq\min\{[x]_w,[x]_{z'}\},\\&\emptyset\neq[x]_{z'}\cap[x]_{w'}\subsetneqq\min\{[x]_{w'},[x]_z\},\end{aligned}$$

(otherwise if $[x]_z\cap[x]_w=\emptyset$, that means $[x]_z\subseteq[x]_{w'}$ which is contradictory to the above $\{x\in U:[x]_z\subseteq[x]_{w'}\}=\emptyset$). By doing so, we can obtain $P(w,z)\neq0$, $P(w',z)\neq0$, $P(w,z')\neq0$ and $P(w',z')\neq0$, because of $P(w,z)=\frac{|[x]_w\cap[x]_z|}{|U|}$.

The proof of the converse proceeds in the same way as the proof of "if $\underline{Z}(W)=\emptyset$ and $\underline{W}(Z)=\emptyset$ then there exists $P_{WZ}=\{P(w,z),P(w',z),P(w,z'),P(w',z')\}$ such that none of the elements equal zero" does. □

For example, consider the data shown in Table 1 associated with the characterization of flu. From $\underline{M}(H)=\{p2,p5\}$ and $\underline{H}(M)=\{p1,p4,p6\}$, we get $P(h_0,m_0)=0$. From $\underline{T}(M)=\{p3,p4,p6\}$ and $\underline{M}(T)=\{p2,p5\}$, there holds $P(t_1,m_0)=0=P(t_3,m_0)$. From $\underline{T}(H)=\{p4\}$ and $\underline{H}(T)=\emptyset$, there is $P(h_1,t_3)=0$. From $\underline{H}(F)=\emptyset=\underline{F}(H)$, we have $P(f_1,h_1)=\frac{1}{3}=P(f_1,h_0)$, $P(f_0,h_1)=\frac{1}{6}=P(f_0,h_0)$.

Theorem 2 tells us that when variables *W* and *Y* are independent, i.e., *W* is not derivable from *Z* and *Z* is not derivable from *W*, values of the joint probability of *W* and *Z* are greater than zero. When only one of the dependencies between two variables (i.e., one variable is or is not derivable from the other, the reverse is unknown) are given, our attention is the change in conditional probability of these two variables, formulated as the following theorem.

Theorem 3.

(i) *If* $\underline{Y}(Z)=\emptyset$, *then* $0<P(z|y), P(z'|y), P(z|y'), P(z'|y')<1$ *and vice versa.*
(ii) *If* $\underline{Y}(Z)\neq\emptyset$, *then at least one of* $P(z|y)$, $P(z'|y)$, $P(z|y')$, $P(z'|y')$ *equals 1 or 0, and vice versa.*

Table 1
An information system about patients suffering from flu.

Patient	*Headache (H)*	*Muscle-pain (M)*	*Temperature (T)*	*Flu (F)*
p1	*no*	*yes*	*high*	*yes*
p2	*yes*	*no*	*high*	*yes*
p3	*yes*	*yes*	*veryhigh*	*yes*
p4	*no*	*yes*	*normal*	*no*
p5	*yes*	*no*	*high*	*no*
p6	*no*	*yes*	*veryhigh*	*yes*

Proof. It is obvious that from $\underline{Y}(Z)=\emptyset$ we can obtain

$$\underline{Y}(Z)=\{x\in U:[x]_Y\subseteq[x]_Z\}=\emptyset,\ \text{i.e., } [x]_Y\nsubseteq[x]_Z,$$

hence $0\neq|[x]_y\cap[x]_z|<|[x]_y|$, $0\neq|[x]_y\cap[x]_{z'}|<|[x]_y|$, $0\neq|[x]_{y'}\cap[x]_z|<|[x]_{y'}|$ and $0\neq|[x]_{y'}\cap[x]_{z'}|<|[x]_{y'}|$ in the light of $[x]_y+[x]_{y'}=[x]_z+[x]_{z'}$. Therefore, $0<P(z|y)=\frac{|[x]_y\cap[x]_z|}{|[x]_y|}<1$, $0<P(z'|y)=\frac{|[x]_y\cap[x]_{z'}|}{|[x]_y|}<1$, $0<P(z|y')=\frac{|[x]_{y'}\cap[x]_z|}{|[x]_{y'}|}<1$ and $0<P(z'|y')=\frac{|[x]_{y'}\cap[x]_{z'}|}{|[x]_{y'}|}<1$. In this way, the proof of the opposite direction is easily understood.

Concerning $\underline{Y}(Z)\neq\emptyset$, there is $\underline{Y}(Z)=\{x\in U:[x]_Y\subseteq[x]_Z\}\neq\emptyset$, which means there holds at least one of $[x]_y\subseteq[x]_z$ or $[x]_y\subseteq[x]_{z'}$ or $[x]_{y'}\subseteq[x]_z$ or $[x]_{y'}\subseteq[x]_{z'}$. Let $[x]_y\subseteq[x]_z$, then we can get $0<|[x]_y\cap[x]_z|=|[x]_y|$ and $|[x]_y\cap[x]_{z'}|=0$. Thus, $P(z|y)=1$ and $P(z'|y)=0$. Proving the opposite side can be done in the same way. □

Remark 2. It is worth noting that $\underline{Y}(Z)=\emptyset$ does not necessarily imply $\underline{Z}(Y)=\emptyset$, in other words, given that *Z* is not derivable from *Y*, it is unsure whether *Y* is derivable from *Z*. This crucially depends on the number of attribute values for each attribute (e.g., *Y* and *Z*) in terms of rough set theory. The deep understanding of this issue is as follows.

By means of the definition of $\underline{Y}(Z)$, it is clear that $\underline{Y}(Z)=\{x\in U:[x]_Y\subseteq[x]_Z\}=\emptyset$, that is to say, $[x]_y\nsubseteq[x]_z$, $[x]_y\nsubseteq[x]_{z'}$, $[x]_{y'}\nsubseteq[x]_z$ and $[x]_{y'}\nsubseteq[x]_{z'}$ hold at the same time. Further, $\emptyset\neq[x]_y\cap[x]_z\subsetneqq[x]_z$, $\emptyset\neq[x]_y\cap[x]_{z'}\subsetneqq[x]_{z'}$, $\emptyset\neq[x]_{y'}\cap[x]_z\subsetneqq[x]_z$ and $\emptyset\neq[x]_{y'}\cap[x]_{z'}\subsetneqq[x]_{z'}$ can be validated as a result of $[x]_y+[x]_{y'}=[x]_z+[x]_{z'}$, in other words, $[x]_z\nsubseteq[x]_y$, $[x]_{z'}\nsubseteq[x]_y$, $[x]_z\nsubseteq[x]_{y'}$ and $[x]_{z'}\nsubseteq[x]_{y'}$, and hence we can get $\underline{Z}(Y)=\{x\in U:[x]_Z\subseteq[x]_Y\}=\emptyset$. It seems that $\underline{Y}(Z)=\emptyset$ entails $\underline{Z}(Y)=\emptyset$, but it is not always the case.

Take $U/Z=\{[x]_{z_0},[x]_{z_1},[x]_{z_2}\}$ and $U/Y=\{[x]_{y'},[x]_y\}$ for example. From $\underline{Y}(Z)=\emptyset$ we can infer that $[x]_y$ is not included in $[x]_{z_0}$, $[x]_{z_1}$ and $[x]_{z_2}$, and $[x]_{y'}$ does likewise. Therefore, we obtain the following possibilities: $[x]_{z_0}\subsetneqq[x]_y$ and $[x]_{z_1}+[x]_{z_2}\supsetneqq[x]_{y'}$ (here the positions of $[x]_{z_0}$, $[x]_{z_1}$ and $[x]_{z_2}$ are fully interchangeable, also the part of $[x]_y$ is interchangeable with that of $[x]_{y'}$), which means $\underline{Z}(Y)\neq\emptyset$; $[x]_{z_0}$ goes across the areas of $[x]_y$ and $[x]_{y'}$ but is not included in each of them, so do $[x]_{z_1}$ and $[x]_{z_2}$, which suggests $\underline{Z}(Y)=\emptyset$.

Having considered the above analysis, it is not difficult to find that assuming $U/Y=\{[x]_{y_0},[x]_{y_1},\dots,[x]_{y_n}\}$ ($n+1$ represents the number of values of attributes *Y*) and $U/Z=\{[x]_{z_0},[x]_{z_1},\dots,[x]_{z_m}\}$ ($m+1$ represents the number of values of attributes *Z*), for $n\geq1$ and $m>1$ there hold two possibilities of $\underline{Z}(Y)\neq\emptyset$ and $\underline{Z}(Y)=\emptyset$ using $\underline{Y}(Z)=\emptyset$ in that we can always construct the case of $[x]_{z_0}\subsetneqq[x]_{y_0}$ with the rest of $[x]_{z_i}$ $(i=1,2,\dots,m)$ satisfying $[x]_{y_j}\nsubseteq[x]_{z_i}$ $(j=1,2,\dots,n)$, and the case of $\emptyset\neq[x]_{z_i}\cap[x]_{y_j}$ and $[x]_{z_i}\nsubseteqq[x]_{y_j}$; only $\underline{Z}(Y)=\emptyset$ holds with $n\geq1$ and $m=1$.

Similarly, $\underline{Y}(Z)\neq\emptyset$ does not necessarily imply $\underline{Z}(Y)\neq\emptyset$. As to the case of $n=1$ and $m\geq1$, there is only $\underline{Z}(Y)\neq\emptyset$ in that we can always get one of z_j $(j=1,\dots,m)$ such that $[x]_{z_j}\subseteq[x]_{y_1}$ from $[x]_Y\subseteq[x]_Z$, without loss of generality, assuming that $[x]_{y_0}\subseteq[x]_{z_0}$. When setting $n>1$ but $m\geq1$, however, there exists $\underline{Z}(Y)=\emptyset$ or $\underline{Z}(Y)\neq\emptyset$. The reason behind this result is that for $n>1$ and $m=1$, $\underline{Z}(Y)=\emptyset$ can be obtained due to $[x]_{z_0}\nsubseteq[x]_{y_0}$ and $[x]_{z_1}\nsubseteq[x]_{y_i}$ $(i=1,\dots,n)$ from $[x]_{y_0}\subseteq[x]_{z_0}$ besides the result of $\underline{Z}(Y)\neq\emptyset$ with $[x]_{z_1}\subseteq[x]_{y_n}$ and $[x]_{z_0}\nsubseteq[x]_{y_i}$ $(i=0,\dots,n-1)$ or $[x]_{z_0}=[x]_{y_0}$ and $[x]_{z_1}\nsubseteq[x]_{y_i}$ $(i=1,\dots,n)$. For the case of $n>1$ and $m>1$, assuming $[x]_{y_0}\subsetneqq[x]_{z_0}$, then we can construct $[x]_{z_j}\nsubseteqq[x]_{y_i}$ $(j=0,1,\dots,m;i=0,1,\dots,n)$, which entails $\underline{Z}(Y)=\emptyset$; or at least $[x]_{z_m}\subseteq[x]_{y_n}$ with none of $[x]_{z_j}$ $(j=0,1,\dots,m-1)$ included in $[x]_{y_i}$ $(i=0,1,\dots,n-1)$, which implies $\underline{Z}(Y)\neq\emptyset$.

Note that Theorems 1–3 are established as the value of attribute variable is binary. As for n-tuple attribute values, based on the analysis of Remark 2, there hold Theorem 2 and (ii) of Theorem 3, the details are given as follows:

Corollary 1. *Let $(U, \mathbf{A})$ be an information system, U and $\mathbf{A}$ are finite, nonempty sets called the universe and the set of attributes, respectively. Let W, Y and Z be any three different variables from $\mathbf{A}$, each can be a single attribute or a combination of attributes. Suppose $x \in U$.*

(a) *Let $U/W = \{[x]_{w_0}, \ldots, [x]_{w_n}\}(n \geq 1)$ and $U/Z = \{[x]_{z_0}, \ldots, [x]_{z_m}\}(m \geq 1)$. The joint probability function $(j = 0, \ldots, n; i = 0, \ldots, m)$*

$$P(W, Z) = \{P(w_0, z_0), P(w_0, z_1), \ldots, P(w_j, z_i), \ldots, P(w_n, z_m)\}$$

exists and none of the elements are equal to zero if and only if $\underline{Z}(W) = \emptyset$ and $\underline{W}(Z) = \emptyset$.

(b) *Let $U/Y = \{[x]_{y_0}, \ldots, [x]_{y_n}\}(n \geq 1)$ and $U/Z = \{[x]_{z_0}, \ldots, [x]_{z_m}\}(m \geq 1)$. If $\underline{Y}(Z) \neq \emptyset$, then at least one of $\{P(z_0|y_0), \ldots, P(z_m|y_n)\}$ equals 1 or 0, and vice versa.*

(c) *Let $U/Y = \{[x]_{y_0}, \ldots, [x]_{y_n}(n \geq 1)\}$ and $U/Z = \{[x]_{z_0}, [x]_{z_1}\}$. If $\underline{Y}(Z) = \emptyset$, then $0 < P(z_0|y_0), P(z_1|y_0), \ldots, P(z_1|y_n) < 1$ and vice versa.*

(i) of Theorem 3 is not valid when $n \geq 1$ and $m > 1$ because there always exists one of $\{P(z_0|y_0), \ldots, P(z_m|y_n)\}$ that equals zero.

Theorem 4. *Let W, Y and Z be any three different attributes from $\mathbf{A}$, each can be a single attribute or a combination of attributes, and let their attribute values be binary, denoted as $\{w, w'\}$, $\{y, y'\}$ and $\{z, z'\}$. The causal effect $P(y|do(w))$ can be identified with confounding bias via the adjustment formula*

$$P(Y = y|do(W = w)) = \sum_{z} P(Y = y|W = w, Z = z)P(Z = z)$$

and likewise $P(y|do(w'))$, $P(y'|do(w))$ and $P(y'|do(w'))$ whenever there exists a variable Z which satisfies the following conditions:

(a) *There exists "$\underline{Z}(W) = \emptyset = \underline{W}(Z)$, $\underline{W}(Y) \cap \underline{Y}(Z) = \emptyset$, $\underline{W}(Y) \cap \underline{Y}(W) = \emptyset$, $\underline{Z}(Y) \cap \underline{Y}(Z) = \emptyset$ and $\underline{Z}(Y) \cap \underline{Y}(W) = \emptyset$" or "$\underline{Z}(W) \neq \emptyset$, $\underline{W}(Z) = \emptyset$, $\underline{W}(Y) \cap \underline{Y}(Z) = \emptyset$, $\underline{W}(Y) \cap \underline{Y}(W) = \emptyset$, $\underline{Z}(Y) \cap \underline{Y}(Z) = \emptyset$ and $\underline{Z}(Y) \cap \underline{Y}(W) = \emptyset$";*

(b) *There exist $P(W) \neq P(W|Z)$ and $P(Y|W, Z) \neq P(Y|W)$;*

(c) *$\underline{Z}(W) = \emptyset = \underline{W}(Z)$ and $\underline{Z}(WY) = \emptyset = \underline{WY}(Z)$ with $[x]_{wy} \neq \emptyset$.*

Proof. Based on the analysis of Theorem 1, it is easy to verify that the graph among Z, W and Y is acyclic. However, Z does not meet the conditions of the back-door criterion due to the existence of common ancestor Z, which means W and Y are not stably unconfounded by Lemma 3.

Using Theorem 2, there holds $P(W, Z) > 0$ and $P(W, Y, Z) > 0$ when Z satisfies the condition (c). If Z fulfils the conditions (a)-(c), then by application of the product rule of probability, the joint distribution

$$P(Z, W, Y) = P(Y|W, Z)P(W|Z)P(Z)\,(P(Z, W, Y) > 0)$$

holds. Therefore, the distribution generated by an intervention $do(W = w)$ on the endogenous variable W is given by $P(Y, Z|do(W = w)) = P(Y|w, Z)P(Z)$, amounting to removing the equation $P(W|Z)$ from $P(Z, W, Y)$ and substituting w for W in the remaining equations. Further, there holds

$$P(Y|do(W = w)) = \sum_{Z} P(Y, Z|do(W = w)) = \sum_{Z} P(Y|w, Z)P(Z).$$

In view of $P(Y|W) = \sum_Z P(Y|W, Z)P(Z|W)$, there exists $P(Y|do(W = w)) \neq P(Y|w)$ due to the condition (b'), which implies the confounding bias exists between these two distributions, defined by the difference between them. □

Consider that Theorem 2 is not affected by the dimensionality of the attribute values. Theorems 1 and 4 still hold for n-tuple attribute values. Based on the analysis above, we can easily see that $\underline{Z}(W) = \emptyset = \underline{W}(Z)$ guarantees that the intersections of the equivalence classes induced by attribute W and the equivalence classes induced by attribute Z are not empty and the inclusion relations between the equivalence classes of attribute W and the equivalence classes of attribute Z do not exist, which means the probabilities $P(W|Z)$, $P(Z|W)$, $P(W, Z)$ are greater than zero. So it provides neither the information on the marginal independence of Z and W in probabilistic terms (i.e., $P(W) = P(W|Z)$), that is, the probability of W will not be affected by the discovery of Z, nor the information on the conditional independence of W and Y given Z (i.e., $P(Y|W, Z) = P(Y|W)$).

5. The identification of causation based on rough sets

Building upon the above conditions of estimating the effects of the interventions, we now turn our attention to make an estimate of causation (i.e., the likelihood that one event was the cause of another) from data table and we concentrate particularly on the identifiability of the counterfactual quantities "probability of necessity" (*PN*) and "probability of sufficiency" (*PS*) using the specific data tables with different highlights as shown in the sequel.

5.1. The construction of causation

Assuming data tables have been defined or collected and W, Y are different attributes and $\{w, w'\}$, $\{y, y'\}$ represent the corresponding attribute values, we now determine the causal quantities $PN(w, y)$ (the probability that w is a necessary cause of y) and $PS(w, y)$ (the probability that w is a sufficient cause of y) in three simple steps:

Step 1. Select an attribute Z such that: for the unbiased estimate, $\underline{Z}(W) = \emptyset = \underline{W}(Z)$, together with "either $P(W) = P(W|Z)$ or $P(Y|W, Z) = P(Y|W)$"; for the biased estimate, $\underline{Z}(W) = \emptyset = \underline{W}(Z)$, $\underline{Z}(WY) = \emptyset = \underline{WY}(Z)$, $[x]_{wy} \neq \emptyset$ plus $P(W) \neq P(W|Z)$ and $P(Y|W, Z) \neq P(Y|W)$. Meanwhile, $\underline{W}(Y) \cap \underline{Y}(Z) = \emptyset$, $\underline{W}(Y) \cap \underline{Y}(W) = \emptyset$, $\underline{Z}(Y) \cap \underline{Y}(Z) = \emptyset$ and $\underline{Z}(Y) \cap \underline{Y}(W) = \emptyset$.

Step 2. Compute the probabilities $P(y|do(w))$, $P(y|do(w'))$ and $P(y'|do(w'))$ using the adjustment formula $P\big(y|do(w)\big) = \sum_z P(y|w, z)P(z)$.

Step 3. Estimate the counterfactual quantities *PN* and *PS* by Lemma 4, namely

$$\max\left\{0, \frac{P(y) - P(y|do(w'))}{P(w, y)}\right\} \leq PN \leq$$
$$\min\left\{1, \frac{P(y'|do(w')) - P(w', y')}{P(w, y)}\right\},$$
$$\max\left\{0, \frac{P(y|do(w)) - P(y)}{P(w', y')}\right\} \leq PS \leq \min\left\{1, \frac{P(y|do(w)) - P(w, y)}{P(w', y')}\right\}.$$

5.2. Numerical examples

The illustrations of the construction of causal explanations are provided by considering some different data tables. Table 1 comes from Pawlak (2004) and shows an illustrative data table used for rough set based data analysis by Pawlak and his colleagues, and so the causal relationships between the variables in this table will be fully tested. Table 2 is a decision table (see Düntsch & Gediga, 1997; Yamaguchi, 2009) used to test the different influences between the two as the attribute dependency degree of two attributes gets zero. As for the causal explanation of the attributes, this table is used to identify the causation of two attributes for which their dependency degree is zero. Table 3 from Wang et al. (2003) describes the vertebrate world with 20 kinds of vertebrates and takes *Bird* as the decision attribute and *Fly* as one of the condition attributes, considering that the relation between *Bird* and *Fly* is a famous issue on non-monotonic reasoning. Table 4 was first created by Cendrowska (1987) and is located at the UCI Machine Learning repository, available at http://archive.ics.uci.edu/ml/datasets/Lenses. It contains 24 instances with four conditional attributes and one decision attribute and all the attributes have nominal values.

Example 1. Consider information system about patients suffering from flu shown in Table 1. The following is a description of each attribute: *Headache* - H has the following values "$h_1 = yes$ and $h_0 = no$"; *Muscle-pain* - M has the following values "$m_1 = yes$ and $m_0 = no$"; *Temperature* - T has the following values "$t_1 = very\ high$, $t_2 = high$ and $t_3 = normal$"; *Flu* - F has the following values "$f_1 = yes$ and $f_0 = no$".

In rough set theory, by Definition 1 we have

$U/\{T\} = \{\{p1, p2, p5\}_{[x]_{t_2}}, \{p3, p6\}_{[x]_{t_1}}, \{p4\}_{[x]_{t_3}}\}$,
$U/\{H\} = \{\{p1, p4, p6\}_{[x]_{h_0}}, \{p2, p3, p5\}_{[x]_{h_1}}\}$,
$U/\{M\} = \{\{p1, p3, p4, p6\}_{[x]_{m_1}}, \{p2, p5\}_{[x]_{m_0}}\}$,
$U/\{F\} = \{\{p1, p2, p3, p6\}_{[x]_{f_1}}, \{p4, p5\}_{[x]_{f_0}}\}$.

According to Definitions 3 and 4, Theorems 1 and 2, we get

(1) $\underline{H}(F) = \emptyset = \underline{F}(H)$ with $P(F) = P(F|H)$ and $P(H) = P(H|F)$;
(2) $\underline{M}(F) = \emptyset = \underline{F}(M)$ with $P(F) \neq P(F|M)$ and $P(M) \neq P(M|F)$.

For the case of (1), the choice of Z can be H or F, then the causal effects of H on M or on T can be identified via the adjustment for F due to $P(M|H, F) \neq P(M|H)$ and $P(T|H, F) \neq P(T|H)$ plus $\underline{H}(M) \cap \underline{M}(F) = \{p1, p4, p6\} \cap \emptyset = \emptyset$, $\underline{M}(H) \cap \underline{H}(M) = \{p2, p5\} \cap \{p1, p4, p6\} = \emptyset$, $\underline{M}(F) \cap \underline{F}(M) = \emptyset$, $\underline{F}(M) \cap \underline{M}(H) = \emptyset \cap \{p2, p5\} = \emptyset$ and $\underline{H}(T) \cap \underline{T}(F) = \emptyset \cap \{p3, p4, p6\} = \emptyset$, $\underline{T}(H) \cap \underline{H}(T) = \{p4\} \cap \emptyset = \emptyset$, $\underline{T}(F) \cap \underline{F}(T) = \{p3, p4, p6\} \cap \emptyset = \emptyset$, $\underline{F}(T) \cap \underline{T}(H) = \emptyset \cap \{p4\} = \emptyset$; the causal effects of F on M or on T can be identified via the adjustment for H due to $P(M|H, F) \neq P(M|F)$ and $P(T|H, F) \neq P(T|F)$ plus $\underline{F}(M) \cap \underline{M}(H) = \emptyset \cap \{p2, p5\} = \emptyset$, $\underline{M}(F) = \emptyset = \underline{F}(M)$, $\underline{M}(H) \cap \underline{H}(M) = \emptyset$, $\underline{H}(M) \cap \underline{M}(F) = \emptyset$ and $\underline{F}(T) \cap \underline{T}(H) = \emptyset \cap \{p4\} = \emptyset$, $\underline{T}(F) \cap \underline{F}(T) = \emptyset$, $\underline{T}(H) \cap \underline{H}(T) = \emptyset$, $\underline{H}(T) \cap \underline{T}(F) = \emptyset$. More specifically, we have

$P(m_1|do(h_1)) = \sum_F P(m_1|h_1, F)P(F) = \frac{1}{3} = P(m_1|h_1)$,
$P(t_1|do(h_1)) = \sum_F P(t_1|h_1, F)P(F) = \frac{1}{3} = P(t_1|h_1)$,
$P(t_1|do(f_1)) = \sum_H P(t_1|f_1, H)P(H) = \frac{1}{2} = P(t_1|f_1)$,
$P(m_1|do(f_1)) = \sum_H P(m_1|f_1, H)P(H) = \frac{3}{4} = P(m_1|f_1)$.

The calculation of PS and PN is given as follows: for $PN(h_1, m_1) = P(M_{h_0} = m_0|H = h_1, M = m_1)$ and $PS(h_1, m_1) = P(M_{h_1} = m_1|H = h_0, M = m_0)$, there exist

$$\frac{P(m_1) - P\big(m_1|do(h_0)\big)}{P(h_1, m_1)} = -2, \quad \frac{P\big(m_0|do(h_0)\big) - P(h_0, m_0)}{P(h_1, m_1)} = 0,$$

$$0 = \max\left\{0, \frac{P(m_1) - P\big(m_1|do(h_0)\big)}{P(h_1, m_1)}\right\} \leq PN \leq \min\left\{1, \frac{P\big(m_0|do(h_0)\big) - P(h_0, m_0)}{P(h_1, m_1)}\right\} = 0,$$

$$0 = \max\left\{0, \frac{P(m_1|do(h_1)) - P(m_1)}{P(h_0, m_0)}\right\} \leq PS \leq \min\left\{1, \frac{P(m_1|do(h_1)) - P(h_1, m_1)}{P(h_0, m_0)}\right\} = 0.$$

We see that $PN(h_1, m_1) = 0 = PS(h_1, m_1)$, which gives the impression that *Headache* was not a necessary and sufficient cause of muscle pain, that is, the patient will suffer from a pain in his/her muscles even if he does not have a headache. There is no causal relationship between *Headache* and $Muscle-pain$, in spite of the consistency $P(m_1|do(h_1)) \geq P(h_1, m_1)$ and no confounding $P\big(m_1|do(h_1)\big) = P(m_1|h_1)$. While the attribute M depends roughly from H in a degree of $\frac{1}{2}$ by employing $\gamma_{\{H\}}(\{M\}) = \frac{|\underline{H}(M)|}{|U|} = \frac{1}{2}$ (this quantity in fact measures the dependency of m_1 from h_0 due to $\underline{H}(M) = \{p1, p4, p6\} = [x]_{h_0} \subseteq [x]_{m_1}$, i.e., $\gamma_{\{h_0\}}(\{m_1\}) = \frac{1}{2}$ while the dependency of m_1 from h_1 is 0, i.e., $\gamma_{\{h_1\}}(\{m_1\}) = 0$), based on rough set theory. It is easy to obtain that $PN(h_0, m_1) = \frac{2}{3}$ and $PS(h_0, m_1) = 1$.

Similarly, $0 \leq PN(h_1, t_1) \leq 1$ and $0 \leq PS(h_1, t_1) \leq \frac{1}{2}$, meaning that *Headache* was not a necessary and sufficient cause of very high temperature, considering the $PN > \frac{1}{2}$ criterion (Tian & Pearl, 2000) and the fact that $0 \leq PS \leq \frac{1}{2}$ means that the probabilities of causation cannot be determined from statistical data (Pearl, 2009). While the attribute H is not derivable from T by employing $\gamma_{\{H\}}(\{T\}) = \frac{|\underline{H}(T)|}{|U|} = 0$, based on rough set theory.

$PN(f_1, t_1) = 1$ and $PS(f_1, t_1) = \frac{1}{2}$, reflecting that, we can be 100% sure that having flu was necessary for very high temperature. Meanwhile, there is 50% chance that a patient without flu and very high temperature would have suffered from very high temperature had he or she been gotten flu. While the attribute F is not derivable from T by employing $\gamma_{\{F\}}(\{T\}) = \frac{|\underline{F}(T)|}{|U|} = 0$, based on rough set theory.

$\frac{1}{3} \leq PN(f_1, m_1) \leq \frac{2}{3}$ and $\frac{1}{2} \leq PS(f_1, m_1) \leq 1$, which permit us to infer that the probability that a patient would not suffer from muscle pain in the absence of flu, given that this patient had flu and muscle pain, ranges from $\frac{1}{3}$ to $\frac{2}{3}$. The chance that having flu would have been sufficient for producing muscle pain is greater than 50%. However, in rough set theory, the attribute F is not derivable from M by employing $\gamma_{\{F\}}(\{M\}) = \frac{|\underline{F}(M)|}{|U|} = 0$.

For the case of (2), the choice of Z can be M or F. Assume Z is F. For the effect of M on H or on T, one have $P(H|M) \neq P(H|M, F)$ (e.g., $P(h_1|m_1) = \frac{1}{4}$, $P(h_1|m_1, f_1) = \frac{1}{3}$) and $P(T|M) \neq P(T|M, F)$ (e.g., $P(t_1|m_1) = \frac{1}{2}$, $P(t_1|m_1, f_1) = \frac{2}{3}$), plus $\underline{MH}(F) = \{p3\}$, $\underline{F}(MH) = \emptyset = \underline{F}(MT)$ and $\underline{MT}(F) = \{p1, p3, p4, p6\}$, then by Theorems 1 and 4, $P(h_1|do(m_1))$ and $P(t_1|do(m_1))$ are not identifiable by the adjustment formula (1) for F.

Assume Z is M. For the effect of F on H or on T, there is $P(H|F) \neq P(H|M, F)$ (e.g., $P(h_1|f_1) = \frac{1}{2}$, $P(h_1|m_1, f_1) = \frac{1}{3}$) and $P(T|F) \neq P(T|M, F)$ (e.g., $P(t_1|f_1) = \frac{1}{2}$, $P(t_1|m_1, f_1) = \frac{2}{3}$), plus $\underline{FH}(M) = \{p1, p4, p5, p6\}$, $\underline{M}(FH) = \emptyset = \underline{M}(FT)$ and $\underline{FT}(M) = \{p3, p4, p5, p6\}$, therefore by Theorems 1 and 4, $P(h_1|do(f_1))$ and $P(t_1|do(f_1))$ are not identifiable by the adjustment formula (1) for M.

Example 2. Consider another illustration of the Düntsch's decision table (Yamaguchi, 2009) (see Table 2), where $\underline{W}(Z) = \underline{Z}(W) = \underline{W}(Y) = \underline{Y}(W) = \underline{Y}(Z) = \underline{Z}(Y) = \emptyset$, namely W, Y and Z are independent from rough set theory. When $\underline{W}(Y) = \emptyset$, i.e., Y is not derivable from W, whether there is no causal links between W and Y becomes the key to the analysis. For the data in Table 2, using Definition 4, the degrees of dependency between d and c_1, c_2 are zero, i.e., $\gamma_{\{c_1\}}(\{d\}) = \frac{|\underline{c_1}(d)|}{|U|} = 0$ and $\gamma_{\{c_2\}}(\{d\}) = \frac{|\underline{c_2}(d)|}{|U|} = 0$.

Table 2
Düntsch's decision table.

U	c_1	c_2	d
x_1	0	0	0
x_2	0	2	0
x_3	0	2	0
x_4	1	1	0
x_5	1	0	1
x_6	1	2	1
x_7	1	2	1
x_8	0	1	1

Table 3
An information system of the vertebrate world.

U	Gregarious (G)	Fly (F)	Egg (E)	Lung (L)	Bird (B)	Animal name
1	*no*	*yes*	*yes*	*yes*	*yes*	Vulture, Pheasant
2	*yes*	*yes*	*yes*	*yes*	*yes*	Egret, Latham, Scoter Shelduck, Sparrow
3	*yes*	*no*	*yes*	*yes*	*yes*	Penguin, Ostrich
4	*no*	*no*	*no*	*yes*	*no*	Opossum, Mink
5	*no*	*no*	*yes*	*yes*	*no*	Toad, Platypus,Viper, Turtle
6	*no*	*no*	*yes*	*no*	*no*	Dogfish
7	*yes*	*no*	*no*	*yes*	*no*	Reindeer, Seal
8	*yes*	*no*	*yes*	*no*	*no*	Hairtail
9	*yes*	*yes*	*no*	*yes*	*no*	Fruit bat

To start with, based on rough set theory we have
$U/\{d\} = \{\{x_1, x_2, x_3, x_4\}_{[x]_{d=0}}, \{x_5, x_6, x_7, x_8\}_{[x]_{d=1}}\}$,
$U/\{c_1\} = \{\{x_1, x_2, x_3, x_8\}_{[x]_{c_1=0}}, \{x_4, x_5, x_6, x_7\}_{[x]_{c_1=1}}\}$,
$U/\{c_2\} = \{\{x_1, x_5\}_{[x]_{c_2=0}}, \{x_2, x_3, x_6, x_7\}_{[x]_{c_2=2}}, \{x_4, x_8\}_{[x]_{c_2=1}}\}$,
$\underline{d}(c_1) = \emptyset = \underline{d}(c_2)$, $\underline{c_1}(d) = \emptyset = \underline{c_2}(d)$, $\underline{c_1}(c_2) = \emptyset = \underline{c_2}(c_1)$, together with $P(c_1) = P(c_1|c_2)$, $P(c_2) = P(c_2|c_1)$, but $P(d|c_1) \neq P(d|c_1, c_2)$ (e.g., $P(d = 1|c_1 = 1) = \frac{3}{4}$, $P(d = 1|c_1 = 1, c_2 = 1) = 0$), $P(d|c_2) \neq P(d|c_2, c_1)$ (e.g., $P(d = 1|c_2 = 1) = \frac{1}{2}$, $P(d = 1|c_2 = 1, c_1 = 1) = 0$). Therefore one can compute the effect of c_1 on d by the adjustment for c_2 and the effect of c_2 on d by the adjustment for c_1. More specifically,
$P(d = 1|do(c_1 = 1)) = \sum_{c_2} P(d = 1|c_1 = 1, c_2)P(c_2) = \frac{3}{4} = P(d = 1|c_1 = 1)$,
$P(d = 1|do(c_1 = 0)) = \sum_{c_2} P(d = 1|c_1 = 0, c_2)P(c_2) = \frac{1}{4} = P(d = 1|c_1 = 0)$,
$P(d = 0|do(c_1 = 0)) = \sum_{c_2} P(d = 0|c_1 = 0, c_2)P(c_2) = \frac{3}{4} = P(d = 0|c_1 = 0)$,
$P(d = 0|do(c_1 = 1)) = \sum_{c_2} P(d = 0|c_1 = 1, c_2)P(c_2) = \frac{1}{4} = P(d = 0|c_1 = 1)$,
$P(d = 0|do(c_2 = 0)) = \sum_{c_1} P(d = 0|c_2 = 0, c_1)P(c_1) = \frac{1}{2} = P(d = 0|c_2 = 0)$,
$P(d = 1|do(c_2 = 0)) = \sum_{c_1} P(d = 1|c_2 = 0, c_1)P(c_1) = \frac{1}{2} = P(d = 1|c_2 = 0)$,
$P(d = 0|do(c_2 = 1)) = \sum_{c_1} P(d = 0|c_2 = 1, c_1)P(c_1) = \frac{1}{2} = P(d = 0|c_2 = 1)$,
$P(d = 1|do(c_2 = 1)) = \sum_{c_1} P(d = 1|c_2 = 1, c_1)P(c_1) = \frac{1}{2} = P(d = 1|c_2 = 1)$,
$P(d = 0|do(c_2 = 2)) = \sum_{c_1} P(d = 0|c_2 = 2, c_1)P(c_1) = \frac{1}{2} = P(d = 0|c_2 = 2)$,
$P(d = 1|do(c_2 = 2)) = \sum_{c_1} P(d = 1|c_2 = 2, c_1)P(c_1) = \frac{1}{2} = P(d = 1|c_2 = 2)$.

By employing Lemma 4, there exist
$\frac{2}{3} \le PN(c_1 = 1, d = 1) \le 1$, $\frac{2}{3} \le PS(c_1 = 1, d = 1) \le 1$, $0 \le PN(c_1 = 0, d = 1) \le 1$ and $0 \le PS(c_1 = 0, d = 1) \le 1$; $\frac{2}{3} \le PN(c_1 = 0, d = 0) \le 1$, $\frac{2}{3} \le PS(c_1 = 0, d = 0) \le 1$, $0 \le PN(c_1 = 1, d = 0) \le 1$ and $0 \le PS(c_1 = 1, d = 0) \le 1$;
$0 \le PN(c_2 = 0, d = 0) \le 1$, $0 \le PS(c_2 = 0, d = 0) \le \frac{1}{3}$, $0 \le PN(c_2 = 1, d = 0) \le 1$, $0 \le PS(c_2 = 1, d = 0) \le \frac{1}{3}$, $0 \le PN(c_2 = 2, d = 0) \le 1$ and $0 \le PS(c_2 = 2, d = 0) \le \frac{1}{3}$;
$0 \le PN(c_2 = 0, d = 1) \le 1$, $0 \le PS(c_2 = 0, d = 1) \le \frac{1}{3}$, $0 \le PN(c_2 = 1, d = 1) \le 1$, $0 \le PS(c_2 = 1, d = 1) \le \frac{1}{3}$, $0 \le PN(c_2 = 2, d = 1) \le 1$ and $0 \le PS(c_2 = 2, d = 1) \le \frac{1}{3}$.

It is easy to find that the above causal explanation of c_1, c_2 and d provides a more specific and more convincing argument than the attribute dependency degrees of the three $\gamma_{\{c_1\}}(\{d\}) = 0.5625$ and $\gamma_{\{c_2\}}(\{d\}) = 0.3958$ from Yamaguchi (2009) (different from the dependency in Definition 4). We see that the probability that $d = 1$ would not have occurred in the absence of $c_1 = 1$ given that $c_1 = 1$ and $d = 1$ did in fact occur is greater than $\frac{2}{3}$, in other words, there is at least 67% chance that $c_1 = 1$ was necessary for the production of $d = 1$. The probability that setting $c_1 = 1$ would produce $d = 1$ in a situation where $d \neq 1$ and $c_1 \neq 1$ in fact occur is also greater than $\frac{2}{3}$, meaning that there is at least 67% chance that $c_1 = 1$ was sufficient for the production of $d = 1$. Similarly, there is at least 67% chance that $c_1 = 0$ was necessary and sufficient for the production of $d = 0$. Note that under the condition of no-confounding ($P(y|do(x)) = P(y|x)$) the lower bound for *PN* often needs to meet the $PN > \frac{1}{2}$ criterion (Tian & Pearl, 2000), and $0 \le PS \le \frac{1}{2}$ means that the probabilities of causation cannot be determined from statistical data (Pearl, 2009). Accordingly, the causal relation between c_2 and d as well as $c_1 = 0$ ($c_1 = 1$) and $d = 1$ ($d = 0$) cannot be determined from Table 2.

Example 3. Take another classic example of information system on vertebrates with 20 different animals from Wang et al. (2003) shown in Table 3, where *Bird* is the decision attribute, {*Gregarious, Fly, Egg, Lung*} are the condition attributes.

Referring to Table 3, let $F = f_1$ and $F = f_0$ represent, respectively, it can fly and cannot fly, $B = b_1$ and $B = b_0$ represent it is a bird and not a bird, $G = g_1$ and $G = g_0$ represent it is gregarious and not gregarious, $E = e_1$ and $E = e_0$ represent laying eggs and not laying eggs and $L = l_1$ and $L = l_0$ represent using lung for breathing and not using lung for breathing. Based on rough set theory, it is obvious that:
$U/\{Bird\} = \{\{1, 2, 3\}_{[x]_{b_1}}, \{4, 5, 6, 7, 8, 9\}_{[x]_{b_0}}\}$,
$U/\{Fly\} = \{\{1, 2, 9\}_{[x]_{f_1}}, \{3, 4, 5, 6, 7, 8\}_{[x]_{f_0}}\}$,
$U/\{Egg\} = \{\{1, 2, 3, 5, 6, 8\}_{[x]_{e_1}}, \{4, 7, 9\}_{[x]_{e_0}}\}$,
$U/\{Lung\} = \{\{1, 2, 3, 4, 5, 7, 9\}_{[x]_{l_1}}, \{6, 8\}_{[x]_{l_0}}\}$,
$U/\{Gregarious\} = \{\{1, 4, 5, 6\}_{[x]_{g_0}}, \{2, 3, 7, 8, 9\}_{[x]_{g_1}}\}$,
$U/\{Egg, Lung\} = \{\{1, 2, 3, 5\}_{[x]_{e_1, l_1}}, \{6, 8\}_{[x]_{e_1, l_0}}, \{4, 7, 9\}_{[x]_{e_0, l_1}}\}$,
$U/\{Gregarious, Egg\} = \{\{2, 3, 8\}_{[x]_{e_1, g_1}}, \{4\}_{[x]_{e_0, g_0}}, \{1, 5, 6\}_{[x]_{e_1, g_0}}, \{7, 9\}_{[x]_{e_0, g_1}}\}$,
$U/\{Gregarious, Lung\} = \{\{2, 3, 7, 9\}_{[x]_{l_1, g_1}}, \{1, 4, 5\}_{[x]_{l_1, g_0}}, \{6\}_{[x]_{l_0, g_0}}, \{8\}_{[x]_{l_0, g_1}}\}$,
$U/\{Gregarious, Lung, Egg\} = \{\{1, 5\}_{[x]_{l_1, e_1, g_0}}, \{2, 3\}_{[x]_{l_1, e_1, g_1}}, \{4\}_{[x]_{l_1, e_0, g_0}}, \{6\}_{[x]_{l_0, e_1, g_0}}, \{7, 9\}_{[x]_{l_1, e_0, g_1}}, \{8\}_{[x]_{l_0, e_1, g_1}}\}$.

By Theorems 1 and 2, we obtain

(1) $\underline{Fly}(Gregarious) = \emptyset = \underline{Gregarious}(Fly)$ with $P(F) \neq P(F|G)$ (e.g., $P(f_1) = \frac{8}{20}$, $P(f_1|g_1) = \frac{6}{11}$) and $P(G) \neq P(G|F)$ (e.g., $P(g_1) = \frac{11}{20}$, $P(g_1|f_1) = \frac{6}{8}$), $P(B|F) \neq P(B|F, G)$ and $P(B|G) \neq P(B|G, F)$ (e.g., $P(b_1|f_1) = \frac{7}{8}$, $P(b_1|f_1, g_1) = \frac{5}{6}$, $P(b_1|g_1) = \frac{7}{11}$);
(2) $\underline{Fly}(Egg) = \emptyset = \underline{Egg}(Fly)$ with $P(F) \neq P(F|E)$ (e.g., $P(f_1|e_1) = \frac{7}{15}$) and $P(E) \neq P(E|F)$ (e.g., $P(e_1) = \frac{15}{20}$, $P(e_1|f_1) = \frac{7}{8}$), $P(B|F) \neq P(B|F, E)$ and $P(B|E) \neq P(B|E, F)$ (e.g., $P(b_1|f_1, e_1) = 1$, $P(b_1|e_1) = \frac{9}{15}$);
(3) $\underline{Fly}(Bird) = \emptyset = \underline{Bird}(Fly)$ with $P(F) \neq P(F|B)$ (e.g., $P(f_1|b_1) = \frac{7}{9}$) and $P(B) \neq P(B|F)$ (e.g., $P(b_1) = \frac{9}{20}$, $P(b_1|f_1) = \frac{7}{8}$), $P(G|B, F) \neq P(G|B) \neq P(G|F)$ (e.g., $P(g_1|b_1, f_1) = \frac{5}{7}$, $P(g_1|f_1) = \frac{6}{11}$, $P(g_1|b_1) = \frac{7}{9}$), $P(E|B, F) \neq P(E|B) \neq P(E|F)$ (e.g., $P(e_1|f_1, b_0) = 0$,

$P(e_1|f_1) = \frac{7}{8}$, $P(e_1|b_0) = \frac{6}{11}$), $P(L|B, F) \neq P(L|B) \neq P(L|F)$ (e.g., $P(l_1|f_0, b_0) = \frac{4}{5}$, $P(l_1|f_0) = \frac{5}{6}$, $P(l_1|b_0) = \frac{9}{11}$);

(4) $\underline{Gregarious}(Bird) = \emptyset = \underline{Bird}(Gregarious)$ with $P(G) \neq P(G|B)$ (e.g., $P(g_1) = \frac{11}{20}$, $P(g_1|b_1) = \frac{7}{9}$) and $P(B) \neq P(B|G)$ (e.g., $P(b_1) = \frac{9}{20}$, $P(b_1|g_1) = \frac{7}{11}$), $P(F|B) \neq P(F|B, G)$ and $P(F|G) \neq P(F|G, B)$ (e.g., $P(f_1|b_1) = \frac{7}{9}$, $P(f_1|b_1, g_1) = \frac{5}{7}$, $P(f_1|g_1) = \frac{6}{11}$), $P(E|B, G) \neq P(E|B) \neq P(E|G)$ (e.g., $P(e_1|b_0) = \frac{6}{11}$, $P(e_1|g_1) = \frac{8}{11}$, $P(e_1|b_0, g_1) = \frac{1}{4}$), $P(L|B, G) \neq P(L|B) \neq P(L|G)$ (e.g., $P(l_1|b_0) = \frac{9}{11}$, $P(l_1|g_1) = \frac{10}{11}$,$P(l_1|b_0, g_1) = \frac{3}{4}$).

(5) $\underline{Egg}(Gregarious) = \emptyset = \underline{Gregarious}(Egg)$, $P(G) \neq P(G|E)$, $P(B|G, E) \neq P(B|G)$, $P(B|G, E) \neq P(B|E)$ (e.g., $P(b_1|g_1) = \frac{7}{11}$, $P(b_1|e_1) = \frac{9}{15}$, $P(b_1|g_1, e_1) = \frac{7}{8}$);

(6) $\underline{Lung}(Gregarious) = \emptyset = \underline{Gregarious}(Lung)$, $P(G) \neq P(G|L)$, $P(B|G, L) \neq P(B|G)$, $P(B|G, L) \neq P(B|L)$ (e.g., $P(b_1|g_1) = \frac{7}{11}$, $P(b_1|l_1) = \frac{7}{18}$, $P(b_1|g_1, l_1) = \frac{7}{10}$).

Further, according to Theorem 4, we conclude that

1. the effect $P(b|do(f))$ is not identifiable by the adjustment formula for *Gregarious* in view of $\underline{Fly, Bird}(Gregarious) = \{3, 9\}$ and $\underline{Gregarious}(Fly, Bird) = \emptyset$.
2. the effect $P(b|do(f))$ is not identifiable by the adjustment formula for *Egg* due to $\underline{Fly, Bird}(Egg) = \{1, 2, 3, 9\}$ and $\underline{Egg}(Fly, Bird) = \emptyset$.
3. the effects $P(g|do(f))$, $P(e|do(f))$ and $P(l|do(f))$ are not identifiable by the adjustment formula for *Bird* due to $\underline{Fly, Gregarious}(Bird) = \{1, 4, 5, 6\}$, $\underline{Fly, Egg}(Bird) = \{1, 2, 4, 7, 9\}$, $\underline{Fly, Lung}(Bird) = \{6, 8\}$ and $\underline{Bird}(Fly, Gregarious) = \emptyset = \underline{Bird}(Fly, Egg) = \underline{Bird}(Fly, Lung)$;
   the effects $P(b|do(g))$, $P(g|do(b))$, $P(b|do(e))$, $P(e|do(b))$ and $P(l|do(b))$ are not identifiable by the adjustment formula for *Fly* due to $\underline{Bird, Gregarious}(Fly) = \{1, 4, 5, 6\}$, $\underline{Bird, Egg}(Fly) = \{5, 6, 8\}$, $\underline{Bird, Lung}(Fly) = \{6, 8\}$ and $\underline{Fly}(Bird, Gregarious) = \emptyset = \underline{Fly}(Bird, Egg) = \underline{Fly}(Bird, Lung)$.
4. the effects $P(f|do(g))$, $P(e|do(g))$ and $P(l|do(g))$ are not identifiable by the adjustment formula for *Bird* because of $\underline{Gregarious, Fly}(Bird) = \{1, 4, 5, 6\}$, $\underline{Gregarious, Egg}(Bird) = \{4, 7, 9\}$, $\underline{Gregarious, Lung}(Bird) = \{6, 8\}$ and $\underline{Bird}(Gregarious, Fly) = \emptyset = \underline{Bird}(Gregarious, Egg) = \underline{Bird}(Gregarious, Lung)$;
   $P(f|do(b))$ is not identifiable by the adjustment formula for *Gregarious* $\underline{Bird, Fly}(Gregarious) = \{3, 9\}$ and $\underline{Gregarious}(Bird, Fly) = \emptyset$;
   $P(e|do(b))$, $P(b|do(e))$, $P(b|do(l))$ and $P(l|do(b))$ are not identifiable via the adjustment formula for *Gregarious* with confounding bias due to $[x]_{b_1, e_0} = \emptyset = [x]_{b_1, l_0}$ (which implies $P(b_1, e_0, g) = 0 = P(b_1, l_0, g)$) in spite of $\underline{Gregarious}(Bird, Egg) = \emptyset = \underline{Bird, Egg}(Gregarious)$ and $\underline{Gregarious}(Bird, Lung) = \emptyset = \underline{Bird, Lung}(Gregarious)$.
5. the effect $P(b|do(g))$ is not identifiable by the adjustment formula for *Egg* and *Lung* due to $\underline{Bird, Gregarious}(Egg) = \{1, 2, 3\} = \underline{Bird, Gregarious}(Lung)$, $\underline{Egg}(Bird, Gregarious) = \emptyset = \underline{Lung}(Bird, Gregarious)$.

To put it another way, when omitting the number of the animals in the column of "Animal name" and recalculating the probabilities, we have $P(F) = P(F|E)$ but $P(B|F, E) \neq P(B|F) \neq P(B|E)$ (e.g., $P(b_1|f_1, e_1) = 1$, $P(b_1|f_1) = \frac{2}{3}$, $P(b_1|e_1) = \frac{1}{2}$), $\underline{Fly}(Bird) \cap \underline{Bird}(Egg) = \emptyset \cap \{1, 2, 3\} = \emptyset$, $\underline{Fly}(Bird) = \emptyset = \underline{Bird}(Fly)$, $\underline{Egg}(Bird) \cap \underline{Bird}(Egg) = \{4, 7, 9\} \cap \{1, 2, 3\} = \emptyset$, $\underline{Egg}(Bird) \cap \underline{Bird}(Fly) = \{4, 7, 9\} \cap \emptyset = \emptyset$, $\underline{Egg}(Bird) \cap \underline{Bird}(Egg) = \emptyset$, $\underline{Fly}(Bird) \cap \underline{Bird}(Fly) = \emptyset$, and further, using Theorems 1 and 2, one can calculate the causal effect of *Fly* on *Bird* without confounding bias via the adjustment for *Egg*, namely $P(b_1|do(f_1)) = \sum_E P(b_1|f_1, E)P(E) = P(b_1|f_1, e_1)P(e_1) + P(b_1|f_1, e_0)P(e_0) = \frac{2}{3} = P(b_1|f_1)$ and the causal effect of *Egg* on *Bird* without confounding bias via the adjustment for *Fly*, i.e., $P(b_1|do(e_1)) = \sum_F P(b_1|e_1, F)P(F) = P(b_1|e_1, f_1)P(f_1) + P(b_1|e_1, f_0)P(f_0) = \frac{1}{2} = P(b_1|e_1)$. Accordingly, one get

$\frac{3}{4} \leq PN(f_1, b_1) = P(B_{f_0} = b_0|F = f_1, B = b_1) \leq 1$,
$\frac{3}{5} \leq PS(f_1, b_1) = P(B_{f_1} = b_1|F = f_0, B = b_0) \leq \frac{4}{5}$,
$PN(e_1, b_1) = P(B_{e_0} = b_0|E = e_1, B = b_1) = 1$,
$PS(e_1, b_1) = P(B_{e_1} = b_1|E = e_0, B = b_0) = \frac{1}{2}$,
$PN(e_0, b_0) = P(B_{e_1} = b_1|E = e_0, B = b_0) = \frac{1}{2}$,
$PS(e_0, b_0) = P(B_{e_0} = b_0|E = e_1, B = b_1) = 1$.

Further the degree of dependency between *Bird* and *Fly* is zero, i.e., $\gamma_{\{Fly\}}(\{Bird\}) = \frac{|\underline{Fly}(Bird)|}{|U|} = 0 = \gamma_{\{Bird\}}(\{Fly\})$ due to $\underline{Fly}(Bird) = \emptyset = \underline{Bird}(Fly)$. The degree of dependency between *Bird* and *Egg* is $k = \gamma_{\{Egg\}}(\{Bird\}) = \frac{|\underline{Egg}(Bird)|}{|U|} = \frac{1}{3}$, i.e., $k = \gamma_{e_1}(b_1) = \frac{|\underline{e_1}(b_1)| = |[x]_{e_1} \subseteq [x]_{b_1}|}{|U|} = 0$, $k = \gamma_{e_0}(b_0) = \frac{|\underline{e_0}(b_0)| = |[x]_{e_0} \subseteq [x]_{b_0}|}{|U|} = \frac{1}{3}$.

Using the proposed results to the data in Table 3, we get the interventional interpretation and counterfactual analysis for the conditional attributes *Fly* and *Egg* and the decision attribute *Bird*, learning more information than that conveyed by attribute dependency. Additionally we find from Table 3 that the column of "Animal name" has no effect on the lower approximation but can affect statistical independence. This might relate to the presence of the possible sample selection bias (Bareinboim & Pearl, 2016; Bareinboim & Tian, 2015) and the results given by omitting "Animal name" column may be viewed as a reference in data analysis and inference.

Example 4. Take another dataset for fitting contact lenses from the UCI Machine Learning repository. The detailed description of each attribute is as follows (see Table 4):

Age of the patient -a: it has the following values: (1) young, (2) pre-presbyopic, (3) presbyopic.
Spectacle prescription -b: it has the following values: (1) myope, (2) hypermetrope.
Astigmatic -c: it has the following values: (1) no, (2) yes.
Tear production rate -d: it has the following values: (1) reduced, (2) normal.
Decision -e: it is a decision attribute that contains the following values: (1) hard contact lenses, (2) soft contact lenses, (3) no contact lenses.

According to Definitions 3 and 4, Theorems 1 and 2, it is easily obtained that $\underline{a}(b) = \emptyset = \underline{b}(a)$, $\underline{a}(c) = \emptyset = \underline{c}(a)$, $\underline{a}(d) = \emptyset = \underline{d}(a)$, $P(a) = P(a|b)$, $P(a) = P(a|c)$, $P(a) = P(a|d)$, $P(e|ab) \neq P(e|a) \neq P(e|b)$, $P(e|ac) \neq P(e|a) \neq P(e|c)$, $P(e|ad) \neq P(e|a) \neq P(e|d)$, $\underline{a}(e) = \emptyset = \underline{e}(a)$, $\underline{e}(b) = \emptyset = \underline{b}(e)$, $\underline{c}(e) = \emptyset$, $\underline{e}(c) \neq \emptyset$, $\underline{d}(e) = \emptyset$, $\underline{e}(d) \neq \emptyset$, $\gamma_{\{a\}}(\{e\}) = 0 = \gamma_{\{b\}}(\{e\}) = \gamma_{\{c\}}(\{e\}) = \gamma_{\{d\}}(\{e\})$. Further we get the unbiased estimate of $P(e|do(a))$, $P(e|do(b))$, $P(e|do(c))$ and $P(e|do(d))$. By Definition 10 and Lemma 4, plus $PN > \frac{1}{2}$ and $PS > \frac{1}{2}$, it is concluded that

$PN(a = 1, e = 1) \in [\frac{1}{2}, 1]$, $PS(a = 1, e = 1) \in [\frac{1}{7}, \frac{2}{7}]$; $PN(b = 1, e = 1) \in [\frac{2}{3}, 1]$, $PS(b = 1, e = 1) \in [\frac{2}{11}, \frac{3}{11}]$;

$PN(c = 2, e = 1) = 1$, $PS(c = 2, e = 1) = \frac{1}{3}$, $PN(c = 1, e = 2) = 1$, $PS(c = 1, e = 2) = \frac{5}{12}$;

$PN(d = 2, e = 1) = 1$, $PS(d = 2, e = 1) = \frac{1}{3}$, $PN(d = 2, e = 2) = 1$, $PS(d = 2, e = 2) = \frac{5}{12}$, $PN(d = 1, e = 3) = \frac{3}{4}$, $PS(d = 1, e = 3) = 1$.

For the combination of attributes, it is easily found that $\underline{ab}(c) = \emptyset = \underline{c}(ab)$, $\underline{ab}(d) = \emptyset = \underline{d}(ab)$, $\underline{abc}(d) = \emptyset = \underline{d}(abc)$, $P(ab) = P(ab|c) = P(ab|d)$, $P(abc) = P(abc|d)$, $P(e|ab, c) \neq P(e|ab)$, $P(e|ab, d) \neq P(e|ab)$, $P(e|abc, d) \neq P(e|abc)$, $\underline{ab}(e) = \emptyset = \underline{e}(ab)$, $\underline{abc}(e) \neq \emptyset$, $\underline{e}(abc) = \emptyset$, $\underline{abc}(e) \cap \underline{e}(d) = \emptyset$, $\underline{d}(e) = \emptyset = \underline{e}(abc)$, $\underline{bc}(a) = \emptyset = \underline{a}(bc)$, $\underline{bc}(e) = \emptyset = \underline{e}(bc)$, $P(bc) = P(bc|a)$, $P(e|bc, a) \neq P(e|bc)$, $\underline{bcd}(a) = \emptyset = \underline{a}(bcd)$, $\underline{bcd}(e) \neq \emptyset$, $\underline{e}(bcd) = \emptyset$,

Table 4
Decision table for fitting contact lenses.

U	a	b	c	d	e	U	a	b	c	d	e	U	a	b	c	d	e
1	1	1	2	2	1	9	3	2	1	2	2	17	2	2	2	1	3
2	1	2	2	2	1	10	1	1	1	1	3	18	2	2	2	2	3
3	2	1	2	2	1	11	1	1	2	1	3	19	3	1	1	1	3
4	3	1	2	2	1	12	1	2	1	1	3	20	3	1	1	2	3
5	1	1	1	2	2	13	1	2	2	1	3	21	3	1	2	1	3
6	1	2	1	2	2	14	2	1	1	1	3	22	3	2	1	1	3
7	2	1	1	2	2	15	2	1	2	1	3	23	3	2	2	1	3
8	2	2	1	2	2	16	2	2	1	1	3	24	3	2	2	2	3

$P(bcd) = P(bcd|a)$, $P(e|bcd,\ a) \neq P(e|bcd)$, $\underline{bd}(a) = \emptyset = \underline{a}(bd)$, $\underline{bd}(e) \neq \emptyset$, $\underline{e}(bd) = \emptyset$, $P(bd) = P(bd|a)$, $P(e|bd,\ a) \neq P(e|bd)$, $\underline{cd}(a) = \emptyset = \underline{a}(cd)$, $\underline{cd}(e) \cap \underline{e}(cd) = \emptyset$, $\underline{a}(e) \cap \underline{e}(cd) = \emptyset$, $P(cd) = P(cd|a)$, $P(e|cd,\ a) \neq P(e|cd)$, $\gamma_{\{ab\}}(\{e\}) = 0 = \gamma_{\{bc\}}(\{e\})$, $\gamma_{\{bd\}}(\{e\}) = \frac{1}{2}$, $\gamma_{\{cd\}}(\{e\}) = \frac{1}{2}$, $\gamma_{\{abc\}}(\{e\}) = \frac{1}{4}$, $\gamma_{\{abd\}}(\{e\}) = \frac{1}{2}$, $\gamma_{\{bcd\}}(\{e\}) = \frac{3}{4}$, and further the unbiased estimate of $P(e|do(ab))$, $P(e|do(abc))$, $P(e|do(bc))$, $P(e|do(bd))$, $P(e|do(cd))$ and $P(e|do(bcd))$ can be identified. Using Definition 10 and Lemma 4, plus $PN > \frac{1}{2}$ and $PS > \frac{1}{2}$, there is

$PN(a = 1 \quad b = 1, e = 1) \in [0.408, 1]$, $PS(a = 1 \quad b = 1, e = 1) \in [\frac{2}{17}, \frac{5}{17}]$, the same ranges for $PN(a = 1\ b = 2, e = 1)$ and $PS(a = 1\ b = 2, e = 1)$, $PN(a = 2\ b = 1, e = 1)$ and $PS(a = 2\ b = 1, e = 1)$, $PN(a = 3\ b = 1, e = 1)$ and $PS(a = 3\ b = 1, e = 1)$;

$PN(a = 1\ \ b = 1\ \ c = 2, e = 1) \in [0.744, 1]$, $PS(a = 1\ \ b = 1\ \ c = 2, e = 1) \in [\frac{8}{19}, \frac{11}{19}]$, the same ranges for $PN(a = 1\ b = 2\ c = 2, e = 1)$ and $PS(a = 1\ b = 2\ c = 2, e = 1)$, $PN(a = 2\ b = 1\ c = 2, e = 1)$ and $PS(a = 2\ b = 1\ c = 2, e = 1)$, $PN(a = 3\ b = 1\ c = 2, e = 1)$ and $PS(a = 3\ b = 1\ c = 2, e = 1)$;

$PN(a = 1\ \ b = 1\ \ c = 1, e = 2) \in [0.624, 1]$, $PS(a = 1\ \ b = 1\ \ c = 1, e = 2) \in [\frac{7}{18}, \frac{11}{18}]$, with the equal bounds for $PN(a = 1\ b = 2\ c = 1, e = 2)$ and $PS(a = 1\ b = 2\ c = 1, e = 2)$, $PN(a = 2\ b = 1\ c = 1, e = 2)$ and $PS(a = 2\ b = 1\ c = 1, e = 2)$, $PN(a = 2\ b = 2\ c = 1, e = 2)$ and $PS(a = 2\ b = 2\ c = 1, e = 2)$, $PN(a = 3\ b = 2\ c = 1, e = 2)$ and $PS(a = 3\ b = 2\ c = 1, e = 2)$;

$PN(a = 2\ b = 2\ c = 2, e = 3) \approx 0.408$, $PS(a = 2\ b = 2\ c = 2, e = 3) = 1$, with the equal bounds for $PN(a = 3\ b = 2\ c = 2, e = 3)$ and $PS(a = 3\ b = 2\ c = 2, e = 3)$, $PN(a = 3\ b = 1\ c = 1, e = 3)$ and $PS(a = 3\ b = 1\ c = 1, e = 3)$, $PN(a = 1\ b = 1\ d = 1, e = 3)$ and $PS(a = 1\ b = 1\ d = 1, e = 3)$, $PN(a = 1\ b = 2\ d = 1, e = 3)$ and $PS(a = 1\ b = 2\ d = 1, e = 3)$, $PN(a = 2\ b = 1\ d = 1, e = 3)$ and $PS(a = 2\ b = 1\ d = 1, e = 3)$, $PN(a = 2\ b = 2\ d = 1, e = 3)$ and $PS(a = 2\ b = 2\ d = 1, e = 3)$, $PN(a = 3\ b = 1\ d = 1, e = 3)$ and $PS(a = 3\ b = 1\ d = 1, e = 3)$, $PN(a = 3\ b = 2\ d = 1, e = 3)$ and $PS(a = 3\ b = 2\ d = 1, e = 3)$;

$PN(a = 1\ \ b = 1\ \ d = 2, e = 2) \in [0.624, 1]$, $PS(a = 1\ \ b = 1\ \ d = 2, e = 2) \in [\frac{7}{18}, \frac{11}{18}]$, the same ranges for $PN(a = 1\ b = 2\ d = 2, e = 2)$ and $PS(a = 1\ b = 2\ d = 2, e = 2)$, $PN(a = 2\ b = 1\ d = 2, e = 2)$ and $PS(a = 2\ b = 1\ d = 2, e = 2)$, $PN(a = 2\ b = 2\ d = 2, e = 2)$ and $PS(a = 2\ b = 2\ d = 2, e = 2)$, $PN(a = 3\ b = 2\ d = 2, e = 2)$ and $PS(a = 3\ b = 2\ d = 2, e = 2)$;

$PN(a = 1\ \ b = 1\ \ d = 2, e = 1) \in [0.744, 1]$, $PS(a = 1\ \ b = 1\ \ d = 2, e = 1) \in [\frac{8}{19}, \frac{11}{19}]$, the same ranges for $PN(a = 1\ b = 2\ d = 2, e = 1)$ and $PS(a = 1\ b = 2\ d = 2, e = 1)$, $PN(a = 2\ b = 1\ d = 2, e = 1)$ and $PS(a = 2\ b = 1\ d = 2, e = 1)$, $PN(a = 3\ b = 1\ d = 2, e = 1)$ and $PS(a = 3\ b = 1\ d = 2, e = 1)$;

$PN(b = 1\ c = 2, e = 1) \in [\frac{8}{9}, 1]$, $PS(b = 1\ c = 2, e = 1) \in [\frac{8}{17}, \frac{9}{17}]$, $PN(b = 2\ c = 1, e = 2) \in [0.776, 1]$, $PS(b = 2\ c = 1, e = 2) \in [\frac{7}{16}, \frac{9}{16}]$;

$PN(b = 1\ c = 2\ d = 2, e = 1) = 0.952$, $PS(b = 1\ c = 2\ d = 2, e = 1) = 1$, $PN(b = 2\ c = 1\ d = 2, e = 2) = 0.904$, $PS(b = 1\ c = 1\ d = 2, e = 2) = 1$;

$PN(b = 1\ c = 1\ d = 1, e = 3) = 0.432$, $PS(b = 1\ c = 1\ d = 1, e = 3) = 1$, with the equal bounds for $PN(b = 1\ c = 2\ d = 1, e = 3)$ and $PS(b = 1\ c = 2\ d = 1, e = 3)$, $PN(b = 2\ c = 1\ d = 1, e = 3)$ and $PS(b = 2\ c = 1\ d = 1, e = 3)$, $PN(b = 2\ c = 2\ d = 1, e = 3)$ and $PS(b = 2\ c = 2\ d = 1, e = 3)$;

$PN(b = 1\ d = 1, e = 3) = 0.5$, $PS(b = 1\ d = 1, e = 3) = 1$, $PN(b = 2\ \ d = 1, e = 3) = 0.5$, $PS(b = 2\ \ d = 1, e = 3) = 1$, $PN(b = 1\ \ d = 2, e = 1) = \in [\frac{8}{9}, 1]$, $PS(b = 1\ d = 2, e = 1) \in [\frac{8}{17}, \frac{9}{17}]$, $PN(b = 2\ d = 2, e = 2) = \in [\frac{7}{9}, 1]$, $PS(b = 2\ d = 2, e = 2) \in [\frac{7}{16}, \frac{9}{16}]$;

$PN(c = 2\ d = 1, e = 3) = 0.5$, $PS(c = 2\ d = 1, e = 3) = 1$, $PN(c = 1\ \ d = 1, e = 3) = 0.5$, $PS(c = 1\ \ d = 1, e = 3) = 1$, $PN(c = 1\ \ d = 2, e = 2) = 1$, $PS(c = 1\ d = 2, e = 2) = \frac{15}{18}$, $PN(c = 2\ d = 2, e = 1) = 1$, $PS(c = 2\ d = 2, e = 1) = \frac{12}{18}$.

Results from Table 4 show us the interpretation of causation between the individual condition attributes, their combinations and the decision attribute. If the causation between an individual condition attribute and the decision attribute can be identified, then the causation between its combinations and this decision also can be generally identified. Meanwhile, data in Table 4 have discrete values and are obtained from UCI Machine Learning Repository where exist many data sets containing continuous attributes. In this case, we need to first obtain a new decision table by discretization to convert real-valued attributes into nominal/symbolic ones, since rough set theory is good at dealing with discrete attributes. After discretization, the theorems become available for the data sets.

To summarize: we conclude from these examples that the estimate of the causal quantity is significantly different from the degree of dependency between variables defined in the context of rough sets. In terms of the specific attribute values, causal quantity measures the relation between the attributes (i.e., the necessary and sufficient causes of one attribute to produce the other attribute) compared to the way of dependency (i.e., the derivability of one attribute from the other attribute). When the degree of dependency comes to zero, the causal effect of the occurrence of one on the occurrence of the other might exist. This issue relates closely to the inclusion operation "$\subseteq$" of the lower approximation used in defining the dependency, which is a stronger condition (i.e., the asymmetry holds unless the equivalence classes induced by the attributes are equal) in comparison with "$\cap$" operation of conditional probability. Concerning the correlation between the variables, dependency and conditional probability from rough set theory do the same thing but from a different viewpoint. However, causal explanation for variables gets deep to the heart of data analysis, inference, decision making, machine intelligence research and so on. Research also has found that properties of the lower approximation can steer us towards the identification of causal effects under the intervention via the adjustment formula with/without confounding bias for the attributes in data tables. The role of the lower approximation lies in identifying the variables which can be used to establish the graphical structure (embedded in the back-door criterion and stable no-confounding criterion) through the dependency between variables, considering that rough sets operate directly on the attribute-value table and there is no graph in the original data table while such criteria work straight on the (directed) acyclic graph. As the examples exhibited, not all the causation of attributes can be estimated using the proposed results. This depends mainly on the dissatisfaction with the conditions for the use of the adjustment formula when setting the covariates (having the potential for adjustment) as a single variable. How to tackle this issue requires further efforts, for example, study the situation where graph structures embedded in the back-door criterion with the covariates (a subset of which can be selected for adjustment) viewed as a set of attribute variables in the rough set framework by referring to the work on Pearl's back-door criterion (Maathuis & Colombo, 2015; Pearl, 2009) and the work on the graphs involved in this criterion (Enright, Madden, & Madden, 2013; Ramasso & Denoeux, 2014) or develop the representation of the notions of intervention and counterfactual in the pure rough set terminology by drawing inspiration possibly from the "*do*" operator and its generalizations (Benferhat, 2010; Boukhris et al., 2013; Pearl, 2009) along with the application of rough sets to directed graphs (Chiaselotti, Ciucci, Gentile, & Infusino, 2017; Chiaselotti, Gentile, & Infusino, 2018).

6. Conclusions

The causal explanations behind data play an important role in problem solving, everyday reasoning, judgment, and communication about the world. Pearl's profound insight of the combination of graphical models and probability achieves mathematical computation of causality. Following this idea, we provide conditions for estimating cause–effect relationships from an attribute-value information system with complete data and discrete values by fusing the lower approximation of the set into the mathematical model of causality. Theorem 1 presents the conditions of uncovering causal effects under the intervention with no-confounding bias by the adjustment formula, which extends the back-door adjustment with a single covariate from rough sets' angle. It is strengthened by Theorems 2 and 3, verifying that the combination of the lower approximation and the probability can be used to decide what measurements are needed for identifiability. Theorem 4 depicts the conditions of the effect estimate under the intervention with confounding bias by the adjustment formula. Since rough sets do not use any parameters outside data and do reason starting with the primitive data, the method developed in this paper can help to understand and manipulate the identification of causation about attribute-value tables. Also this is a simple attempt to analyze causality in rough set theory by applying causal quantities of causal inference, which expands the applicability of structural causal models, and paves the way for portraying interventions and counterfactuals in rough set terms, which enriches the content of rough set theory.

Causality is shown in Lake, Ullman, Tenenbaum, and Gershman (2017) as one of the core ingredients for building AI (artificial intelligence) systems with more human-like learning and thought, and the machine learning systems in Pearl (2019) that can reason about interventions and counterfactuals can serve as the basis for strong AI. We close this section by presenting some speculative thoughts on the future research. An interesting topic is to apply the results of identifying causation between attributes in the process of learning from examples, i.e., learning characteristic features of each concept from features of all given instances of the concept and deriving decision rules from the given examples. The issue of powerful understanding for AI systems, such as understanding scenes and generating natural language captions, is a promising research direction that might be solved by employing the proposed results to glue the objects and their use together or to relate the actions with the goals and intentions. Another topic is to use the proposed tools to creativity action and concept approximation, such as constructing new concepts by gluing some existing concepts with coherence and purpose. It is also promising to infuse the proposed tools for causality within the rough set framework into the intelligent systems based on granular computing (Pedrycz, 2013) and the existing machine learning models. When the attribute-value data table is given with missing data, the discovering of causality between attributes is a challenging work, along with finding how lack of knowledge would affect the systems' ability to learn. Another worthy directions include the impact of newly observed instances on the already acquired causal knowledge, the possible integration of the proposed results of causation into emotional research for intelligent machines, the measure of the complexity of time and space for the proposed tools when the number of attributes within incomplete/complete data table becomes very large, and other related issues to AI system research.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Credit authorship contribution statement

Ning Yao: Conceptualization, Methodology, Validation, Formal analysis, Data curation, Writing - original draft, Writing - review & editing. **Duoqian Miao:** Conceptualization, Formal analysis, Writing - review & editing, Supervision, Funding acquisition. **Witold Pedrycz:** Methodology, Formal analysis, Supervision, Writing - review & editing, Validation. **Hongyun Zhang:** Formal analysis, Writing - review & editing, Funding acquisition. **Zhifei Zhang:** Writing - review & editing, Data curation, Investigation.

Acknowledgments

We are very grateful to the editors and two anonymous reviewers for their incisive comments and for providing insightful and extremely helpful suggestions on the early version of this manuscript to improve our work. This research was supported in parts by the National Key R&D Program of China (grant number 213); National Natural Science Foundation of China (grant numbers 61673301, 61573255, 61573259, 61673299, 61603278); and Major Project of Ministry of Public Security (grant number 20170004).

References

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences, 113*(27), 7345–7352.
Bareinboim, E., & Tian, J. (2015). Recovering causal effects from selection bias. In *Proceedings of the 29th AAAI conference on artificial intelligence* (pp. 3475–3481). AAAI Press.
Benferhat, S. (2010). Interventions and belief change in possibilistic graphical models. *Artificial Intelligence, 174*(2), 177–189. doi:10.1016/j.artint.2009.11.012.
Benferhat, S., & Smaoui, S. (2007). Possibilistic causal networks for handling interventions: a new propagation algorithm. In *Proceedings of the 22nd national conference on artificial intelligence* (pp. 373–378). AAAI Press.
Boukhris, I., Elouedi, Z., & Benferhat, S. (2013). Dealing with external actions in belief causal networks. *International Journal of Approximate Reasoning, 54*(8), 978–999. doi:10.1016/j.ijar.2013.02.009.
Cendrowska, J. (1987). PRISM: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies, 27*(4), 349–370. doi:10.1016/S0020-7373(87)80003-2.
Chiaselotti, G., Ciucci, D., Gentile, T., & Infusino, F. (2017). Rough set theory and digraphs. *Fundamenta Informaticae, 153*(4), 291–325. doi:10.3233/FI-2017-1542.
Chiaselotti, G., Gentile, T., & Infusino, F. (2018). Granular computing on information tables: Families of subsets and operators. *Information Sciences, 442*, 72–102. doi:10.1016/j.ins.2018.02.046.
Ciucci, D., Chiaselotti, G., Gentile, T., & Infusino, F. (2016). Generalizations of rough set tools inspired by graph theory. *Fundamenta Informaticae, 148*(1–2), 207–227. doi:10.3233/FI-2016-1431.
Düntsch, I., & Gediga, G. (1997). Statistical evaluation of rough set dependency analysis. *International Journal of Human-Computer Studies, 46*(5), 589–604. doi:10.1006/ijhc.1996.0105.

Düntsch, I., & Gediga, G. (2001). Roughian: Rough information analysis. *International Journal of Intelligent Systems, 16*(1), 121–147. doi:10.1002/1098-111X(200101)16:13.0.CO;2-Z.

Enright, C. G., Madden, M. G., & Madden, N. (2013). Bayesian networks for mathematical models: Techniques for automatic construction and efficient inference. *International Journal of Approximate Reasoning, 54*(2), 323–342. doi:10.1016/j.ijar.2012.10.004.

Lake, B., Ullman, T., Tenenbaum, J., & Gershman, S. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences, 40*, e253. doi:10.1017/S0140525X16001837.

Lingras, P., & Haider, F. (2015). Partially ordered rough ensemble clustering for multigranular representations. *Intelligent Data Analysis, 19*(s1), S103–S116. doi:10.3233/IDA-150772.

Maathuis, M. H., & Colombo, D. (2015). A generalized back-door criterion. *The Annals of Statistics, 43*(3), 1060–1088. doi:10.1214/14-AOS1295.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated and edited by D.M. Dąbrowska and T.P. Speed from the Polish original (1990). *Statistical Science, 5*(4), 465–472.

Pawlak, Z. (1991). *Rough sets: Theoretical aspects of reasoning about data*. Kluwer Academic Publishers.

Pawlak, Z. (2004). Some issues on rough sets. In J. F. Peters, A. Skowron, J. W. Grzymała-Busse, B. Kostek, R. W. Świniarski, & M. S. Szczuka (Eds.), *Transactions on rough sets i* (pp. 1–58). Springer Berlin Heidelberg.

Pawlak, Z., & Skowron, A. (2007). Rudiments of rough sets. *Information Sciences, 177*(1), 3–27. doi:10.1016/j.ins.2006.06.003.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Pearl, J. (2010). The foundations of causal inference. *Sociological Methodology, 40*(1), 75–149.

Pearl, J. (2015). Causal thinking in the twilight zone. *Observational Studies, 1*, 200–204.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM, 62*, 54–60.

Pedrycz, W. (2013). *Granular computing : Analysis and design of intelligent systems*. CRC Press.

Peters, J. F., & Skowron, A. (2006). Some contributions by Zdzisław Pawlak. In G.-Y. Wang, J. F. Peters, A. Skowron, & Y. Yao (Eds.), *Rough sets and knowledge technology* (pp. 1–11). Springer Berlin Heidelberg.

Qian, J., Miao, D., Zhang, Z., & Yue, X. (2014). Parallel attribute reduction algorithms using mapreduce. *Information Sciences, 279*, 671–690. doi:10.1016/j.ins.2014.04.019.

Ramasso, E., & Denoeux, T. (2014). Making use of partial knowledge about hidden states in hmms: An approach based on belief functions. *IEEE Transactions on Fuzzy Systems, 22*(2), 395–405. doi:10.1109/TFUZZ.2013.2259496.

Raza, M. S., & Qamar, U. (2018). A heuristic based dependency calculation technique for rough set theory. *Pattern Recognition, 81*, 309–325. doi:10.1016/j.patcog.2018.04.009.

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

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. Springer New York.

Tian, J., & Pearl, J. (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence, 28*(1–4), 287–313.

Tran, A. D., Arch-Int, S., & Arch-Int, N. (2018). A rough set approach for approximating differential dependencies. *Expert Systems with Applications, 114*, 488–502. doi:10.1016/j.eswa.2018.06.025.

Wang, J., Zhao, M., Zhao, K., & Han, S. (2003). Multilevel data summarization from information systems: A "rule + exception" approach. *AI Communications, 16*(1), 17–39.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20*, 557–585.

Wu, W. Z., Qian, Y., Li, T. J., & Gu, S. M. (2017). On rule acquisition in incomplete multi-scale decision tables. *Information Sciences, 378*, 282–302. doi:10.1016/j.ins.2016.03.041.

Yamaguchi, D. (2009). Attribute dependency functions considering data efficiency. *International Journal of Approximate Reasoning, 51*(1), 89–98. doi:10.1016/j.ijar.2009.08.002.

Yao, J., & Azam, N. (2015). Web-based medical decision support systems for three-way medical decision making with game-theoretic rough sets. *IEEE Transactions on Fuzzy Systems, 23*, 3–15. doi:10.1109/TFUZZ.2014.2360548.

Yao, N., Miao, D., Zhang, Z., & Lang, G. (2016). Probabilistic estimation for generalized rough modus ponens and rough modus tollens. In V. Flores, F. Gomide, A. Janusz, C. Meneses, D. Miao, G. Peters, … Y. Yao (Eds.), *Rough sets* (pp. 166–176). Springer International Publishing.

Yao, Y. (2011). The superiority of three-way decisions in probabilistic rough set models. *Information Sciences, 181*(6), 1080–1096. doi:10.1016/j.ins.2010.11.019.

Zhang, X., & Miao, D. (2016). Quantitative/qualitative region-change uncertainty/certainty in attribute reduction: Comparative region-change analyses based on granular computing. *Information Sciences, 334–335*, 174–204. doi:10.1016/j.ins.2015.11.037.

Zhou, J., & Miao, D. (2011). β-Interval attribute reduction in variable precision rough set model. *Soft Computing, 15*(8), 1643–1656. doi:10.1007/s00500-011-0693-4.

Ziarko, W. (1993). Variable precision rough set model. *Journal of Computer and System Science, 46*(1), 39–59. doi:10.1016/0022-0000(93)90048-2.

Ziarko, W. (2007). Dependencies in structures of decision tables. In M. Kryszkiewicz, J. F. Peters, H. Rybinski, & A. Skowron (Eds.), *Rough sets and intelligent systems paradigms* (pp. 113–121). Springer Berlin Heidelberg.

Ziarko, W. (2008). Probabilistic approach to rough sets. *International Journal of Approximate Reasoning, 49*(2), 272–284. doi:10.1016/j.ijar.2007.06.014.