Discovering most informative rules from data

Gabriele Kern-Isberner
FernUniversität Hagen, Dept. of Computer Science,
P.O. Box 940, D-58084 Hagen, Germany,
gabriele.kern-isberner@fernuni-hagen.de

Abstract

The search for most relevant relationships in data is a central concern of knowledge discovery. In this paper, relationships are represented by rules (or, more formally, by conditionals), and relevance is defined with respect to an information-theoretical principle. We show how to extract structural conditional information from numerical relationships in a non-heuristic, mathematically founded way. The basic idea is to consider knowledge discovery as a process inverse to inductive knowledge representation, and to use a group-theoretical machinery for disentangling highly complex interactions of rules.

1 Introduction

Commonsense and expert knowledge is most generally expressed by rules, connecting a precondition and a conclusion by an *if-then*-construction. If-then-rules are more formally denoted as *conditionals*. The crucial point with conditionals is that they carry generic knowledge which can be applied to different situations. This makes them most interesting objects in Artificial Intelligence, in theoretical as well as in practical respect. For instance, a salesman who has general knowledge about the preferences of his customers can use this knowledge when consulting any new customer. So in e-commerce, rule bases describing preferences of customers are most important for modelling intelligent sales assistants.

Typically, two central problems have to be solved in practical applications:

- Where do the rules come from? How to extract rules from statistical data?
- How should rules be represented? How to propagate and combine conditional knowledge?

Both of these problems can be dealt with separately, but it is most rewarding to combine them, that is,

- to discover rules which are most relevant *with respect to some inductive representation formalism*, and
- to build up a "best" model from the discovered rules that can be used for queries.

This paper will present an approach to discover rules which are most relevant with respect to the *maximum entropy-methods*. Since entropy measures the amount of information, this approach can be considered as aiming to find the *most informative rules* in data. The basic idea is to exploit numerical relationships as manifestations of interactions of underlying conditional knowledge. Our approach differs from usual knowledge discovery and data mining methods in that it takes explicitly inductive representation, or inference, respectively, into consideration. It is not based on observing conditional independencies, but aims at learning relevant conditional dependencies in a non-heuristic way. As a further novelty, our method does not compute single, isolated rules, but yields as a result a set of rules in taking into account highly complex interactions of rules.

Notations. Let $\mathcal{L}$ be a propositional language over a finite alphabet $a, b, c \ldots$. Let Ω denote the set of possible worlds for $\mathcal{L}$, i.e. Ω is the complete set of interpretations of $\mathcal{L}$. Throughout this paper, we will write $\overline{A}$ instead of $\neg A$, and AB instead of $A \wedge B$, for formulas $A, B \in \mathcal{L}$. Conditionals $(B|A)$ represent statements of the form "*If A then B*", expressing a relationship between the *antecedent* $A \in \mathcal{L}$ and the *consequent* $B \in \mathcal{L}$. $(\mathcal{L} \mid \mathcal{L})$ denotes the set of all conditionals $(B|A)$ with $A, B \in \mathcal{L}$. *Single-elementary conditionals* are conditionals whose antecedents are conjunctions of literals, and whose consequents consist of one single literal. Let P be a probability distribution over $\mathcal{L}$, and let $(B|A) \in (\mathcal{L} \mid \mathcal{L})$ and $x \in [0,1]$. Then P *represents* $(B|A)\,[x]$, $P \models (B|A)\,[x]$, iff $P(A) > 0$ and $P(AB) = xP(A)$. P represents a set $\mathcal{R}^* = \{(B_1|A_1)\,[x_1], \ldots, (B_n|A_n)\,[x_n]\}$ of probabilistic conditionals iff $P \models (B_i|A_i)\,[x_i]$ for all $1 \leqslant i \leqslant n$.

1.1 Modelling based on the principle of maximum entropy

Suppose a set $\mathcal{R}^* = \{(B_1|A_1)\,[x_1], \ldots, (B_n|A_n)\,[x_n]\}$ of probabilistic conditionals is given. For instance, $\mathcal{R}^*$ may describe the knowledge available to a physician when he has to make a diagnosis. Or, $\mathcal{R}^*$ may express commonsense knowledge like "*Students are young with a probability of (about) 80 %*" and "*Singles (i.e. unmarried people) are young with a probability of (about) 70 %*", the latter knowledge being formally expressed by $\mathcal{R}^* = \{(young|student)[0.8], (young|single)[0.7]\}$.

Usually, these rule bases represent incomplete knowledge, in that there are a lot of probability distributions apt to represent them. So learning, or inductively representing, respectively, the rules means to take them as a set of conditional constraints and to select a unique probability distribution as a "best" model which can be used for queries and further inferences. Paris [15] investigates several inductive representation techniques and proves that the *principle of maximum entropy, (ME-principle)* yields the only method to represent incomplete knowledge in an unbiased way, satisfying a set of postulates describing sound commonsense reasoning. The entropy $H(P)$ of a probability distribution P is defined as $H(P) = -\sum_\omega P(\omega) \log P(\omega)$ and measures the amount of indeterminateness inherent in P. Applying the principle of maximum entropy then means to select the unique distribution $P^* = ME(\mathcal{R}^*)$ that maximizes $H(P)$ subject to $P \models \mathcal{R}^*$. In this way, the ME-method ensures that no further information is added, so that the knowledge $\mathcal{R}^*$ is represented most faithfully. $ME(\mathcal{R}^*)$ can be written in the form

$$ME(\mathcal{R}^*)(\omega) = \alpha_0 \prod_{\substack{1 \leqslant i \leqslant n \\ \omega \models A_i B_i}} \alpha_i^{1-x_i} \prod_{\substack{1 \leqslant i \leqslant n \\ \omega \models A_i \overline{B_i}}} \alpha_i^{-x_i} \tag{1}$$

with the α_i's being chosen appropriately so as to satisfy all of the conditional constraints in $\mathcal{R}^*$ (cf. [9]).

The ME-principle provides a most convenient and founded method to represent incomplete probabilistic knowledge (efficient implementations of ME-systems can be found via `www.fernuni-hagen.de/BWLOR/forsch.htm` and `www.pit-systems.de`). Unlike Bayesian networks, no external (and often unjustified) independence assumptions have to be made, and only relevant conditional dependencies are part of the knowledge base. Bayesian networks need a lot of probabilities being specified. If one has to model the dependencies, for instance, between two diseases, D_1, D_2, and two symptoms, S_1, S_2, one has to quantify all probabilities $P(s_j|d_i)$, where s_j and d_i, respectively, is any one of $S_j, \neg S_j$ and $D_i, \neg D_i$, for $i, j = 1, 2$. But not only the large amount of probabilities necessary to build up Bayesian networks are a problem. Although a physician will usually be capable to quantify $P(S_j|D_i)$ from his expert knowledge, he will hardly be able to say something informed about $P(S_j|\neg D_i)$ – what is the probability of a symptom given that the disease is *not* present? In an ME-environment, the expert has only to list whatever relevant conditional probabilities he is aware of. Moreover, the two basic ingredients for Bayesian networks, namely the set of conditional probabilities and the independence assumptions, specify *complete probabilistic knowledge*, thereby detracting from the flexible power of generic conditional information. ME-modelling, on the other hand, preserves the generic nature of conditionals by minimizing the amount of information being added.

Nevertheless, modelling ME-rule bases has to be done carefully so as to ensure that *all* relevant dependencies are taken into account. This task can be difficult and troublesome. Usually, the modelling rules are based somehow on statistical data. To date, however, no tool is known that helps to find ME-optimal set of rules. In [10], we presented a general approach to solve the *inverse representation problem* which we will now describe for the ME-framework.

1.2 Searching for structures of knowledge

The most usual approach to discover "interesting" rules from data is to look for rules with a significantly high (conditional) probability and a concise antecedent [13, 1, 14]. Basing relevance on frequencies, however, is sometimes unsatisfactory and inadequate, in particular in complex domains like medicine. Furthermore, the rules are considered as isolated pieces of knowledge, no interaction between rules can be taken into account. In order to obtain more structured information, one often searches for causal relationships by investigating conditional independencies and thus non-interactivity between sets of variables [4, 16, 7, 2]. Some of these algorithms also make use of optimization criteria which are based on entropy [8, 6]. Although causality is undoubtedly most important for human understanding, it seems to be too rigid a concept to represent human knowledge in an exhaustive way. For instance, a person suffering from a flu is certainly sick ($P(sick|flu) = 1$), and she often will complain about headache ($P(headache|flu) = 0.9$). Then we have $P(headache|flu) = P(headache|flu \wedge sick)$, but we would surely expect $P(headache|\neg flu) \neq P(headache|\neg flu \wedge sick)$! Although the first equality suggests a conditional independence between *sick* and *headache*, due to the causal dependency between *headache* and *flu*, the second inequality shows this to be (of course) false. Furthermore, a physician might also state some conditional probability involving sick and headache, so that we obtain a complex network of rules. Each of these rules will be considered relevant by the expert, but none will be found when searching for conditional independencies! So, what actually are the "structures of knowledge" by which conditional dependencies (not independencies!) manifest themselves in data? What are the "footprints" conditionals leave on probabilities after they have been learned inductively?

To answer this question, we first take a structural look on conditionals, bare of numerical values, that is, we focus on the set $\mathcal{R} = \{(B_1|A_1), \ldots, (B_n|A_n)\}$ of measure-free conditionals. Conditionals are quite extraordinary pieces of knowledge, flexible and dynamic yet difficult to handle. First, they are objects of a non-boolean nature, demanding for non-classical representation techniques. Second, they are different from material implication in mathematics, in general not describing certain knowledge but plausible conclusions. An adequate way to model its non-classical uncertainty is to represent a conditional $(B|A)$ as a *generalized indicator function* on worlds, setting

$$(B|A)(\omega) = \begin{cases} 1 & : \quad \omega \models AB \\ 0 & : \quad \omega \models A\overline{B} \\ u & : \quad \omega \models \overline{A} \end{cases} \tag{2}$$

where u stands for *unknown* (cf. [5, 3]). When we consider (finite) sets of conditionals $\mathcal{R} = \{(B_1|A_1), \ldots, (B_n|A_n)\}$, we have to modify (2) appropriately to identify the impact of each conditional in $\mathcal{R}$ on worlds in Ω. So to each conditional $(B_i|A_i)$ in $\mathcal{R}$, we associate two abstract symbols $\mathbf{a}_i^+, \mathbf{a}_i^-$, symbolizing a (possibly) positive effect on verifying worlds and a (possibly) negative effect on falsifying worlds:

$$\sigma_i(\omega) = \begin{cases} \mathbf{a}_i^+ & \text{if} \quad (B_i|A_i)(\omega) = 1 \\ \mathbf{a}_i^- & \text{if} \quad (B_i|A_i)(\omega) = 0 \\ 1 & \text{if} \quad (B_i|A_i)(\omega) = u \end{cases} \tag{3}$$

$\mathbf{a}_1^+, \mathbf{a}_1^-, \ldots, \mathbf{a}_n^+, \mathbf{a}_n^-$ are taken as generators of a (free abelian) group $\mathcal{F}_\mathcal{R} = \langle \mathbf{a}_1^+, \mathbf{a}_1^-, \ldots, \mathbf{a}_n^+, \mathbf{a}_n^- \rangle$, so that the neutral group element, 1, corresponds to the neutral behavior of the conditionals on worlds (i.e. $(B_i|A_i)(\omega) = u$). Since all $\sigma_i(\omega)$ are group elements, we can form (arbitrary) products. The function $\sigma_\mathcal{R} : \Omega \to \mathcal{F}_\mathcal{R}$, defined by

$$\sigma_\mathcal{R}(\omega) = \prod_{1 \leqslant i \leqslant n} \sigma_i(\omega) = \prod_{\substack{1 \leqslant i \leqslant n \\ \omega \models A_i B_i}} \mathbf{a}_i^+ \prod_{\substack{1 \leqslant i \leqslant n \\ \omega \models A_i \overline{B}_i}} \mathbf{a}_i^- \tag{4}$$

describes the all-over effect of $\mathcal{R}$ on ω. $\sigma_\mathcal{R}(\omega)$ is called *(a representation of) the conditional structure of* ω *with respect to* $\mathcal{R}$.

Example 1 Let $\mathcal{R} = \{(c|a), (c|b)\}$, where a, b, c are atoms. We associate $\mathbf{a}_1^+, \mathbf{a}_1^-$ with the first conditional, $(c|a)$, and $\mathbf{a}_2^+, \mathbf{a}_2^-$ with the second one, $(c|b)$. Since $\omega = abc$ verifies both conditionals, we obtain $\sigma_\mathcal{R}(abc) = \mathbf{a}_1^+ \mathbf{a}_2^+$. In the same way, e.g., $\sigma_\mathcal{R}(ab\overline{c}) = \mathbf{a}_1^- \mathbf{a}_2^-$, $\sigma_\mathcal{R}(a\overline{b}c) = \mathbf{a}_1^+$ and $\sigma_\mathcal{R}(\overline{a}b\overline{c}) = \mathbf{a}_2^-$. ∎

Notice the striking similarity between (4) and (1) – in (1), the abstract symbols $\mathbf{a}_i^+, \mathbf{a}_i^-$ of (4) have been replaced by the numerical values $\alpha_i^{1-x_i}$ and $\alpha_i^{-x_i}$, respectively (α_0 is simply a normalizing factor). Therefore, the ME-distribution $ME(\mathcal{R}^*)$ follows the conditional structure of worlds with respect to the conditionals in $\mathcal{R}^*$ and is thus most adequate to represent probabilistic conditional knowledge. The α_i's bear the crucial conditional information, and $\alpha_i^{1-x_i}, \alpha_i^{-x_i}$ are the "footprints" left on the probabilities when ME-learning $\mathcal{R}^*$ (also cf. [9]). In the following, we will formalize these ideas and prepare the theoretical ground for the data mining techniques to be presented in the next two sections.

To comply with the group theoretical structure of $\mathcal{F}_{\mathcal{R}}$, we introduce the free abelian group $\widehat{\Omega} := \langle \omega \mid \omega \in \Omega \rangle$, generated by all $\omega \in \Omega$, and consisting of all words $\widehat{\omega} = {\omega_1}^{r_1} \ldots {\omega_m}^{r_m}$ with $\omega_1, \ldots, \omega_m \in \Omega$ and integers $r_1, \ldots r_m$. We will often use fractional representations for the elements of $\widehat{\Omega}$, that is, for instance, we will write $\dfrac{\omega_1}{\omega_2}$ instead of $\omega_1 \omega_2^{-1}$. Now $\sigma_{\mathcal{R}}$ may be extended to $\widehat{\Omega}$ in a straightforward manner by setting $\sigma_{\mathcal{R}}(\widehat{\omega}) = \sigma_{\mathcal{R}}(\omega_1)^{r_1} \ldots \sigma_{\mathcal{R}}(\omega_m)^{r_m}$, yielding a *homomorphism of groups* $\sigma_{\mathcal{R}} : \widehat{\Omega} \to \mathcal{F}_{\mathcal{R}}$. Similarly, we extend probability distributions P to homomorphisms $P : \widehat{\Omega} \to (\mathbb{R}^+, \cdot)$ from $\widehat{\Omega}$ into the non-negative real numbers by setting $P({\omega_1}^{r_1} \ldots {\omega_m}^{r_m}) = P(\omega_1)^{r_1} \cdot \ldots \cdot P(\omega_m)^{r_m}$. Having the same conditional structure defines an equivalence relation $\equiv_{\mathcal{R}}$ on $\widehat{\Omega}$: $\widehat{\omega}_1 \equiv_{\mathcal{R}} \widehat{\omega}_2$ iff $\sigma_{\mathcal{R}}(\widehat{\omega}_1) = \sigma_{\mathcal{R}}(\widehat{\omega}_2)$, i.e. iff $\widehat{\omega}_1 \widehat{\omega}_2^{-1} \in ker\ \sigma_{\mathcal{R}} := \{\widehat{\omega} \in \widehat{\Omega} \mid \sigma_{\mathcal{R}}(\widehat{\omega}) = 1\}$. Thus the kernel of $\sigma_{\mathcal{R}}$ plays an important part in identifying the conditional structure of elements $\widehat{\omega} \in \widehat{\Omega}$.

Besides the explicit representation of knowledge by $\mathcal{R}$, also the implicit normalizing constraint $P(\top|\top) = 1$ has to be taken into account. It is easy to check that $ker\ \sigma_{(\top|\top)} = \widehat{\Omega}_0 := \{\widehat{\omega} = {\omega_1}^{r_1} \cdot \ldots \cdot {\omega_m}^{r_m} \in \widehat{\Omega} \mid \sum_{j=1}^m r_j = 0\}$. Two elements $\widehat{\omega}_1 = \omega_1^{r_1} \ldots \omega_m^{r_m}$, $\widehat{\omega}_2 = \nu_1^{s_1} \ldots \nu_p^{s_p} \in \widehat{\Omega}$ are equivalent modulo $\widehat{\Omega}_0$, $\widehat{\omega}_1 \equiv_{\top} \widehat{\omega}_2$, iff $\widehat{\omega}_1 \widehat{\Omega}_0 = \widehat{\omega}_2 \widehat{\Omega}_0$, i.e. iff $\sum_{1 \leqslant j \leqslant m} r_j = \sum_{1 \leqslant k \leqslant p} s_k$. This means that $\widehat{\omega}_1$ and $\widehat{\omega}_2$ are equivalent modulo $\widehat{\Omega}_0$ iff they both are a (cancelled) product of the same number of generators, each generator being counted with its corresponding exponent. Set $ker_0\ \sigma_{\mathcal{R}} := ker\ \sigma_{\mathcal{R}} \cap \widehat{\Omega}_0 = ker\ \sigma_{\mathcal{R} \cup \{(\top|\top)\}}$.

Now, this formalization allows us to make clear what it means that a distribution P "follows the conditional structures" imposed by $\mathcal{R}$ in the next definition. To make it a very concise one, we focus on the principal case of positive distributions (for the general definition, cf. [12]).

Definition 2 Suppose P is a (positive) probability distribution, and let $\mathcal{R} = \{(B_1|A_1), \ldots, (B_n|A_n)\}$ be a set of conditionals. P is *(conditionally) indifferent with respect to* $\mathcal{R}$ iff $P(\widehat{\omega}_1) = P(\widehat{\omega}_2)$, whenever $\widehat{\omega}_1 \equiv_{\mathcal{R}} \widehat{\omega}_2$ and $\widehat{\omega}_1 \equiv_{\top} \widehat{\omega}_2$, for all $\widehat{\omega}_1, \widehat{\omega}_2 \in \widehat{\Omega}$.

If P is indifferent with respect to $\mathcal{R}$, then it does not distinguish between elements $\widehat{\omega}_1 \equiv_{\top} \widehat{\omega}_2$ with the same conditional structure with respect to $\mathcal{R}$. Conversely, any deviation $P(\widehat{\omega}) \neq 1$ can be explained by the conditionals in $\mathcal{R}$ acting on $\widehat{\omega}$ in a non-balanced way. Note that the notion of indifference only aims at observing conditional structures, without making use of any probabilities associated with the conditionals. The following proposition shows, that conditional indifference establishes a connection between the kernels $ker_0\ \sigma_{\mathcal{R}}$ and $ker_0\ P := \{\widehat{\omega} \in \widehat{\Omega}_0 \mid P(\widehat{\omega}) = 1\}$ which will be crucial to elaborate conditional structures:

Proposition 3 *A probability distribution P is indifferent with respect to a (finite) set $\mathcal{R} \subseteq (\mathcal{L} \mid \mathcal{L})$ of conditionals iff $ker_0\ \sigma_{\mathcal{R}} \subseteq ker_0\ P$.*

Any ME-distribution is indifferent with respect to its generating set of conditionals:

Proposition 4 *Let $\mathcal{R}^*$ be a (finite) set of probabilistic conditionals with structural (i.e. measure-free) counterpart $\mathcal{R} \subset (\mathcal{L} \mid \mathcal{L})$, and let $P^* = ME(\mathcal{R}^*)$ the ME-distribution generated by $\mathcal{R}^*$. Then P^* is indifferent with respect to $\mathcal{R}$.*

The proof of this proposition is straightforward by observing (1).

Example 5 We continue Example 1. Here we observe

$$\sigma_{\mathcal{R}}\left(\frac{abc \cdot \overline{a}\overline{b}\overline{c}}{a\overline{b}c \cdot \overline{a}bc}\right) = \frac{\sigma_{\mathcal{R}}(abc) \cdot \sigma_{\mathcal{R}}(\overline{a}\overline{b}\overline{c})}{\sigma_{\mathcal{R}}(a\overline{b}c) \cdot \sigma_{\mathcal{R}}(\overline{a}bc)} = \frac{\mathbf{a}_1^+\mathbf{a}_2^+ \cdot 1}{\mathbf{a}_1^+ \cdot \mathbf{a}_2^+} = 1,$$

that is, $\frac{abc \cdot \overline{a}\overline{b}\overline{c}}{a\overline{b}c \cdot \overline{a}bc} \in ker_0\ \sigma_{\mathcal{R}}$. Then any ME-representation $P^* = ME(\{(c|a)[x], (c|b)[y]\})$ with $x, y \in [0,1]$ will fulfill $P^*\left(\frac{abc \cdot \overline{a}\overline{b}\overline{c}}{a\overline{b}c \cdot \overline{a}bc}\right) = 1$, i.e. $P^*(abc)P^*(\overline{a}\overline{b}\overline{c}) = P^*(a\overline{b}c)P^*(\overline{a}bc)$. ■

1.3 Data mining and group theory – a strange connection?

The concept of conditional structures, however, is not only an algebraic means to judge well-behavedness with respect to conditionals [11]. As group elements, they make conditional effects on worlds computable and thereby allows us to study interactions between different conditionals in $\mathcal{R}^*$. On (multis)sets of worlds (i.e. elements of $\widehat{\Omega}$), we may observe cancellations or accumulations of conditional impacts which are reflected by the corresponding ME-representation (see Example 5 above). Conversely, finding a set of rules which is able to represent a given probability distribution P via ME-methods can be done by elaborating numerical relationships in P, interpreting them as manifestations of underlying conditional dependencies. The procedure to discover appropriate sets of rules is sketched in the following and will be explained in more detail in the next section:

- Start with a set $\mathcal{B}$ of single-elementary rules the length of which is considered to be large enough to capture all relevant dependencies. Ideally, $\mathcal{B}$ would consist of rules whose antecedents have maximal length (i.e. number of variables -1).
- Search for numerical relationships in P by investigating which products of probabilities match, in order to calculate $ker_0\ P$.
- Compute the corresponding conditional structures with respect to $\mathcal{B}$, yielding equations of group elements in $\mathcal{F}_{\mathcal{B}}$.
- Solve these equations by forming appropriate factor groups of $\mathcal{F}_{\mathcal{B}}$.

- Building these factor groups correspond to eliminating and joining the basic conditionals in $\mathcal{B}$ to make their information more concise, in accordance with the numerical structure of P. Actually, the antecedents of the conditionals in $\mathcal{B}$ are shortened so as to comply with the numerical relationships in P.

As strange as this connection between knowledge discovery and group theory might appear at first sight, it is obvious from an abstract and methodological point of view: Considering knowledge discovery as an operation inverse to inductive knowledge representation, the use of group theoretical means to realize invertability is bold, but straightforward. Moreover, the joint impact of conditionals and their interactions can be symbolized by products and quotients, respectively. Their handling in a group theoretical structure allows a systematic disentangling of highly complex conditional interaction, thereby presenting a completely new approach to discover "structures of knowledge".

2 Discovering conditional structures in data

In this section,we will present an approach to computing sets $\mathcal{R}$, or $\mathcal{R}^*$, respectively, of conditionals that are apt to generate some given (positive) probability function P via ME-presentation. The method to be presented is guided by the following idea: If P is the result of inductively representing a set $\mathcal{R}^*$ of conditionals by applying the ME-principle, $P = ME(\mathcal{R}^*)$, then P is necessarily indifferent with respect to $\mathcal{R}$, i.e. $ker_0\ \sigma_{\mathcal{R}} \subseteq ker_0\ P$ by Proposition 3. Ideally, we would have P to represent $\mathcal{R}$ faithfully, that is,

$$P \models \mathcal{R} \quad \text{and} \quad ker_0\ P = ker_0\ \sigma_{\mathcal{R}}. \tag{5}$$

Assuming faithfulness means presupposing that no equation $P(\widehat{\omega}) = 1$ is fulfilled accidentally, but that any of these equations is induced by $\mathcal{R}$. Thus the structures of the conditionals in $\mathcal{R}$ become manifest in the elements of $ker_0\ P$, that is, in elements $\widehat{\omega} \in \widehat{\Omega}$ with $P(\widehat{\omega}) = 1$. As a further prerequisite, we will assume that this knowledge inherent to P is representable by a set of single-elementary conditionals. This restriction should not be considered as a heavy drawback, bearing in mind the expressibility of single-elementary conditionals.

So assume $\mathcal{R}^* = \{(b_1|A_1)[x_1], \ldots, (b_n|A_n)[x_n]\}$ is an existing, but hidden set of single-elementary conditionals, such that (5) holds. Let us further suppose that $ker_0\ P$ is known from exploiting numerical relationships. Without loss of generality, only to simplify notation, we assume all consequents b_i to be positive literals. Since conditional indifference is a structural notion, we omit the quantifications x_i of the conditionals in what follows. Let $\sigma_{\mathcal{R}} : \widehat{\Omega} \to \mathcal{F}_{\mathcal{R}} = \langle \mathbf{a}_1^+, \mathbf{a}_1^-, \ldots, \mathbf{a}_n^+, \mathbf{a}_n^- \rangle$ denote a conditional structure homomorphism with respect to $\mathcal{R}$.

Our method is a *bottom-up approach* generalizing conditionals in accordance with the conditional structure revealed by $ker_0\ P$. We start with considering basic single-elementary conditionals, which are single-elementary conditionals with antecedents of maximal length. For each atom $v \in \mathcal{L}$, choose an arbitrary, but fixed numbering of the remaining atoms $w \neq v$, $(w_0, w_1, \ldots, w_{\#(atoms)-1})$. Then *basic single-*

elementary conditionals are conditionals of the form

$$\psi_{v,l} = (v \mid \bigwedge_j w_j^{\epsilon_j}) \tag{6}$$

with $\epsilon_j \in \{0,1\}, w_j^1 := w_j, w_j^0 := \overline{w_j}, 0 \leqslant j \leqslant \#(atoms) - 1$ and $l = \sum_j \epsilon_j 2^j$. We will abbreviate the antecedent of $\psi_{v,l}$ by $C_{v,l}$. Let

$$\mathcal{B} = \{\psi_{v,l} \mid v \text{ atom in } \mathcal{L}, 0 \leqslant l \leqslant 2^{\#(atoms)-1} - 1\}$$

denote the set of all basic single-elementary conditionals in $(\mathcal{L} \mid \mathcal{L})$, and let $\mathcal{F}_\mathcal{B} = \langle \mathbf{b}_{v,l}^+, \mathbf{b}_{v,l}^- \mid v \text{ atom in } \mathcal{L}, 0 \leqslant l \leqslant 2^{\#(atoms)-1} - 1\rangle$ be the free abelian group corresponding to $\mathcal{B}$ with conditional structure homomorphism $\sigma_\mathcal{B} : \widehat{\Omega} \to \mathcal{F}_\mathcal{B}$,

$$\sigma_\mathcal{B} = \prod_{v,l} \sigma_{v,l}, \ \sigma_{v,l}(\omega) = \begin{cases} \mathbf{b}_{v,l}^+, & \text{if } \omega = C_{v,l} v \\ \mathbf{b}_{v,l}^-, & \text{if } \omega = C_{v,l} \overline{v} \\ 1, & \text{else} \end{cases}$$

Note that $\sigma_\mathcal{B}$ and $\mathcal{F}_\mathcal{B}$ are known, whereas $\sigma_\mathcal{R}$ and $\mathcal{F}_\mathcal{R}$ are not. We only know the kernel, $ker_0\ \sigma_\mathcal{R}$, of $\sigma_\mathcal{R}$, which is, by assuming faithfulness, the same as the kernel, $ker_0\ P$, of P. Now, to establish a connection between what is obvious ($\mathcal{B}$) and what is searched for ($\mathcal{R}$), we define a homomorphism $g : \mathcal{F}_\mathcal{B} \to \mathcal{F}_\mathcal{R}$ via

$$g(\mathbf{b}_{v,l}^\pm) = \prod_{\substack{1 \leqslant i \leqslant n \\ b_i = v, C_{v,l} \models A_i}} \mathbf{a}_i^\pm, \tag{7}$$

Actually, g is a "phantom" which is not explicitly given, but only known to exist. Its crucial meaning for the knowledge discovery task is revealed by the following theorem and its corollary:

Theorem 6 *Let $g : \mathcal{F}_\mathcal{B} \to \mathcal{F}_\mathcal{R}$ be as in (7). Then $\sigma_\mathcal{R} = g \circ \sigma_\mathcal{B}$.*

Corollary 7 *$\widehat{\omega} \in ker_0\ \sigma_\mathcal{R} = ker_0\ P$ iff $\widehat{\omega} \in \widehat{\Omega}_0$ and $\sigma_\mathcal{B}(\widehat{\omega}) \in ker\ g$.*

This means, that numerical relationships observed in P (and represented by elements of $ker_0\ P$) translate into group theoretical equations modulo the kernel of g:

Proposition 8 *Let $\widehat{\omega} = \omega_1^{r_1} \cdot \ldots \cdot \omega_m^{r_m} \in \widehat{\Omega}_0$. Then $\sigma_\mathcal{B}(\omega_1^{r_1} \cdot \ldots \cdot \omega_m^{r_m}) \in ker\ g$ iff for all atoms v in $\mathcal{L}$,*

$$\prod_{\substack{1 \leqslant k \leqslant m \\ \omega_k = C_{v,l} v}} (\mathbf{b}_{v,l}^+)^{r_k}, \quad \prod_{\substack{1 \leqslant k \leqslant m \\ \omega_k = C_{v,l} \overline{v}}} (\mathbf{b}_{v,l}^-)^{r_k} \in ker\ g. \tag{8}$$

So each (generating) element of $ker_0\ \sigma_\mathcal{R}$ gives rise to an equation modulo $ker\ g$ for the generators $\mathbf{b}_{v,l}^+, \mathbf{b}_{v,l}^-$ of $\mathcal{F}_\mathcal{B}$. Moreover, Proposition 8 allows us to split up equations modulo $ker_0\ g$ to handle each atom separately as a consequent of conditionals.

The idea of our data mining procedure is to exploit the relations mod $ker\ g$ holding between the group elements $\mathbf{b}_{v,l} \in \mathcal{F}_\mathcal{B}$ with the aim to define a finite sequence of sets $\mathcal{S}^{(0)}, \mathcal{S}^{(1)}, \ldots$ of conditionals approximating $\mathcal{R}$:

$$ker_0\ \sigma_{\mathcal{S}^{(0)}} \subseteq ker_0\ \sigma_{\mathcal{S}^{(1)}} \subseteq \ldots \subseteq ker_0\ \sigma_\mathcal{R} \tag{9}$$

The set $\mathcal{B}$ of basic single elementary conditionals proves to be an ideal starting point:

Lemma 9 *$\sigma_{\mathcal{B}}$ is injective, i.e. $ker_0\ \sigma_{\mathcal{B}} = \{1\}$.*

So $\sigma_{\mathcal{B}}$ provides the most finely grained conditional structure on $\widehat{\Omega}$: No different elements $\widehat{\omega}_1 \neq \widehat{\omega}_2$ are equivalent with respect to $\mathcal{B}$. So set $\mathcal{S}^{(0)} := \mathcal{B}$ and $g^{(0)} := g$.

By solving the equations modulo $ker\ g^{(t)}$, $t \geqslant 0$, we obtain a finite sequence of sets $\mathcal{S}^{(t)}$ of single-elementary conditionals $\varphi_{v,j}^{(t)} = (v|D_{v,j}^{(t)})$ with corresponding group generators ${\mathbf{s}_{v,j}^{(t)}}^+, {\mathbf{s}_{v,j}^{(t)}}^-$ such that $ker_0\ \sigma_{\mathcal{S}^{(0)}} \subseteq \ldots ker_0\ \sigma_{\mathcal{S}^{(t)}} \subseteq ker_0\ \sigma_{\mathcal{S}^{(t+1)}} \subseteq \ldots \subseteq ker_0\ \sigma_{\mathcal{R}}$, as desired, together with homomorphisms $g^{(t)}$ describing their relationship to $\mathcal{R}$. Each $\mathcal{S}^{(t+1)}$ arises from $\mathcal{S}^{(t)}$ by eliminating conditionals, or by aggregating conditionals with identical consequents in conjoining their antecedents. These operations are induced by group equations modulo $ker_0\ g^{(t)}$ (equivalence modulo $ker_0\ g^{(t)}$ is denoted by $\equiv_{g^{(t)}}$ in the following, and the $+, -$-signs may be omitted): For instance, if $\mathbf{s}_{v,j}^{(t)} \equiv_{g^{(t)}} 1$, then simply eliminate $\varphi_{v,j}^{(t)}$, and if

$$\mathbf{s}_{v,j_0}^{(t)} \equiv_{g^{(t)}} \mathbf{s}_{v,j_1}^{(t)} \ldots \mathbf{s}_{v,j_m}^{(t)}, \tag{10}$$

then

1. eliminate $\varphi_{v,j_0}^{(t)}$ from $\mathcal{S}^{(t)}$;
2. replace each $\varphi_{v,j_k}^{(t)}$ by $\varphi_{v,j_k}^{(t+1)} = \varphi_{v,j_0}^{(t)} \sqcup \varphi_{v,j_k}^{(t)} = (v \mid D_{v,j_0}^{(t)} \vee D_{v,j_k}^{(t)})$, for $1 \leqslant k \leqslant m$; set $D_{v,j_k}^{(t+1)} = D_{v,j_0}^{(t)} \vee D_{v,j_k}^{(t)}$, $1 \leqslant k \leqslant m$;
3. retain all other $\varphi_{w,l}^{(t)}$.

For the formal details and theoretical justifications, cf. [10]. We will briefly illustrate the method by an example.

Example 10 We consider formulas involving the three atomic propositions a - *being a student*, b - *being young*, and c - *being single (i.e. unmarried)*. The distribution P over a, b, c is given as follows:

ω	$P(\omega)$	ω	$P(\omega)$	ω	$P(\omega)$	ω	$P(\omega)$
abc	0.1950	$ab\overline{c}$	0.1758	$a\overline{b}c$	0.0408	$a\overline{b}\overline{c}$	0.0519
$\overline{a}bc$	0.1528	$\overline{a}b\overline{c}$	0.1378	$\overline{a}\overline{b}c$	0.1081	$\overline{a}\overline{b}\overline{c}$	0.1378

Important relationships between probabilities are revealed by $P(\overline{a}b\overline{c}) = P(\overline{a}\overline{b}\overline{c})$, $\frac{P(abc)}{P(ab\overline{c})} = \frac{P(\overline{a}bc)}{P(\overline{a}b\overline{c})}$, $\frac{P(a\overline{b}c)}{P(a\overline{b}\overline{c})} = \frac{P(\overline{a}\overline{b}c)}{P(\overline{a}\overline{b}\overline{c})}$. We list the twelve basic single-elementary conditionals of $\mathcal{B}$ which we start with:

$$\begin{array}{lll}
\psi_{a,0}{=}(a \mid \overline{b}\overline{c}) & \psi_{b,0}{=}(b \mid \overline{a}\,\overline{c}) & \psi_{c,0}{=}(c \mid \overline{a}\overline{b}) \\
\psi_{a,1}{=}(a \mid \overline{b}c) & \psi_{b,1}{=}(b \mid \overline{a}c) & \psi_{c,1}{=}(c \mid \overline{a}b) \\
\psi_{a,2}{=}(a \mid b\overline{c}) & \psi_{b,2}{=}(b \mid a\overline{c}) & \psi_{c,2}{=}(c \mid a\overline{b}) \\
\psi_{a,3}{=}(a \mid bc) & \psi_{b,3}{=}(b \mid ac) & \psi_{c,3}{=}(c \mid ab)
\end{array}$$

with corresponding generators $\mathbf{b}_{v,l}^+, \mathbf{b}_{v,l}^-$ of $\mathcal{F}_{\mathcal{B}}$. The numerical relationships yield the following group equations :

$$\begin{array}{llll}
\mathbf{b}_{a,2}^-\mathbf{b}_{b,0}^+\mathbf{b}_{c,1}^- & = \sigma_{\mathcal{B}}(\overline{a}b\overline{c}) & = \sigma_{\mathcal{B}}(\overline{a}\overline{b}\overline{c}) & = \mathbf{b}_{a,0}^-\mathbf{b}_{b,0}^-\mathbf{b}_{c,0}^- \\
\mathbf{b}_{a,3}^+\mathbf{b}_{b,3}^+\mathbf{b}_{c,3}^+ \cdot \mathbf{b}_{a,2}^-\mathbf{b}_{b,0}^+\mathbf{b}_{c,1}^- & = \sigma_{\mathcal{B}}(abc)\sigma_{\mathcal{B}}(\overline{a}b\overline{c}) & = \sigma_{\mathcal{B}}(ab\overline{c})\sigma_{\mathcal{B}}(\overline{a}bc) & = \mathbf{b}_{a,2}^+\mathbf{b}_{b,2}^+\mathbf{b}_{c,3}^- \cdot \mathbf{b}_{a,3}^-\mathbf{b}_{b,1}^+\mathbf{b}_{c,1}^+ \\
\mathbf{b}_{a,1}^+\mathbf{b}_{b,3}^-\mathbf{b}_{c,2}^+ \cdot \mathbf{b}_{a,0}^-\mathbf{b}_{b,0}^-\mathbf{b}_{c,0}^- & = \sigma_{\mathcal{B}}(a\overline{b}c)\sigma_{\mathcal{B}}(\overline{a}\overline{b}\overline{c}) & = \sigma_{\mathcal{B}}(a\overline{b}\overline{c})\sigma_{\mathcal{B}}(\overline{a}\overline{b}c) & = \mathbf{b}_{a,0}^+\mathbf{b}_{b,2}^-\mathbf{b}_{c,2}^- \cdot \mathbf{b}_{a,1}^-\mathbf{b}_{b,1}^-\mathbf{b}_{c,0}^+
\end{array}$$

Considering these equations for each atom a, b, c separately and omitting the $\{+,-\}$-signs, we obtain the following equivalences modulo an appropriate subgroup of $\mathcal{F}_{\mathcal{B}}$: $\mathbf{b}_{a,0} \equiv \mathbf{b}_{a,1} \equiv \mathbf{b}_{a,2} \equiv \mathbf{b}_{a,3}$, $\mathbf{b}_{c,0} \equiv \mathbf{b}_{c,1} \equiv \mathbf{b}_{c,2} \equiv \mathbf{b}_{c,3}$, $\mathbf{b}_{b,0} \equiv 1$, $\mathbf{b}_{b,3} \equiv \mathbf{b}_{b,1}\mathbf{b}_{b,2}$.

Eliminating $\psi_{b,0}$ and joining conditionals according to these equations results in the set $\mathcal{S} = \{(a|\top), (c|\top), (b|a), (b|c)\}$ of (unquantified) conditionals. Associating the proper probabilities with these structural conditionals, we obtain $\mathcal{S}^* = \{(a|\top)[0.4635], (c|\top)[0.4967], (b|a)[0.8], (b|c)[0.7]\}$ as an ME-generating set for P, i.e. $P = ME(\mathcal{S}^*)$. ■

3 Conclusion

In this paper, a new method for discovering conditional dependencies from data has been developed. This method is based on information-theoretical concepts and group-theoretical techniques, taking knowledge discovery as an operation inverse to inductive knowledge representation. By investigating relationships between the numerical values of a probability distribution P, the effects of conditionals are analyzed and isolated, and conditionals are joined suitably so as to fit the knowledge structures inherent to P.

The applicability of the method presented in this paper neither depends on the presupposition of P being a faithful respresentation nor on having a complete description of the kernel $ker_0\ P$ available: Each numerical relationship found amongst the values of P corresponds to an element of $ker_0\ P$ and may be used to set up equations for the corresponding group elements. If P fails to be a faithful representation of some suitable set of conditionals, then too many equations will have to be solved trivially. In this case, backtracking will be necessary, undoing the last joining of conditionals.

Our method is a bottom-up approach, starting from a set $\mathcal{B}$ of most basic rules. Ideally, $\mathcal{B}$ contains conditionals of a simple structure and of maximal length (cf. Example 10) which seems to make the handling of $\mathcal{B}$ impossible. By applying the mining method to the *null-worlds* (i.e. the possible worlds with probability 0, or the empty cells in a contingency table, respectively), however, $\mathcal{B}$ can be shrunken to a tractable size in advance.

References

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. Fast discovery of association rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in knowledge discovery and data mining*, pages 307–328. MIT Press, Cambridge, Mass., 1996.

[2] W. Buntine. A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 8(2):195–210, 1996.

[3] P.G. Calabrese. Deduction and inference using conditional logic and probability. In I.R. Goodman, M.M. Gupta, H.T. Nguyen, and G.S. Rogers, editors, *Conditional Logic in Expert Systems*, pages 71–100. Elsevier, North Holland, 1991.

[4] G.F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9:309–347, 1992.

[5] B. DeFinetti. *Theory of Probability*, volume 1,2. John Wiley and Sons, New York, 1974.

[6] D. Geiger. An entropy-based learning algorithm of bayesian conditional trees. In *Proceedings Eighth Conference on Uncertainty in Artificial Intelligence*, pages 92–97, 1992.

[7] D. Heckerman. Bayesian networks for knowledge discovery. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in knowledge discovery and data mining.* MIT Press, Cambridge, Mass., 1996.

[8] E. Herskovits and G. Cooper. Kutató: An entropy-driven system for construction of probabilistic expert systems from databases. Technical Report KSL-90-22, Knowledge Systems Laboratory, 1990.

[9] G. Kern-Isberner. Characterizing the principle of minimum cross-entropy within a conditional-logical framework. *Artificial Intelligence*, 98:169–208, 1998.

[10] G. Kern-Isberner. Solving the inverse representation problem. In *Proceedings 14th European Conference on Artificial Intelligence, ECAI'2000*, pages 581–585, Berlin, 2000. IOS Press.

[11] G. Kern-Isberner. Conditional preservation and conditional indifference. *Journal of Applied Non-Classical Logics*, 2001. (to appear).

[12] G. Kern-Isberner. *Conditionals in nonmonotonic reasoning and belief revision.* Springer, Lecture Notes in Artificial Intelligence, 2001. (to appear).

[13] G. Kern-Isberner and H.P. Reidmacher. Interpreting a contingency table by rules. *International Journal of Intelligent Systems*, 11(6), 1996.

[14] N. Megiddo and R. Srikant. Discovering predictive association rules. In *Proceedings of the 4th International Conference on Knowledge Discovery in Databases and Data Mining*, 1998.

[15] J.B. Paris. *The uncertain reasoner's companion – A mathematical perspective.* Cambridge University Press, 1994.

[16] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search.* Number 81 in Lecture Notes in Statistics. Springer, New York Berlin Heidelberg, 1993.