A Knowledge-Intensive Approach for Semi-Automatic Causal Subgroup Discovery

Martin Atzmueller and Frank Puppe

Abstract This paper presents a methodological view on knowledge-intensive causal subgroup discovery implemented in a semi-automatic approach. We show how to identify causal relations between subgroups by generating an extended causal subgroup network utilizing background knowledge. Using the links within the network we can identify causal relations, but also relations that are potentially confounded and/or effect-modified by external (confounding) factors. In a semi-automatic approach, the network and the discovered relations are presented to the user as an intuitive visualization. The applicability and benefit of the presented technique is illustrated by examples from a case-study in the medical domain.

1 Introduction

Subgroup discovery (e.g., [17, 9, 10, 2]) is a powerful approach for exploratory and descriptive data mining to obtain an overview of the interesting dependencies between a specific target (dependent) variable and usually many explaining (independent) variables; for example, the risk of coronary heart disease (target variable) is significantly higher in the subgroup of smokers with a positive family history than in the general population.

When interpreting and applying the discovered relations, it is often necessary to consider the patterns in a causal context. However, considering an association with a causal interpretation can often lead to incorrect results, due to the basic tenet of statistical analysis that association does not imply causation (cf., [5]): A subgroup

Martin Atzmueller

University of Würzburg, Department of Computer Science VI, Am Hubland, 97074 Würzburg, Germany, e-mail: atzmueller@informatik.uni-wuerzburg.de

Frank Puppe

University of Würzburg, Department of Computer Science VI, Am Hubland, 97074 Würzburg, Germany, e-mail: puppe@informatik.uni-wuerzburg.de

may not be causal for the target group, and thus can be suppressed by other causal groups and thus become irrelevant. Then, the causal subgroups are better suited for characterizing the target concept, and already explain the relations captured by the suppressed subgroup. Furthermore, the estimated *effect*, i.e., the quality of the subgroup may be due to associations with other *confounding* factors that were not considered in the quality computation. For instance, a relatively high/low quality of a subgroup may only be due to other variables that are associated with the independent variables and are a direct cause of the (dependent) target variable. Then, it is necessary to identify potential confounders, and to measure or to control their influence concerning the subgroup and the target concept. Let us assume, for example, that ice cream consumption and murder rates are highly correlated. However, this does not necessarily mean that ice cream incites murder or that murder increases the demand for ice cream. Instead, both ice cream and murder rates might be joint effects of a common cause, namely, hot weather.

In this paper, we present an approach for the semi-automatic detection of (true) causal subgroups and potentially confounded and/or effect-modified relations. We apply known subgroup patterns in a knowledge-intensive process as background knowledge that can be incrementally refined: The applied patterns represent subgroups that are *acausal*, i.e., have no causes, and subgroup patterns that are known to be directly causally related to other (target) subgroups. Additionally, both types of patterns can be combined, for example, in the medical domain certain variables such as *Sex* have no causes, and it is known that they are causal risk factors for certain diseases.

Using the patterns contained in the background knowledge, and a set of subgroups for analysis, we can construct a causal net containing relations between the subgroups. In a semi-automatic process, this network can be interactively inspected and analyzed by the user: It directly provides a visualization of the causal relations between the subgroups, and also provides for a possible explanation of these. By traversing the relations in the network, we can then identify causal relations, potential confounding and/or effect-modification. The approach has been implemented as a plugin for the VIKAMINE¹ system.

The rest of the paper is organized as follows: First, we discuss the background of subgroup discovery, the concept of confounding, and basic constraint-based causal analysis methods in Section 2. After that we present the knowledge-intensive causal discovery approach for detecting causal and confounding/effect-modified relations in Section 3. Exemplary results of the application of the presented approach are given in Section 4 using data from a fielded system in the medical domain. Section 5 discusses the features, limitations, and experiences with the application of the presented approach. Finally, Section 6 concludes the paper with a summary of the presented work and points out interesting directions for future work.

¹ <http://vikamine.sourceforge.net>

2 Background

In this section, we first introduce the necessary notions concerning the used knowledge representation, before we define the setting for subgroup discovery. After that, we introduce the concept of confounding, criteria for its identification, and describe basic constraint-based techniques for causal subgroup analysis.

2.1 Basic Definitions

Let Ω_A be the set of all attributes. For each attribute $a \in \Omega_A$ a range $dom(a)$ of values is defined. The set $\mathcal{V}_A$ is assumed to be the (universal) set of attribute values of the form $(a = v)$, where $a \in \Omega_A$ is an attribute and $v \in dom(a)$ is an assignable value. We consider nominal attributes only so that numeric attributes need to be discretized accordingly.

Let CB be the case base (data set) containing all available cases (instances): A case $c \in CB$ is given by the n-tuple

$$c = ((a_1 = v_1), (a_2 = v_2), \dots, (a_n = v_n))$$

of $n = |\Omega_A|$ attribute values, $v_i \in dom(a_i)$ for each a_i .

2.2 Subgroup Discovery

The main application areas of subgroup discovery (e.g., [17, 9, 10, 2]) are exploration and descriptive induction, to obtain an overview of the relations between a (dependent) target variable and a set of explaining (independent) variables. As in the *MIDOS* approach [17], we consider subgroups that have the most unusual (distributional) characteristics with respect to the concept of interest given by a binary target variable. Therefore, not necessarily complete relations but also partial relations, i.e., (small) subgroups with "interesting" characteristics can be sufficient.

Subgroup discovery mainly relies on the subgroup description language, the quality function, and the search strategy. Often, heuristic methods (e.g., [10]) but also efficient exhaustive algorithms (e.g., the SD-Map algorithm [2]) are applied. The description language specifies the individuals belonging to the subgroup. For a common single-relational propositional language a subgroup description can be defined as follows:

Definition 1 (Subgroup Description). A subgroup description $sd = \{e_1, e_2, \dots, e_n\}$ is defined by the conjunction of a set of selectors $e_i = (a_i, V_i)$: Each of these are selections on domains of attributes, $a_i \in \Omega_A, V_i \subseteq dom(a_i)$. We define Ω_E as the set of all selectors and Ω_{sd} as the set of all possible subgroup descriptions.

A quality function measures the interestingness of the subgroup and is used to rank these. Typical quality criteria include the difference in the distribution of the target variable concerning the subgroup and the general population, and the subgroup size.

Definition 2 (Quality Function). Given a particular target variable $t \in \Omega_E$, a quality function $q : \Omega_{sd} \times \Omega_E \rightarrow R$ is used in order to evaluate a subgroup description $sd \in \Omega_{sd}$, and to rank the discovered subgroups during search.

Several quality functions were proposed (cf., [17, 9, 10, 2]), e.g., the functions q_{BT} and q_{RG} :

$$q_{BT} = \frac{(p - p_0) \cdot \sqrt{n}}{\sqrt{p_0 \cdot (1 - p_0)}} \cdot \sqrt{\frac{N}{N - n}}, \quad q_{RG} = \frac{p - p_0}{p_0 \cdot (1 - p_0)}, n \geq \mathcal{T}_{Supp},$$

where p is the relative frequency of the target variable in the subgroup, p_0 is the relative frequency of the target variable in the total population, $N = |CB|$ is the size of the total population, and n denotes the size of the subgroup.

In contrast to the quality function q_{BT} (the classic binomial test), the quality function q_{RG} only compares the target shares of the subgroup and the total population measuring the *relative gain*. Therefore, a support threshold $\mathcal{T}_{Supp}$ is necessary to discover significant subgroups.

The result of subgroup discovery is a set of subgroups. Since subgroup discovery methods are not necessarily covering algorithms, the discovered subgroups can overlap significantly, and their estimated quality (effect) might be confounded by external variables. In order to reduce the redundancy of the subgroups and to identify potential confounding factors, methods for causal analysis can then be applied.

2.3 The Concept of Confounding

Confounding can be described as a bias in the estimation of the effect of the subgroup on the target concept due to attributes affecting the target concept that are not contained in the subgroup description [11, 12]. Thus, confounding is caused by a lack of comparability between the subgroup and the complementary group due to a difference in the distribution of the target concept caused by other factors.

2.3.1 Simpson's Paradox.

An extreme case for confounding is presented by *Simpson's Paradox* [15]: The (positive) effect (association) between a given variable X and a variable T is countered by a negative association given a third factor F , i.e., X and T are negatively correlated in the subpopulations defined by the values of F . For binary variables X, T, F this can be formulated as

$$P(T|X) > P(T|\neg X), P(T|X, F) < P(T|\neg X, F), P(T|X, \neg F) < P(T|\neg X, \neg F),$$

i.e., the event X increases the probability of T in a given population while it decreases the probability of T in the subpopulations restricted by F and $\neg F$.

As an example, let us assume that there is a positive correlation between the event X that describes *people that do not consume soft drinks* and T specifying the diagnosis *diabetes* (see Figure 1).

Combined	T	$\neg T$	Σ	Rate (T)
X	25	25	50	50%
$\neg X$	20	30	50	40%
Σ	45	55	100	

Fig. 1 Example: Simpson’s Paradox – aggregate counts.

This association implies that people not consuming soft drinks are affected more often by diabetes (50% non-soft-drinkers vs. 40% soft-drinkers). However, this is due to age, if older people (given by F) consume soft drinks less often than younger people, and if diabetes occurs more often for older people, inverting the effect (see Figure 2).

Restricted on F	T	$\neg T$	Σ	Rate (T)
X	24	16	40	60%
$\neg X$	8	2	10	80%
Σ	32	18	50	

Restricted on $\neg F$	T	$\neg T$	Σ	Rate (T)
X	1	9	10	10%
$\neg X$	12	28	40	30%
Σ	13	37	50	

Fig. 2 Example: Simpson’s Paradox – grouped counts.

2.3.2 Criteria for Identifying Confounders.

There are three criteria that can be used to identify a **confounding factor** F [11], given the factors X contained in a subgroup description and a target concept T :

1. A confounding factor F must be a *cause* for the target concept T , e.g., an independent risk factor for a certain disease.
2. The factor F must be associated/correlated with the set of subgroup (factors) X , i.e., there needs to be a (statistically significant) association between F and X .
3. A confounding factor F must *not* be (causally) affected by the subgroup (factors) X , i.e., there must not be a causal dependency between X and F .

However, these criteria are only necessary but not sufficient to identify confounders. If purely automatic methods are applied for detecting confounding, then

such approaches may label some variables as confounders incorrectly, e.g., if the real confounders have not been measured, or if their contributions cancel out.

Thus, user interaction is rather important for validating confounded relations. Furthermore, the identification of confounding requires causal (background) knowledge since confounding is itself a causal concept [12]. Such background knowledge can be formalized in an interactive approach as discussed below.

2.3.3 Proxy Factors and Effect Modification.

There are two phenomena that are closely related to confounding. First, a factor may only be associated with the subgroup but may be the real cause for the target concept. Then, the subgroup is only a **proxy factor**. Another situation is given by **effect modification**: Then, a third factor F does not necessarily need to be associated with the subgroup described by the factors X ; F can be an additional factor that increases the effect of X in a certain subpopulation only, pointing to new subgroup descriptions that are interesting by themselves.

2.4 Constraint-Based Methods for Causal Subgroup Analysis

In general, the philosophical concept of causality refers to the set of all particular 'cause-and-effect' or 'causal' relations. A subgroup is causal for the target group, if in an ideal experiment [5] the probability of an object not belonging to the subgroup to be a member of the target group increases or decreases when the characteristics of the object are changed such that the object becomes a member of the subgroup.

For example, the probability that a patient survives (target group) increases if the patient received a special treatment (subgroup). Then, a redundant subgroup, that is, for example, conditionally independent from the target group given another subgroup, can be suppressed. Then the relation can already be completely explained by the conditioning subgroup and the independent subgroup does not provide any significant insight.

For causal analysis the subgroups are represented by binary variables that are true for an object (case) if it is contained in the subgroup, and false otherwise. For constructing a causal subgroup network, constraint-based methods are particularly suitable because of scalability reasons (cf., [5, 14]) since they only depend on simple statistical (in-)dependence tests, e.g., the χ^2 -test for independence.

However, constraint-based methods make several assumptions (cf., [5]) with respect to the data and the correctness of the statistical tests. The crucial condition is the *Markov condition* depending on the assumption that the data can be expressed by a Bayesian network: Let X be a node in a causal Bayesian network, and let Y be any node that is not a descendant of X in the causal network. Then, the Markov condition holds if X and Y are independent conditioned on the parents of X .

The CCC and CCU rules [14] described below constrain the possible causal models by applying simple statistical tests: For subgroups s_1, s_2, s_3 represented by binary variables the χ^2 -test for independence is utilized for testing their independence $ID(s_1, s_2)$, dependence $D(s_1, s_2)$ and conditional independence $CondID(s_1, s_2 | s_3)$, as shown below. For the tests user-selectable thresholds are applied, e.g., $\mathcal{T}_1 = 1$, $\mathcal{T}_2 = 3.84$, or higher, respectively:

$$\begin{aligned} ID(s_1, s_2) &\longleftrightarrow \chi^2(s_1, s_2) < \mathcal{T}_1, & D(s_1, s_2) &\longleftrightarrow \chi^2(s_1, s_2) > \mathcal{T}_2, \\ CondID(s_1, s_2 | s_3) &\longleftrightarrow \chi^2(s_1, s_2 | s_3 = 0) + \chi^2(s_1, s_2 | s_3 = 1) < 2 \cdot \mathcal{T}_1 \end{aligned}$$

Thus, the decision of (conditional) (in-)dependence is threshold-based, which is a problem causing potential errors if very many tests are performed (e.g., [16]). Therefore, we propose a semi-automatic approach featuring interactive analysis of the inferred relations. The CCC rule requires three pairwise dependent (*correlated*) variables while the CCU rule requires two dependence relations and one independence relation.

Definition 3 (CCC Rule). Let X, Y, Z denote three variables that are pairwise dependent, i.e., $D(X, Y), D(X, Z), D(Y, Z)$; let X and Z become independent when conditioned on Y . In the absence of hidden and confounding variables we may infer that one of the following causal relations exists between X, Y and Z : $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \rightarrow Z$, $X \leftarrow Y \leftarrow Z$. However, if X has no causes, then the first relation is the only one possible, even in the presence of hidden and confounding variables.

Definition 4 (CCU Rule). Let X, Y, Z denote three variables: X and Y are dependent ($D(X, Y)$), Y and Z are dependent ($D(Y, Z)$), X and Z are independent ($ID(X, Z)$), but X and Z become dependent when conditioned on Y ($CondD(X, Z | Y)$). In the absence of hidden and confounding variables, we may infer that X and Z cause Y .

3 Semi-Automatic Causal Subgroup Discovery and Analysis

The approach for knowledge-intensive causal subgroup discovery is embedded in an incremental semi-automatic process. In the following, we first introduce the process model for causal subgroup discovery and analysis. After that, we present the necessary background knowledge for effective causal analysis. Next, we describe a method for constructing an extended causal subgroup network, and discuss how to identify confounded and effect-modified relations. After that, we discuss visualizations for supporting the analysis.

3.1 Process Model for Causal Subgroup Discovery and Analysis

The process model includes a discovery, (causal) analysis, evaluation and validation, and finally a knowledge extraction and formalization step. The individual steps of the process for semi-automatic causal subgroup discovery are shown in Figure 3 and discussed below.

1. First, the user applies a standard *subgroup discovery* approach, e.g., [10, 2]. The result of this step is a set of the most interesting subgroups. Optionally, background knowledge contained in the knowledge base can also be applied during subgroup discovery (e.g., as discussed in [3]).
2. Next, *causal analysis* using appropriate background knowledge for a detailed analysis of the discovered associations is applied. Using constraint-based techniques for causal analysis, a (partial) causal subgroup network containing the discovered interesting subgroups is constructed.
3. In the *evaluation and validation* phase, the user assesses the (partial) causal network: Since the relations contained in the network can be wrong due to various statistical errors, inspection and validation of the causal relations is essential in order to obtain valid results. After inspection and validation the final results (as a selection of the discovered subgroup patterns) are obtained.
4. The user can extend and/or tune the applied background knowledge during the *knowledge extraction* step: Then, the knowledge base can be updated in incremental fashion by including further background knowledge, e.g., based upon the discovery results. The formalized background knowledge can then be incrementally (re-)applied and integrated into the process.

Fig. 3 Process Model for Knowledge-Intensive Causal Subgroup Analysis.

3.2 Extended Causal Subgroup Analysis

Detecting causal relations, i.e., (true) causal associations, confounding, and effect modification using causal subgroup analysis consists of the following two main steps that can be iteratively applied.

1. First, we generate a causal subgroup network considering a target group T , a user-selected set of subgroups U , a set of confirmed (causal) and potentially confounding factors C for any included group, a set of unconfirmed potentially confounding factors P given by subgroups significantly dependent with the target group, and additional background knowledge described below. In addition to causal links, the generated network also contains (undirected) associations between the variables.
2. In the next step, we traverse the network and mark the potential confounded and/or effect-modified relations. The causal network and the proposed relations are then presented to the user for subsequent interpretation and analysis. After the factors have been confirmed the background knowledge can be extended.

In the following sections we discuss these steps in detail. First, we describe the elements of the background knowledge that are utilized for causal analysis.

3.3 Background Knowledge for Causal Analysis

For the causal analysis step we first need to generate an extended causal network capturing the relations between the subgroups represented by binary variables.

In order to effectively generate the network, we need to include background knowledge provided by the user, for example, by a domain specialist. In the medical domain, for example, a lot of background knowledge is already available and can be directly integrated in the analysis phase. Examples include (causal) relations between diseases, and the relations between diseases and their respective findings. Additionally, it is often required to include the available background knowledge since otherwise only known relations are rediscovered (cf., [3]), valid results can even not be obtained, or too many results are proposed that need to be validated manually. The applicable background consists of two basic elements, i.e., acausal factors (variables) and causal relations between variables:

1. Acausal factors: These include factors represented by subgroups that have no causes; in the medical domain, e.g., the subgroup $Age \geq 70$ or the subgroup $Sex = male$ have no causes, while the subgroup $BMI = underweight$ has certain causes.
2. (Direct) causal relations: Such relations include subgroups that are not only dependent but (directly) causal for the target group/target variable and/or other subgroups. In the medical domain, for example, the subgroup $Body-Mass-Index (BMI)=overweight$ is directly causal for the subgroup $Gallstones=probable$.

Depending on the domain, for example, considering the medical domain, it is often relatively easy to provide acausal information. Direct and indirect causal relations are often also easy to acquire, and can be acquired 'on-the-fly' when applying the presented process. However, in some domains, e.g., in the medical domains, it is often difficult to provide non-ambiguous directed relationships between certain

variables: One disease can cause another disease and vice versa, under different circumstances. In such cases, the relations should be formalized with respect to both directions, and can still be exploited in the method discussed below. Then, the interpretation performed by the user is crucial in order to obtain valid and ultimately interesting results.

3.4 Constructing an Extended Causal Subgroup Network

Algorithm 1 summarizes how to construct an extended causal subgroup network, based on a technique for basic causal subgroup analysis described in [9, 8]. When applying the algorithm, the relations contained in the network can be wrong due to various statistical errors (cf., [5]), especially for the CCU rule (cf., [14]). Therefore, after applying the algorithm, the resulting causal net is presented to the user for interactive analysis.

The first step (lines 1-5) of the algorithm determines for each subgroup pair (including the target group) whether they are independent, based on the inductive principle that the dependence of subgroups is necessary for their causality.

In the next step (lines 6-10) we determine for any pair of subgroups whether the first subgroup s_1 is suppressed by a second subgroup s_2 , i.e., if s_1 is conditionally independent from the target group T given s_2 . The χ^2 -measure for the target group and s_1 is calculated both for the restriction on s_2 and its complementary subgroup. If the sum of the two test-values is below a threshold, then we can conclude that subgroup s_1 is conditionally independent from the target group. Conditional independence is a sufficient criterion, since the target distribution of s_1 can be explained by the target distribution in s_2 , i.e., by the intersection. Since similar subgroups could symmetrically suppress each other, the subgroups are ordered by quality; then subgroups with a nearly identical extension (and lower quality) can be eliminated.

The next two steps (lines 11-18) check conditional independence between each pair of subgroups given the target group or a third subgroup, respectively. For each pair of conditionally independent groups, the separating (conditioning) group is noted. Then, this separator information is exploited in the next steps, i.e., independencies or conditional independencies for pairs of groups derived in the first steps are used to exclude any causal links between the groups. The conditioning steps (lines 6-18) can optionally be iterated in order to condition on combinations of variables (pairs, triples). However, the decisions taken further (in the CCU and CCC rules) may become statistically weaker justified due to smaller counts in the considered contingency tables (e.g., [5, 8]).

Direct causal links (line 19) are added based on background knowledge, i.e., given subgroup patterns that are causal for specific subgroups. In the last step (lines 24-26) we also add conditional associations for dependent subgroups that are not conditionally independent and thus not suppressed by any other subgroups. Such links can later be useful in order to detect the (true) associations considering a confounding factor.

Algorithm 1 Constructing a causal subgroup net.

Require: Target group T , user-selected set of subgroups U , potentially confounding groups P , background knowledge B containing acausal subgroup information, and known subgroup patterns $C \subseteq B$ that are directly causal for other subgroups. Define $\mathcal{S} = U \cup P \cup C$

- 1: **for all** $s_i, s_j \in \mathcal{S} \cup T, s_i \neq s_j$ **do**
- 2: **if** $\text{approxEqual}(s_i, s_j)$ **then**
- 3: Exclude any causalities for the subgroup with smaller correlation to T
- 4: **if** $ID(s_i, s_j)$ **then**
- 5: Exclude causality: $ex(s_i, s_j) = \text{true}$
- 6: **for all** $s_i, s_j \in \mathcal{S}, s_i \neq s_j$ **do**
- 7: **if** $\neg ex(s_i, T), \neg ex(s_j, T), \text{ or } \neg ex(s_i, s_j)$ **then**
- 8: **if** $\text{CondID}(s_i, T|s_j)$ **then**
- 9: Exclude causality: $ex(s_i, T) = \text{true}$, and include s_j into $\text{separators}(s_i, T)$
- 10: If conditional independencies are symmetric, then select the strongest relation
- 11: **for all** $s_i, s_j \in \mathcal{S}, i < j$ **do**
- 12: **if** $\neg ex(s_i, T), \neg ex(s_j, T), \text{ or } \neg ex(s_i, s_j)$ **then**
- 13: **if** $\text{CondID}(s_i, s_j|T)$ **then**
- 14: Exclude causality: $ex(s_i, s_j) = \text{true}$, and include T into $\text{separators}(s_i, s_j)$
- 15: **for all** $s_i, s_j, s_k \in \mathcal{S}, i < j, i \neq k, j \neq k$ **do**
- 16: **if** $\neg ex(s_i, s_j), \neg ex(s_j, s_k), \text{ or } \neg ex(s_i, s_k)$ **then**
- 17: **if** $\text{CondID}(s_i, s_j|s_k)$ **then**
- 18: Exclude causality: $ex(s_i, s_j) = \text{true}$, and include s_k into $\text{separators}(s_i, s_j)$
- 19: Integrate direct causal links that are not conditionally excluded considering the sets C and B
- 20: **for all** $s_i, s_j, s_k \in \mathcal{S}$ **do**
- 21: Apply the extended CCU rule, using background knowledge
- 22: **for all** $s_i, s_j, s_k \in \mathcal{S}$ **do**
- 23: Apply the extended CCC rule, using background knowledge
- 24: **for all** $s_i, s_j, s_k \in \mathcal{S} \cup \{T\}, i < j, i \neq k, j \neq k$ **do**
- 25: **if** $\neg \text{CondID}(s_i, s_j|s_k)$ **then**
- 26: Integrate association between dependent s_i and s_j that are not conditionally excluded

Extending CCC and CCU using Background Knowledge The CCU and CCC steps (lines 20-23) derive the directions of the causal links between subgroups, based on information derived in the previous steps. In the context of the presented knowledge-intensive approach, we extend the basic CCC and CCU rules including background knowledge both for the derivation of additional links, and for inhibiting links that contradict the background knowledge. We introduce associations instead of causal directions if these are wrong, or if not enough information is available in order to derive the causal directions. The rationale behind this principle is given by the intuition that we want to exploit and provide as much information as possible considering the generated causal net. When identifying potentially confounded relations, we can also often utilize weaker associative information.

For the **extended CCU rule** we use background knowledge for inhibiting acausal directions, since the CCU rule can be disturbed by confounding and hidden variables. The causal or associative links do not necessarily indicate direct associations/causal links but can also point to relations enabled by hidden or confounding variables [14].

For the **extended CCC rule**, we can use the relations inferred by the extended CCU rule for disambiguating between the causal relations, if the *CCU* rule is applied in all possible ways: The non-separating condition (conditional dependency) of the relation identified by the *CCU* rule is not only a sufficient but a necessary condition [14], i.e., for $X \rightarrow Y \leftarrow Z$, with $CondID(X, Z|Y)$. Additionally, we can utilize background knowledge for distinguishing between the causal relations. So, for three variables X, Y, Z with $D(X, Y), D(X, Z), D(Y, Z)$, and $CondID(X, Z|Y)$, if there exists an (inferred) causal link $X \rightarrow Y$ between X and Y , we may identify the relation $X \rightarrow Y \rightarrow Z$ as the true relation. Otherwise, if Y or Z have no causes, then we select the respective relation, e.g., $X \leftarrow Y \rightarrow Z$ for an acausal variable Y .

3.5 Identifying Confounded Relations

A popular method for controlling confounding factors is given by *stratification* [11]: For example, in the medical domain a typical confounding factor is the attribute *age*: We can stratify on age groups such as $age < 30$, $age 30 - 69$, and $age \geq 70$. Then, the subgroup – target relations are measured within the different strata, and compared to the (crude) unstratified measure.

It is easy to see, that in the context of the presented approach stratification for a binary variable is equivalent to conditioning on them: If we assess a conditional subgroup – target relation and the subgroup factors become independent (or dependent), then this indicates potential confounding. After constructing a causal net, we can easily identify such relations. Since the causal directions derived by the extended CCC and CCU rules may be ambiguous, user interaction is crucial: In some domains, e.g., in the medical domains, it is often difficult to provide non-ambiguous directed relationships between certain variables: One disease can cause another disease and vice versa, under different circumstances. The network then also provides an intuitive visualization for the analysis.

In order to identify potentially confounded relations and the corresponding variables, as shown in Figure 4, and described below, we just need to traverse the network. In this way, we collect the interesting and relevant relations.

- **Potential Confounding:** If there is an association between two variables X and Y , and the network contains the relations $C \rightarrow X, C \rightarrow Y$, and there is no causal link between X and Y , i.e., they are conditionally independent given C , then C is a confounder that inhibits the relation between X and Y . This is also true if there is no causal link between C and X but instead an association.
- **Potential Confounding/Effect Modification:** If the network contains the relations $C \rightarrow X, C \rightarrow Y$, and there is also either an association or a causal link between X and Y , then this points to confounding and possible effect modification of the relation between the variables X and Y .
- **Potential Collider (or Confounder):** If there is no (unconditioned) association between two variables X and Y and the network contains the relations $X \rightarrow C$

Fig. 4 Examples of Confounded Relations.

and $Y \rightarrow C$, then C is a potential collider: X and Y become dependent by conditioning on C . The variable C is then no confounder in the classical sense, if the (inferred) causal relations are indeed true. However, such a relation as inferred by the CCU rule can itself be distorted by confounded and hidden variables. The causal directions could also be inverted, if the association between X and Y is just not strong enough as estimated by the statistical tests. In this case, C is a potential confounder. Therefore, manual inspection is crucial in order to detect the true causal relation.

3.6 Supporting Visual Techniques

In order to support the user when analyzing and validating the causal network and the contained relations, we propose several techniques that are integrated into the main visualization for displaying the network itself. The proposed visualization is interactive such that the user can apply changes on the fly and adapt the parameters as needed. Feedback is then implemented instantly. The visualizations are completely integrated in the VIKAMINE (Visual, Interactive, and Knowledge-intensive Analysis and MINing Environment) system (vikamine.sourceforge.net)

and can also be combined with the other visualization capabilities provided by the system. In this way, the causal analysis step is completely embedded within the proposed knowledge-intensive subgroup mining process. The approach for causal analysis is provided by a special plugin for the VIKAMINE system: Starting with the common subgroup discovery setting using the core functionality of VIKAMINE, the results can then be analyzed using the *causal analysis plugin* and can be visualized using the visualization techniques presented below.

In general, the *Visual Information Seeking Mantra* by [13], ‘Overview first, zoom and filter, then details-on-demand’ is an important guideline for visualization methods. In an iterative process, the user is able to subsequently concentrate on the interesting data by filtering irrelevant and redundant data, and focusing (zooming in) on the interesting elements, until finally details are available for an interesting subset of the analyzed objects. Therefore, we implemented the visualizations according to these guidelines. Therefore, we therefore propose the following techniques for the visualization of the causal subgroup network:

- Association filters: Selected associations can be explicitly excluded or included, e.g., with respect to non-causal associations. Then, these associations are marked as undirected lines otherwise. By default (non-causal) associations are excluded from view.
- Color coding of nodes: Acausal nodes, conflicting nodes, conflicting (bidirectional edges), are marked with different (customizable) colors for a better overview.
- Node filters (zooming in on the network): We provide filters for selecting a subset of the nodes of the network. The user can then select a set of specific nodes of interest in order to focus the analysis.

Essentially, the *assessment* of the patterns, i.e., the evaluation and validation of the relations in order to determine their final interestingness need to be facilitated by specialized techniques. Especially the zooming and filtering techniques allow an easier assessment of the relations in the causal network as discussed below.

Figure 5 shows an example of a zoomed network for which several filters have been applied. Compared to Figure 7 it can be easily seen that associations have been filtered. The target node is shown in green color, while two obviously acausal edges are colored in red, since incorrect causal directions have been derived for these (they are only associated, as discussed above). Causal edges are shown in green, while the conflicting bidirectional edges are shown in red. The strength of the association is shown by the width of the edges. Also, in comparison to Figure 7 the subgroups corresponding to the causal nodes are shown with their respective *true positive* and *false positive* counts – as sub-bars within the nodes. In contrast, Figure 7 focuses on the (causal) relations between the nodes.

Fig. 5 Example for a causal graph with filtered and color-coded elements.

4 Examples

For the experiments we utilized a case base containing about 8600 cases taken from the SONOCONSULT system [7] – a medical documentation and consultation system for sonography. The system is in routine use in the DRK-hospital in Berlin/Köpenick, and the collected cases contain detailed descriptions of findings of the examination(s), together with the inferred diagnoses (binary attributes).

The experiments were performed using the VIKAMINE system [1] for semi-automatic knowledge-intensive subgroup discovery, that was extended with a component implementing the presented approach.

Fig. 6 Confounded Relation: Gallstones and Liver cirrhosis.

In the following, we provide some (simplified) examples considering the diagnosis *Gallstones=established* as the target variable. After applying a subgroup discovery method, several subgroups were selected by the user in order to derive a causal subgroup network, and to check the relations with respect to possible confounders. These selected subgroups included, for example, the subgroup *Fatty liver=probable or possible* and the subgroup *Liver cirrhosis=probable*.

A first result is shown in Figure 6: In this network, the subgroup *Liver cirrhosis=probable* is confounded by the variable *Age >= 70*. However, there is still an influence on the target variable considering the subgroup *Liver cirrhosis=probable* shown by the association between the subgroups. This first result indicates confounding and effect modification (the strengths of the association between the nodes is also visualized by the widths of the links). A more detailed result is shown in Figure 7: In this network another potential confounder, i.e., *Sex=female* is included. Then, it becomes clear, that both the subgroup *Fatty liver=probable or possible* and the subgroup *Liver cirrhosis=probable* are confounded by the variables *Sex* and

Fig. 7 Confounded Relations: Gallstones, Liver cirrhosis and Fatty Liver.

Age, and the association (shown in Figure 6) between the subgroup *Liver cirrhosis=probable* and the target group is then no longer present (This example does not consider the removal of the gall-bladder which might have an additional effect concerning a medical interpretation).

It is easy to see that the generated causal subgroup network becomes harder to interpret, if many variables are included, and if the number of connections between the nodes increases. Therefore, the presented filtering and zooming techniques, e.g., in order to exclude (non-causal) associations, and for coloring the nodes and the edges of the network in order to increase their interpretability, are crucial for an effective implementation: Based on the available background knowledge, causal subgroup nodes, and (confirmed) causal directions can be marked. Since the network is first traversed and the potentially confounded relations are reported to the user, the analysis can also be focused towards the respective variables, as a further filtering condition. The user can then analyze selected parts of the network in more detail.

5 Discussion

The presented approach has a rather broad application area in various domains. It provides the means for identifying confounded and/or effect-modified subgroups with respect to a certain target variable. Starting with the results (i.e., the output) of a common subgroup discovery algorithm, a causal network for a selected set of (interesting) subgroups is generated, that can then be traversed. Next, the interesting findings are highlighted for visual inspection by the user. In contrast to approaches that try to automatically detect confounding or try to locate instances of Simpson's Paradox, this is a quite important distinction: The presented method is not able to automatically detect confounded relations with guarantees for the correctness of these findings. Due to statistical errors of the applied tests, various assumptions for these, and the necessary causality conditions this is also a rather hard problem. However, the presented method allows for a semi-automatic approach for detection of confounding and/or effect-modification. This semi-automatic approach provides more reliable results, since the 'human in the loop' still has the option for the necessary

validation of the identified relations, because the causal inferences made by the algorithm may be wrong due to statistical errors. This is especially relevant in the medical domain in which the findings of such an algorithm always need to be validated before being proclaimed as authoritative. In our applications this proved to be crucial for the acceptance of the presented methods.

Since the proposed approach also includes the feature of including background knowledge into the process, knowledge-rich domains, e.g., the medical domain, directly benefit from this powerful option. For such domains, applying background knowledge is not only sufficient, it is often necessary in order to gain the acceptance of the users: Medical doctors, for example, often expect high-quality results that can only be obtained when enough background knowledge has been provided to the (semi-)automatic discovery and analysis methods.

Furthermore, the power of the method is significantly enhanced when applying background knowledge, since many more causal relations can potentially be inferred by the analysis algorithm. The knowledge can then also be incrementally formalized and captured in the knowledge base for future reference. This feature is especially suitable for less knowledge-rich domains, e.g., specialized technical domains, for which there is only little initial background knowledge available.

A special advantage of the presented method is given by the fact, that the method does not work as a black box: Instead the results and findings of the algorithm, i.e., the identified relations and inferred causal links can always be *explained* by the application of the rules (CCC and CCU) and by the given background knowledge. In this way, a trace of the rule applications can provide significant insight in the case of possible contradictions and erroneous decisions of the algorithm. Then, the user can always refer to these rule applications for locating and fixing potential problems, and also for supporting findings of the algorithm.

6 Conclusion

In this paper, we have presented an approach for knowledge-intensive causal subgroup discovery: We utilize background knowledge for establishing an extended causal model of the domain. The constructed network can then be used to identify potential causal relations, and confounded/effect-modified associations. In a semi-automatic approach, these can then be evaluated and validated by the user. During this step the user is supported by appropriate visualization techniques that are key features for an effective approach. Furthermore, the available background knowledge can be incrementally refined and extended.

In the future, we are planning to integrate an efficient approach for detecting confounding that is directly embedded in the subgroup discovery method. Related work in that direction was described, for example, in [6, 4]. Another interesting direction for future work is given by considering further background knowledge for causal analysis and by integrating existing ontologies into the knowledge base. Additionally, the refinement of the contained relations is an interesting issue to consider.

Acknowledgements

This work has been partially supported by the German Research Council (DFG) under grant Pu 129/8-2.

References

1. Atzmueller, M., Puppe, F.: Semi-Automatic Visual Subgroup Mining using VIKAMINE. *Journal of Universal Computer Science* **11**(11), 1752–1765 (2005)
2. Atzmueller, M., Puppe, F.: SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery. In: *Proc. 10th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006)*, pp. 6–17. Springer, Berlin (2006)
3. Atzmueller, M., Puppe, F., Buscher, H.P.: Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery. In: *Proc. 19th Intl. Joint Conference on Artificial Intelligence (IJCAI-05)*, pp. 647–652. Edinburgh, Scotland (2005)
4. Atzmueller, M., Puppe, F., Buscher, H.P.: A semi-automatic approach for confounding-aware subgroup discovery. *International Journal on AI Tools (IJAIT)* **18**(1), 1 – 18 (2009)
5. Cooper, G.F.: A Simple Constraint-Based Algorithm for Efficiently Mining Observational Databases for Causal Relationships. *Data Min. Knowl. Discov.* **1**(2), 203–224 (1997)
6. Fabris, C.C., Freitas, A.A.: Discovering Surprising Patterns by Detecting Occurrences of Simpson’s Paradox. In: *Research and Development in Intelligent Systems XVI*, pp. 148–160. Springer, Berlin (1999)
7. Huettig, M., Buscher, G., Menzel, T., Scheppach, W., Puppe, F., Buscher, H.P.: A Diagnostic Expert System for Structured Reports, Quality Assessment, and Training of Residents in Sonography. *Medizinische Klinik* **99**(3), 117–122 (2004)
8. Kloesgen, W., May, M.: Database Integration of Multirelational Causal Subgroup Mining. Tech. rep., Fraunhofer Institute AIS, Sankt Augustin, Germany (2002)
9. Klösgen, W.: *Handbook of Data Mining and Knowledge Discovery*, chap. 16.3: Subgroup Discovery. Oxford University Press, New York (2002)
10. Lavrac, N., Kavsek, B., Flach, P., Todorovski, L.: Subgroup Discovery with CN2-SD. *Journal of Machine Learning Research* **5**, 153–188 (2004)
11. McNamee, R.: Confounding and Confounders. *Occup. Environ. Med.* **60**, 227–234 (2003)
12. Pearl, J.: *Causality: Models, Reasoning and Inference*, chap. 6.2: Why There is No Statistical Test For Confounding, Why Many Think There Is, and Why They Are Almost Right. Cambridge University Press (2000)
13. Shneiderman, B.: The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: *Proc. IEEE Symp. Visual Languages*, pp. 336–343. Boulder, Colorado (1996)
14. Silverstein, C., Brin, S., Motwani, R., Ullman, J.D.: Scalable Techniques for Mining Causal Structures. *Data Mining and Knowledge Discovery* **4**(2/3), 163–192 (2000)
15. Simpson, E.H.: The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society* **18**, 238–241 (1951)
16. Webb, G.I.: Discovering Significant Patterns. *Machine Learning* **68**(1), 1–33 (2007)
17. Wrobel, S.: An Algorithm for Multi-Relational Discovery of Subgroups. In: *Proc. 1st Europ. Symposium on Principles of Data Mining and Knowledge Discovery*, pp. 78–87. Springer, Berlin (1997)