Knowledge Discovery by Reversing Inductive Knowledge Representation[*]

Gabriele Kern-Isberner[†]
Department of Computer Science
FernUniversitaet Hagen, 58084 Hagen, Germany

Jens Fisseler[‡]
Institute for Robotics and Cognitive Systems
Universitaet Luebeck, 23538 Luebeck, Germany

Abstract

In a very basic sense, the aim of knowledge discovery is to reveal *structures of knowledge* which can be seen as being represented by *structural relationships*. In this paper, we make this notion more precise and present a method how to extract structural information from statistical data. There are two key ideas underlying our approach: First, knowledge discovery is understood as a process which is inverse to inductive knowledge representation. So the relevance of discovered information is judged with respect to the chosen representation method. Second, the link between structural and numerical knowledge is established by an algebraic theory of conditionals, which considers conditionals as agents acting on possible worlds. By applying this theory, we develop an algorithm that computes sets of probabilistic rules from distributions. In particular, we show how sparse information can be dealt with appropriately in our framework. The inductive representation method used here is based on information theory, so that the discovered rules can be considered as being most informative in a strict, formal sense.

Introduction

In a very basic sense, the aim of knowledge discovery is to reveal *structures of knowledge* from data, which make fundamental relationships explicit and focus on relevant aspects. A basic means to formalize such relationships are rules, connecting a precondition and a conclusion by an *if-then-construction*. Such *rules* or *conditionals* are widely used for knowledge representation and reasoning. They should be clearly distinguished from (material) implications, as they are able to represent also *plausible relationships* or *default rules* (for a deeper discussion of this topic and further references, cf. e.g. (Kern-Isberner 2001b; Nute & Cross 2002; Benferhat, Dubois, & Prade 1997)). The crucial point with conditionals is that they carry generic knowledge which can be applied to different situations. This makes them most interesting objects for knowledge representation in general, in theoretical as well as in practical respect. Conditionals can be specified further by degrees of plausibility, possibility, probability and the like. In particular, probability theory provides a solid mathematical framework for conditionals which is often used for statistical knowledge discovery (cf. e.g. (Spirtes, Glymour, & Scheines 1993; Cheeseman & Oldford 1994)). The semantics that is usually associated with conditionals in this field is a frequentistic one, and their relevance is measured in terms of *confidence* and *support* (see, e.g., (Agrawal *et al.* 1996)). When searching for structures in probabilistic data, causality is often appreciated as a most appropriate framework (Pearl 1988; Spirtes, Glymour, & Scheines 1993; Cowell *et al.* 1999). While, on the one hand, frequentistic criteria are sometimes found a bit weak to embody relevance, causality, on the other hand, is too rigid a concept to fit plausible relationships which are also most relevant for human reasoning.

In this paper, we present a method to discover structures imposed by plausible cognitive links from data. Relevance is understood with respect to informativeness, a notion which is based here on solid information-theoretical grounds. In short, our aim is to find *most informative rules from data*. In more detail, we assume the probabilistic distribution provided by the statistical data to be generated from some basic set of conditionals via the *principle of maximum entropy* (*ME*), and we develop an algorithm to find such a generating set of conditionals. The ME-methodology provides techniques to represent incomplete probabilistic information inductively by a probability distribution and allows nonmonotonic, semantic-based inferences (cf. (Jaynes 1983; Paris 1994)). The *structures of knowledge* which ME-representations follow have been made explicit in (Kern-Isberner 1998) and have provided the grounds for developing a new algebraic theory of conditionals (Kern-Isberner 2001b). Different from former, mainly logical approaches to put conditional reasoning in formal terms, here conditionals are considered as agents acting on possible worlds. This theory formalizes precisely what *conditional structures* are, and how they can be used to handle complex interactions of conditionals. In contrast to causal approaches to knowledge representation and discovery, our focus is on *conditional dependencies*, not on *conditional independencies* which are usually assumed to underly the concept of causality. In (Kern-Isberner 2003), it is shown that our concept is strictly more

[*] This work is supported by the German Research Society, DFG, under grant BE 1700/5-1

[†] gabriele.kern-isberner@fernuni-hagen.de

[‡] fisseler@rob.uni-luebeck.de

general than conditional independence.

Our method is a *bottom-up approach*, starting with conditionals with long premisses, and shortening these premises to make the conditionals most expressive but without losing information, in accordance with the information inherent to the data. In particular, the main features of the approach can be described as follows:

- The method is based on statistical information but not on probabilities close to 1; actually, it mostly uses only structural information obtained from the data;
- it is able to disentangle highly complex interactions between conditionals.
- We are going to discover not single, isolated rules but a set of rules, thus taking into regard the collective effects of several conditionals.
- Zero probabilities computed from data are interpreted as missing information, not as certain knowledge.

From a more foundational point of view, this paper presents and elaborates quite an unusual approach to knowledge discovery: Here, knowledge discovery is understood as a process which reverses inductive knowledge representation. We use the ME-principle as a vehicle to represent incomplete probabilistic knowledge inductively, and show how to solve the *inverse maxent problem* (i.e. computing ME-generating conditionals from a probability distribution). This paper continues work begun in (Kern-Isberner 2000; 2001b).

The organization of this paper is as follows: The following section summarizes basic facts concerning probabilistic logic and the maximum entropy approach. Then we sketch the main features of the algebraic theory of conditionals which is based on group theory and provides the grounds for our approach to knowledge discovery; the resulting method is briefly described afterwards. Then we present the CKD-algorithm (*CKD = Conditional Knowledge Discovery*) which is illustrated by an example, and go into implementation details. Finally, we conclude this paper with a summary and an outlook on further and practical work.

Probabilistic logic and maximum entropy

We consider a propositional framework over a finite set $\mathcal{V} = \{V_1, V_2, \ldots\}$ of (multivalued) propositional variables V_i with finite domains. For each variable $V_i \in \mathcal{V}$, the values are denoted by v_i. In generalizing the bivalued propositional framework, we call expressions of the form $V_i = v_i$ *literals*, and abbreviate them by v_i. The language $\mathcal{L}$ consists of all formulas A built by conjoining finitely many literals by conjunction ($\wedge$), disjunction ($\vee$), and negation ($\neg$) in a well-formed way. The conjunction operator, $\wedge$, will usually be omitted, so AB will mean $A \wedge B$, and negation is indicated by barring, i.e. $\overline{A} = \neg A$. An *elementary conjunction* is a conjunction consisting of literals, and a *complete conjunction* is an elementary conjunction where each variable from $\mathcal{V}$ is represented by exactly one value. Let Ω denote the set of complete conjunctions of $\mathcal{L}$. Ω can be taken as the set of *possible worlds* ω, providing a complete description of each possible state, and hence corresponding to elementary events in probability theory.

Conditionals are written in the form $(B|A)$, with antecedents, A, and consequents, B, both formulas in $\mathcal{L}$, and may be read as *if A then B*. Let $(\mathcal{L} \mid \mathcal{L})$ denote the set of all conditionals over $\mathcal{L}$. *Single-elementary conditionals* are conditionals whose antecedents are elementary conjunctions, and whose consequents consist of one single literal.

Let P be a probability distribution over $\mathcal{V}$. Within a probabilistic framework, conditionals can be quantified and interpreted probabilistically via conditional probabilities:

$$P \models (B|A)\,[x] \quad \text{iff} \quad P(A) > 0 \text{ and } P(AB) = xP(A)$$

for $x \in [0,1]$. If $\mathcal{R}^* = \{(B_1|A_1)\,[x_1], \ldots, (B_n|A_n)\,[x_n]\}$ is a set of probabilistic conditionals, then $\mathcal{R} = \{(B_1|A_1), \ldots, (B_n|A_n)\}$ denotes the set of structural (i.e. unquantified) conditionals.

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 %*", this knowledge being formally expressed by $\mathcal{R}^* = \{(\textit{young}\,|\textit{student})[0.8], (\textit{young}\,|\textit{single})[0.7]\}$. Usually, such 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 (Paris 1994) 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 $P^* = \textit{ME}(\mathcal{R}^*)$ 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. $\textit{ME}(\mathcal{R}^*)$ can be written in the form

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

with the α_i's being chosen appropriately so as to satisfy all of the conditional constraints in $\mathcal{R}^*$ (cf. (Kern-Isberner 1998)); $\textit{ME}(\mathcal{R}^*)$ is called the *ME-representation of* $\mathcal{R}^*$.

The ME-principle provides a most convenient and theoretically sound method to represent incomplete probabilistic knowledge.[1] 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. In fact, 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 and transferable 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. So, a method to compute rule sets appropriate for ME-modelling from statistical data is urgently needed.

Conditional structures and conditional indifference

In order to obtain structural information from data, one usually searches for causal relationships by investigating conditional independencies and thus non-interactivity between sets of variables (Cooper & Herskovits 1992; Spirtes, Glymour, & Scheines 1993; Heckerman 1996; Buntine 1996). Some of these algorithms also make use of optimization criteria which are based on entropy (Herskovits & Cooper 1990; Geiger 1992). 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 ($P(\mathit{sick}\,|\mathit{flu}) = 1$), and they often will complain about headache ($P(\mathit{headache}\,|\mathit{flu}) = 0.9$). Then we have

$$P(\mathit{headache}\,|\mathit{flu}) = P(\mathit{headache}\,|\mathit{flu} \wedge \mathit{sick}),$$

but we would surely expect

$$P(\mathit{headache}\,|\neg \mathit{flu}) \neq P(\mathit{headache}\,|\neg \mathit{flu} \wedge \mathit{sick})!$$

Although, from a naïve point of view, 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 wish to state some conditional probability involving *sick* and *headache*, so that we would 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 use the approach developed in (Kern-Isberner 2000; 2001b); all proofs and lots of examples can be found in (Kern-Isberner 2001b). We first take a structural look on conditionals, bare of numerical values, that is, we focus on sets $\mathcal{R} = \{(B_1|A_1), \ldots, (B_n|A_n)\}$ of measure-free conditionals. A well-known approach to model its non-classical uncertainty is to represent a conditional $(B|A)$ as a three-valued indicator function on worlds

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

where u stands for *unknown* (cf., e.g., (DeFinetti 1974; Calabrese 1991)). Two conditionals are *equivalent* iff they yield the same indicator function, so that $(B|A) \equiv (D|C)$ iff $AB \equiv CD$ and $A\overline{B} \equiv C\overline{D}$.

We generalize this approach a bit by associating to each conditional $(B_i|A_i)$ in $\mathcal{R}$ 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:

$$\sigma_i(\omega) = \begin{cases} \mathbf{a}_i^+ & \text{if} \quad \omega \models A_i B_i \\ \mathbf{a}_i^- & \text{if} \quad \omega \models A_i \overline{B_i} \\ 1 & \text{if} \quad \omega \models \overline{A_i} \end{cases} \tag{2}$$

with 1 being the neutral element of the (free abelian) group $\mathcal{F}_\mathcal{R} = \langle \mathbf{a}_1^+, \mathbf{a}_1^-, \ldots, \mathbf{a}_n^+, \mathbf{a}_n^- \rangle$, generated by all symbols $\mathbf{a}_1^+, \mathbf{a}_1^-, \ldots, \mathbf{a}_n^+, \mathbf{a}_n^-$. The function $\sigma_\mathcal{R} : \Omega \rightarrow \mathcal{F}_\mathcal{R}$, defined by

$$\sigma_\mathcal{R}(\omega) = \prod_{1 \leqslant i \leqslant n} \sigma_i(\omega) = \prod_{\substack{1 \leqslant i \leqslant n \\ \omega \models A_i B_i}} \mathbf{a}_i^+ \prod_{\substack{1 \leqslant i \leqslant n \\ \omega \models A_i \overline{B}_i}} \mathbf{a}_i^- \tag{3}$$

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

Example 1 Let $\mathcal{R} = \{(c|a), (c|b)\}$, where A, B, C are bivalued propositional variables with outcomes $\{a, \overline{a}\}, \{b, \overline{b}\}$ and $\{c, \overline{c}\}$, respectively, and let $\mathcal{F}_\mathcal{R} = \langle \mathbf{a}_1^+, \mathbf{a}_1^-, \mathbf{a}_2^+, \mathbf{a}_2^- \rangle$. 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 (3) and (1) – in (1), the abstract symbols $\mathbf{a}_i^+, \mathbf{a}_i^-$ of (3) 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

[1] Efficient implementations of ME-systems can be found via `www.fernuni-hagen.de/BWLOR/forsch.htm` and `www.pit-systems.de`

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. (Kern-Isberner 1998)). In the following, we will put these ideas in formal, algebraic terms and prepare the theoretical ground for the data mining techniques to be presented in this paper.

Let $\widehat{\Omega} := \langle \omega \mid \omega \in \Omega \rangle$ be the free abelian group generated by all $\omega \in \Omega$, and consisting of all products $\widehat{\omega} = {\omega_1}^{r_1} \ldots {\omega_m}^{r_m}$ with $\omega_1, \ldots, \omega_m \in \Omega$ and integers $r_1, \ldots r_m$. Note that, although we speak of *multiplication*, the worlds in such a product are merely juxtaposed, forming a *word* rather than a *product*. With this understanding, a *generalized world* $\widehat{\omega} \in \widehat{\Omega}$ in which only positive exponents occur simply corresponds to a multi-set of worlds. 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

$$\sigma_{\mathcal{R}}({\omega_1}^{r_1} \ldots {\omega_m}^{r_m}) = \sigma_{\mathcal{R}}(\omega_1)^{r_1} \ldots \sigma_{\mathcal{R}}(\omega_m)^{r_m}$$

yielding a *homomorphism of groups* $\sigma_{\mathcal{R}} : \widehat{\Omega} \to \mathcal{F}_{\mathcal{R}}$.

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 $\widehat{\omega} \in \widehat{\Omega}$. *ker* $\sigma_{\mathcal{R}}$ contains exactly all group elements $\widehat{\omega} \in \widehat{\Omega}$ with a balanced conditional structure, that means, where all effects of conditionals in $\mathcal{R}$ on worlds occurring in $\widehat{\omega}$ are completely cancelled. Since $\mathcal{F}_{\mathcal{R}}$ is free abelian, no nontrivial relations hold between the different group generators $\mathbf{a}_1^+, \mathbf{a}_1^-, \ldots, \mathbf{a}_n^+, \mathbf{a}_n^-$ of $\mathcal{F}_{\mathcal{R}}$, so we have $\sigma_{\mathcal{R}}(\widehat{\omega}) = 1$ iff $\sigma_i(\widehat{\omega}) = 1$ for all i, $1 \leqslant i \leqslant n$, and this means

$$\textit{ker}\ \sigma_{\mathcal{R}} = \bigcap_{i=1}^{n} \textit{ker}\ \sigma_i$$

In this way, each conditional in $\mathcal{R}$ contributes to *ker* $\sigma_{\mathcal{R}}$.

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* $\sigma_{(\top|\top)} = \widehat{\Omega}_0$, with

$$\widehat{\Omega}_0 := \{\widehat{\omega} = {\omega_1}^{r_1} \cdot \ldots \cdot {\omega_m}^{r_m} \in \widehat{\Omega} \mid \sum_{j=1}^{m} r_j = 0\}$$

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

$$\textit{ker}_0\ \sigma_{\mathcal{R}} := \textit{ker}\ \sigma_{\mathcal{R}} \cap \widehat{\Omega}_0 = \textit{ker}\ \sigma_{\mathcal{R} \cup \{(\top|\top)\}}$$

In the following, if not stated otherwise, we will assume that all probability distributions are positive. For the methods to be described, this is but a technical prerequisite, permitting a more concise presentation of the basic ideas. The general case may be dealt with in a similar manner (cf. (Kern-Isberner 1999)). Positive distributions P may be extended to homomorphisms $P : \widehat{\Omega} \to (\mathbb{R}^+, \cdot)$ from $\widehat{\Omega}$ into the multiplicative group of non-negative real numbers in a straightforward way by setting

$$P({\omega_1}^{r_1} \ldots {\omega_m}^{r_m}) = P(\omega_1)^{r_1} \cdot \ldots \cdot P(\omega_m)^{r_m}$$

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 both $\widehat{\omega}_1 \equiv_{\mathcal{R}} \widehat{\omega}_2$ and $\widehat{\omega}_1 \equiv_{\top} \widehat{\omega}_2$ hold for $\widehat{\omega}_1, \widehat{\omega}_2 \in \widehat{\Omega}$.

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

$$\textit{ker}_0\ P := \{\widehat{\omega} \in \widehat{\Omega}_0 \mid P(\widehat{\omega}) = 1\}$$

which will be crucial to elaborate conditional structures:

Proposition 3 *A probability distribution P is indifferent with respect to a set $\mathcal{R} \subseteq (\mathcal{L} \mid \mathcal{L})$ of conditionals iff ker$_0$ $\sigma_{\mathcal{R}} \subseteq$ ker$_0$ P.*

If *ker*$_0$ $\sigma_{\mathcal{R}} =$ *ker*$_0$ P, then $P(\widehat{\omega}_1) = P(\widehat{\omega}_2)$ iff $\sigma_{\mathcal{R}}(\widehat{\omega}_1) = \sigma_{\mathcal{R}}(\widehat{\omega}_2)$, for $\widehat{\omega}_1 \equiv_{\top} \widehat{\omega}_2$. In this case, P completely follows the conditional structures imposed by $\mathcal{R}$ – it observes $\mathcal{R}$ *faithfully*.

The next theorem characterizes indifferent probability functions:

Theorem 4 *A (positive) probability function P is indifferent with respect to a set $\mathcal{R} = \{(B_1|A_1), \ldots, (B_n|A_n)\} \subseteq (\mathcal{L} \mid \mathcal{L})$ iff there are positive real numbers $\alpha_0, \alpha_1^+, \alpha_1^-, \ldots, \alpha_n^+, \alpha_n^- \in \mathbb{R}^+$, such that*

$$P(\omega) = \alpha_0 \prod_{\substack{1 \leqslant i \leqslant n \\ \omega \models A_i B_i}} \alpha_i^+ \prod_{\substack{1 \leqslant i \leqslant n \\ \omega \models A_i \overline{B_i}}} \alpha_i^-,\ \omega \in \Omega. \qquad (4)$$

Any ME-distribution is indifferent with respect to its generating set of conditionals, as is obvious by observing (1):

Proposition 5 *Let $\mathcal{R}^*$ be a (finite) set of probabilistic conditionals with structural counterpart $\mathcal{R} \subseteq (\mathcal{L} \mid \mathcal{L})$, and let $P^* =$ ME$(\mathcal{R}^*)$ the ME-distribution generated by $\mathcal{R}^*$. Then P^* is indifferent with respect to $\mathcal{R}$.*

Example 6 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, $\frac{abc \cdot \overline{a}\overline{b}\overline{c}}{a\overline{b}c \cdot \overline{a}bc} \in \mathit{ker}_0\ \sigma_{\mathcal{R}}$. Then any ME-representation $P^* = \mathit{ME}(\{(c|a)[x], (c|b)[y]\})$ 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. $P^*(abc)P^*(\overline{a}\overline{b}\overline{c}) = P^*(a\overline{b}c)P^*(\overline{a}bc)$, no matter what conditional probabilities $x, y \in [0,1]$ have been chosen. ■

In (Kern-Isberner 2003), we investigate the exact relationship between *conditional indifference* and *conditional independence* and show that conditional indifference is the strictly more general concept.

Data mining and group theory – a strange connection?

Before going into more details and presenting the knowledge discovery strategy, let us stop for a moment to contemplate what all this formal machinery is good for. The concept of conditional structures is not only an algebraic means to judge well-behavedness with respect to conditionals (Kern-Isberner 2001a). As group elements, they make conditional effects on worlds transparent and computable and thereby allow 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 6 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 (i.e. simple association 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. #(*variables*) $- 1$).
- Search for numerical relationships in P by investigating which products of probabilities match, in order to calculate $\mathit{ker}_0\ P$.
- 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 nearly 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 offering quite a new (and a bit unusual) view on discovering "structures of knowledge".

Discovering conditional structures in data

In this section, we will describe our approach to knowledge discovery which is based on the group theoretical, algebraic theory of conditionals sketched above. More precisely, we will show how to compute sets $\mathcal{R}$, or $\mathcal{R}^*$, respectively, of conditionals that are apt to generate some given (positive) probability function P via ME-presentation. More details and all proofs can be found in (Kern-Isberner 2001b); the generalization to multivalued variables (instead of bivalued variables) is straightforward.

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, $P = \mathit{ME}(\mathcal{R}^*)$, then P is necessarily indifferent with respect to $\mathcal{R}$, i.e. $\mathit{ker}_0\ \sigma_{\mathcal{R}} \subseteq \mathit{ker}_0\ P$ by Proposition 3. Ideally, we would have P to represent $\mathcal{R}$ faithfully, that is,

$$P \models \mathcal{R} \text{ and } \mathit{ker}_0\ P = \mathit{ker}_0\ \sigma_{\mathcal{R}} \qquad (5)$$

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 $\mathit{ker}_0\ P$, 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 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 $\mathit{ker}_0\ P$ (or parts of it) is known from exploiting numerical relationships. 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 $\mathcal{R}$.

Besides conditional structures, a further notion which is crucial to study and exploit conditional interactions is that of subconditionals: $(D|C)$ is called a *subconditional* of $(B|A)$, and $(B|A)$ is a *superconditional* of $(D|C)$, written as $(D|C) \sqsubseteq (B|A)$, iff $CD \models AB$ and $C\overline{D} \models A\overline{B}$, that is, iff all worlds verifying (falsifying) $(D|C)$ also verify (falsify) $(B|A)$. For any two conditionals $(B|A), (D|C) \in (\mathcal{L} \mid \mathcal{L})$ with $ABC\overline{D} \equiv A\overline{B}CD \equiv \perp$, the supremum $(B|A) \sqcup (D|C)$ in $(\mathcal{L} \mid \mathcal{L})$ with respect to $\sqsubseteq$ exists and is given by

$$(B|A) \sqcup (D|C) \equiv (AB \vee CD|A \vee C)$$

(cf. (Kern-Isberner 1999)). In particular, for two conditionals $(B|A), (B|C)$ with the same consequent, we have

$$(B|A) \sqcup (B|C) \equiv (B|A \vee C)$$

The following lemma provides an easy characterization for the relation $\sqsubseteq$ to hold between single-elementary conditionals:

Lemma 7 *Let $(b|A)$ and $(d|C)$ be two single-elementary conditionals. Then $(d|C) \sqsubseteq (b|A)$ iff $C \models A$ and $b = d$.*

This lemma may be generalized slightly to hold for conditionals $(b|A)$ and $(d|C)$ where A and C are disjunctions of conjunctions of literals not containing b and d, respectively.

From (2), Definition 2 and Proposition 3, it is clear that in an inductive reasoning process such as ME-propagation that results in an indifferent representation of conditional knowledge $\mathcal{R}$, all subconditionals of conditionals in $\mathcal{R}$ also exert the same effects on possible worlds as the corresponding superconditionals. The basic idea is to start with most basic conditionals, and to generalize them step-by-step to superconditionals in accordance with the conditional structure revealed by *ker*$_0$ P. From a theoretical point of view, the most adequate candidates for rules to start with are *basic single-elementary conditionals*, which are single-elementary conditionals with antecedents of maximal length:

$$\psi_{v,l} = (v \mid C_{v,l}) \tag{6}$$

where v is a value of some variable $V \in \mathcal{V}$ and $C_{v,l}$ is an elementary conjunction consisting of literals involving all variables from $\mathcal{V}$ except V. It is clear that considering all such conditionals is intractable, but we are still on theoretical grounds, so let us assume for the moment we could start with the set

$$\mathcal{B} = \{\psi_{v,l} \mid v \in \mathcal{V}, l \text{ suitable}\}$$

of all basic single-elementary conditionals in $(\mathcal{L} \mid \mathcal{L})$, and let $\mathcal{F}_\mathcal{B} = \langle \mathbf{b}_{v,l}^+, \mathbf{b}_{v,l}^- \rangle_{v,l}$ be the free abelian group corresponding to $\mathcal{B}$ with conditional structure homomorphism $\sigma_\mathcal{B} : \widehat{\Omega} \to \mathcal{F}_\mathcal{B}$. 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 (5), 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

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

where the second equality holds due to Lemma 7. g uses the subconditional-relationship in collecting for each basic conditional in $\mathcal{B}$ the effects of the corresponding superconditionals in $\mathcal{R}$. Actually, g is a “phantom” which is not explicitly given, but only assumed to exist. Its crucial meaning for the knowledge discovery task is revealed by the following theorem:

Theorem 8 *Let $g : \mathcal{F}_\mathcal{B} \to \mathcal{F}_\mathcal{R}$ be as in (7). Then*

$$\sigma_\mathcal{R} = g \circ \sigma_\mathcal{B}$$

In particular, $\widehat{\omega} \in$ 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 9 *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 literals v in $\mathcal{L}$,*

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

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 9 allows us to split up equations modulo *ker*$_0$ g to handle each literal separately as a consequent of conditionals, and to separate positive from negative effects. These separations are possible due to the property of the involved groups of being free abelian, and they are crucial to disentangle conditional interactions (cf. also (Kern-Isberner 2001b)).

Now the aim of our data mining procedure can be made more precise: We are going to define a finite sequence of sets $\mathcal{S}^{(0)}, \mathcal{S}^{(1)}, \ldots$ of conditionals approximating $\mathcal{R}$, in the sense that

$$\textit{ker}_0\ \sigma_{\mathcal{S}^{(0)}} \subseteq \textit{ker}_0\ \sigma_{\mathcal{S}^{(1)}} \subseteq \ldots \subseteq \textit{ker}_0\ \sigma_\mathcal{R} \tag{9}$$

The set $\mathcal{B}$ of basic single elementary conditionals proves to be an ideal starting point $\mathcal{S}^{(0)}$:

Lemma 10 *$\sigma_\mathcal{B}$ 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 $\mathcal{B}$.

Step by step, the relations mod *ker* g holding between the group elements are exploited with the aim to construct $\mathcal{S}^{(t+1)}$ from $\mathcal{S}^{(t)}$ by eliminating or joining conditionals by $\sqcup$, in accordance with the equations modulo *ker* g (i.e., by assumption, with the numerical relationships found in P). Each $\mathcal{S}^{(t)}$ is assumed to be a set of conditionals $\varphi_{v,j}^{(t)}$ with a single literal v in the conclusion, and the antecedent $D_{v,j}^{(t)}$ of $\varphi_{v,j}^{(t)}$ is a disjunction of elementary conjunctions not mentioning the variable V. Let $\mathcal{F}_{\mathcal{S}^{(t)}} = \langle {\mathbf{s}_{v,j}^{(t)}}^+, {\mathbf{s}_{v,j}^{(t)}}^- \rangle_{v,j}$ be the free abelian group associated with $\mathcal{S}^{(t)}$, and $\sigma_{\mathcal{S}^{(t)}} : \widehat{\Omega} \to \mathcal{F}_{\mathcal{S}^{(t)}}$ the corresponding structure homomorphism; let $g^{(t)} : \mathcal{F}_{\mathcal{S}^{(t)}} \to \mathcal{F}_\mathcal{R}$ be the homomorphism defined by

$$g^{(t)}({\mathbf{s}_{v,j}^{(t)}}^{\pm}) = \prod_{\substack{1 \leqslant i \leqslant n \\ v = b_i, D_{v,j}^{(t)} \models A_i}} \mathbf{a}_i^{\pm}$$

such that $g^{(t)} \circ \sigma_{\mathcal{S}^{(t)}} = \sigma_\mathcal{R}$. Let $\equiv_{g^{(t)}}$ denote the equivalence relation modulo *ker* $g^{(t)}$, i.e. $\mathbf{s}_1 \equiv_{g^{(t)}} \mathbf{s}_2$ iff $g^{(t)}(\mathbf{s}_1) = g^{(t)}(\mathbf{s}_2)$ for any two group elements $\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{F}_{\mathcal{S}^{(t)}}$. In the following, for ease of notation, we will omit the $+, -$ superscripts on group generators; this is justified, since, by Proposition 9, only one $\{+, -\}$-type of generators is assumed to occur in the equations to be dealt with in the sequel. It is clear that all equations can be transformed such that on either side, only generators with positive exponents occur.

The basic type of equation that arises from *ker*$_0$ P by applying Theorem 8 and the faithfulness assumption (5) is of the form

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

To obtain the new set $\mathcal{S}^{(t+1)}$ by solving this equation, the following steps have to be done:

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

This also includes the case $m = 0$, i.e. $\varphi_{v,j_0}^{(t)} \equiv_{g^{(t)}} 1$; in this case, Step 2 is vacuous and therefore is left out.

It can be shown (cf. (Kern-Isberner 2001b)) that

$$g^{(t+1)} \circ \sigma_{\mathcal{S}^{(t+1)}} = \sigma_{\mathcal{R}}$$

and hence

$$\mathit{ker}_0\ \sigma_{\mathcal{S}^{(t)}} \subseteq \mathit{ker}_0\ \sigma_{\mathcal{S}^{(t+1)}} \subseteq \mathit{ker}_0\ \sigma_{\mathcal{R}}$$

as desired. Moreover, *ker* $g^{(t+1)}$ can be obtained directly from *ker* $g^{(t)}$ by straightforward modifications. Since the considered equation has been solved, it can be eliminated, and other equations may simplify.

Now, that the theoretical background and the basic techniques have been described, we will turn to develop an algorithm for conditional knowledge discovery.

The CKD-algorithm

In this section, we will describe the algorithm *CKD* (= *Conditional Knowledge Discovery*) for mining probabilistic conditionals from statistical data which has been implemented in the CONDOR-system (for an overview, cf. (Beierle & Kern-Isberner 2003)) and is sketched in Figure 1. The resulting set of conditionals will reveal relevant relationships and may serve to represent inductively the corresponding probability distribution via the ME-principle.

```
Algorithm CKD
(Conditional Knowledge Discovery)

Input   A frequency/probability distribution $P$,
        obtained from statistical data,
        only listing entries with positive probabilities,
        together with information on
        variables and appertaining values

Output  A set of probabilistic conditionals

Begin
    % Initialization
    Compute the basic tree of conjunctions
    Compute the list NC of null-conjunctions
    Compute the set $\mathcal{S}_0$ of basic rules
    Compute ker$_0$ $P$
    Compute ker $g$
    Set $\mathcal{K}$ := ker $g$
    Set $\mathcal{S} := \mathcal{S}_0$

    % Main loop
    While equations of type (10) are in $\mathcal{K}$ Do
        Choose $gp \in \mathcal{K}$ of type (10)
        Modify (and compactify) $\mathcal{S}$
        Modify (and reduce) $\mathcal{K}$
    Calculate the probabilities of the conditionals in $\mathcal{S}$
    Return $\mathcal{S}$ and appertaining probabilities
End.
```

Figure 1: The CKD-algorithm

As was already pointed out in the previous section, the set $\mathcal{B}$ of *all* basic single elementary conditionals is intractable and thus may not really serve as a starting point in our algorithm. There is another problem which one usually encounters in data mining problems and which seems to have been neglected hitherto: The frequency distributions calculated from data are mostly not positive – just to the contrary, they would be sparse, full of zeros, with only scattered clusters of non-zero probabilities. This overload of zeros is also a problem with respect to knowledge representation, since a zero in such a frequency distribution often merely means that such a combination has not been *recorded*. The strict probabilistic interpretation of zero probabilities, however, is that such a combination does not *exist* which seems not to be adequate.

Both of these problems – the exponential complexity of the ideal conditional starter set and the sparse and mostly incomplete knowledge provided by statistical data – can be solved in our framework in the following way: The zero values in frequency distributions are taken to be unknown, but equal probabilities, that is, they are treated as non-knowledge without structure. More exactly, let P be the frequency distribution computed from the set of data under consideration. Then, for each two worlds ω_1, ω_2 not occurring in the database and thus being assigned a zero probability, we have $P(\omega_1) = P(\omega_2)$ and hence $\frac{\omega_1}{\omega_2} \in$ *ker*$_0$ P. In this way, all these so-called *null-worlds* contribute to *ker*$_0$ P, and their structure may be theoretically exploited to shrink the starting set of conditionals in advance.

In order to represent missing information in a most concise way, *null-conjunctions* (i.e. elementary conjunctions with frequency 0) have to be calculated as disjunctions of null-worlds. To this end, the *basic tree of conjunctions* is built up. Its nodes are labelled by the names of variables, and the outgoing edges are labelled by the corresponding values, or literals, respectively. The labels of paths going from the root to nodes define elementary conjunctions. So, the leaves of the tree either correspond to complete conjunctions occurring in the database, or to null-conjunctions. These null-conjunctions are collected and aggregated to define a set *NC* of most concise conjunctions of probability 0.

Now we are able to set up a set $\mathcal{S}_0$ of *basic rules* also with the aid of tree-like structures. First, it is important to observe that, due to Proposition 9, conditionals may be separately dealt with according to the literal occurring in their consequents. So $\mathcal{S}_0$ consists of sets $\mathcal{S}_0.v$ of conditionals with consequent v, for each value v of each variable $V \in \mathcal{V}$. Basically, the full trees contain all basic single-elementary conditionals from $\mathcal{B}$, but the trees are pruned with the help of the set *NC* of null-conjunctions. The method to shorten the premises of the rules is the same as has been developed in the previous section with non-zero probabilities, except that now appropriate modifications have to be anticipated, in order to be able to work with a set of rules of acceptable size right from the beginning.

Next, the numerical relationships in P have to be explored to set up *ker*$_0$ P. We only use complete conjunctions with non-zero probabilities for this purpose. So again, we avoid to use missing information. Usually, numerical relationships $P(\widehat{\omega}) = 1$ stemming from learning single-elementary rules can be found between neighboring complete conjunctions (i.e. complete conjunctions that differ in exactly one literal). We construct a *neighbor graph* from P, the vertices of which are the non-null-worlds, labeled by their frequencies or probabilities, and with edges connecting any two neighbors. Then any such relationship $P(\widehat{\omega}) = 1$ corresponds to a cycle of even length (i.e. involving an even number of vertices) in the neighbor graph, such that the cross-product built from the frequencies associated with the vertices, with alternating exponents $+1$ and -1 according to the order of vertices in the cycle, amounts to (a number close to) 1. Therefore, the search for numerical relationships holding in P amounts to searching for such cycles in the neighbor graph. Finally, as the last step of the initialization, *ker* g has to be computed from *ker*$_0$ P with respect to the set $\mathcal{S}_0$ of conditionals.

In the main loop of the algorithm *CKD*, the sets $\mathcal{K}$ of group elements and $\mathcal{S}$ of conditionals are subject to change. In the beginning, $\mathcal{K} = \textit{ker } g$ and $\mathcal{S} = \mathcal{S}_0$; in the end, $\mathcal{S}$ will contain the discovered conditional relationships. More detailed, the products in $\mathcal{K}$ which correspond to equations of type (10) are used to simplify the set $\mathcal{S}$. The modified conditionals induce in turn a modification of $\mathcal{K}$, and this is repeated as long as elements yielding equations of type (10) can be found in $\mathcal{K}$. Note that no probabilities are used in this main loop – only structural information (derived from numerical information) is processed. It is only afterwards, that the probabilities of the conditionals in the final set $\mathcal{S}$ are computed from P, and the probabilistic conditionals are returned.

Although equations of type (10) are the most typical ones, more complicated equations may arise, which need further treatment. The techniques described above, however, are basic to solving *any* group equation. More details will be published in a forthcoming paper. But in many cases, we will find that all or nearly all equations in *ker* g can be solved successfully and hence can be eliminated from $\mathcal{K}$.

We will illustrate our method by the following example. The results shown are found with the help of CONDOR, but the example is simple enough to be calculated "by hand". Nevertheless, it may well serve to show how the algorithm works, in particular, how missing information is dealt with.

Example 11 Suppose in our universe are *animals* (A), *fish* (B), *aquatic beings* (C), *objects with gills* (D) and *objects with scales* (E). The following table may reflect our observations:

object	*freq.*	*prob.*	*object*	*freq.*	*prob.*
$abcde$	59	0.5463	$a\overline{b}cde$	11	0.1019
$abcd\overline{e}$	21	0.1944	$a\overline{b}cd\overline{e}$	9	0.0833
$abc\overline{d}e$	6	0.0556	$abc\overline{d}\overline{e}$	2	0.0185

The set of *null-conjunctions* is calculated as $NC = \{\overline{a}, \overline{c}, \overline{b}\,\overline{d}\}$ – no object matching any one of these partial descriptions occurs in the data base. These null-conjunctions are crucial to set up a starting set $\mathcal{B}$ of basic rules of feasible size:

$$\begin{array}{lllllll} \mathcal{B} = \{ & \phi_{b,1} & = & (b|acde) & \phi_{d,1} & = & (d|abce) \\ & \phi_{b,2} & = & (b|acd\overline{e}) & \phi_{d,2} & = & (d|abc\overline{e}) \\ & \phi_{b,3} & = & (b|\overline{d}) & \phi_{d,3} & = & (d|\overline{b}) \\ & \phi_{e,1} & = & (e|abcd) & \phi_{a,1} & = & (a|\top) \\ & \phi_{e,2} & = & (e|abc\overline{d}) & & & \\ & \phi_{e,3} & = & (e|a\overline{b}cd) & \phi_{c,1} & = & (c|\top)\ \} \end{array}$$

So, the missing information reflected by the set NC null-conjunctions helped to shrink the starting set $\mathcal{B}$ of rules from $5 \cdot 2^4 = 80$ basic single-elementary rules to only 11 conditionals. The next step is to analyze numerical relationships. In this example, we find two numerical relationships between neighboring worlds that are nearly equal:

$$P(a\overline{b}cde) \approx P(a\overline{b}cd\overline{e}) \quad \text{and} \quad P(\frac{abcde}{abcd\overline{e}}) \approx P(\frac{abc\overline{d}e}{abc\overline{d}\overline{e}})$$

The first relationship can be translated into the following structural equations by using $\sigma_{\mathcal{B}}$, according to Theorem 8:

$$\mathbf{b}_{a,1}^{+}\mathbf{b}_{b,1}^{-}\mathbf{b}_{c,1}^{+}\mathbf{b}_{d,3}^{+}\mathbf{b}_{e,3}^{+} \equiv_g \mathbf{b}_{a,1}^{+}\mathbf{b}_{b,2}^{-}\mathbf{b}_{c,1}^{+}\mathbf{b}_{d,3}^{+}\mathbf{b}_{e,3}^{-}$$
$$\Rightarrow\ \mathbf{b}_{b,1}^{-} \equiv_g \mathbf{b}_{b,2}^{-} \text{ and } \mathbf{b}_{e,3}^{+} \equiv_g \mathbf{b}_{e,3}^{-} \equiv_g 1$$

So $\phi_{b,1}$ and $\phi_{b,2}$ are joined to yield $(b|acd)$, and $\phi_{e,3}$ is eliminated. In a similar way, by exploiting the second relationship in P, we obtain $\mathbf{b}_{d,1}^{\pm} \equiv \mathbf{b}_{d,2}^{\pm}$ and $\mathbf{b}_{e,1}^{\pm} \equiv \mathbf{b}_{e,2}^{\pm}$, that is, the corresponding conditionals have to be joined. As a final output, the CKD algorithm returns the following set of conditionals:

cond.	*prob.*	*cond.*	*prob.*
$(a|\top)$	1	$(c|\top)$	1
$(b|\overline{d})$	1	$(d|\overline{b})$	1
$(b|acd)$	0.8	$(d|abc)$	0.91
$(e|abc)$	0.74		

All objects in our universe are aquatic animals which are fish or have gills. Aquatic animals with gills are mostly fish (with a probability of 0.8), aquatic fish usually have gills (with a probability of 0.91) and scales (with a probability of 0.74). ■

Implementation details

In order to be able to test the algorithm, a prototype has been implemented using the functional programming language Haskell. Haskell was chosen as the implementation language because functional programs in general are shorter and thus easier to maintain than their counterparts written in imperative or object-oriented languages. Furthermore, the use of higher-order functions makes it easy to write new functions reusing others, and, in cooperation with Haskell's clear syntax, facilitates to concentrate on the algorithmic details.

For lack of space it is impossible to describe the implementation as a whole. Instead, two crucial parts of the algorithm are presented and discussed in some detail, in order to show certain aspects of the chosen (prototypical) implementation.

As explained in the previous section, one problem when implementing the prototype was the representation of the frequency distribution. The input of the algorithm consists of tabular data, i.e. a table where each column corresponds to one variable $V \in \mathcal{V}$ and every row represents a complete conjunction. The frequency of every complete conjunction can easily be calculated from this table, but the question remains how to represent the frequency distribution in memory. For this purpose, we used a tree-like structure. Given a fixed ordering of the variables $V \in \mathcal{V}$, the internal nodes of this tree are labeled with a variable, where all internal nodes on the same level are labeled with the same variable. Every edge leaving an internal node labeled with variable V_i is labeled with a value $v_i \in [V_i]$. This way, each path from the root node of the tree to one of its leaves defines one complete conjunction, whose frequency is contained in the leaf, and the frequency of a complete conjunction can be computed in time $O(|\mathcal{V}|)$. The frequency of arbitrary conjunctions can also be calculated easily: starting at the root note, the set of literals corresponding to the variable the currently visited node is labeled with is picked out of all literals included in the conjunction. The corresponding subtrees are visited and finally, when reaching the leaves, the particular frequencies are accumulated.

The tree of conjunctions also facilitates the computation of null-conjunctions, which would not be possible using e.g. a tabular representation. Null-conjunctions are represented by leaves with a frequency of 0. These null-conjunctions are collected and aggregated, but one can further accelerate the search for these null-conjunctions (and perhaps also the calculation of the frequencies of conjunctions) by reordering the variables according to certain heuristics. For example, suppose we are given four binary variables A, B, C and D. Suppose further on that the frequency of either complete conjunction $abc\overline{d}$, $\overline{a}bc\overline{d}$, $ab\overline{c}d$ and $\overline{a}b\overline{c}\overline{d}$ is 0. Using the given ordering, one must collect all four null-conjunctions and aggregate them to attain more concise null-conjunctions. But rearranging the variables to B, D, A, C would immediately give the shorter null-conjunction $b\overline{d}$.

Another important part is the computation of *ker*$_0$ P. To do this, one has to construct the neighbor graph of P. This is not that difficult, as the non-null-worlds, which constitute its vertices, can easily be found by traversing the tree of conjunctions. The difficult part is finding cycles of even length, each corresponding to a numerical relationship $P(\widehat{\omega}) = 1$. One possible solution is conducting a slightly modified depth-first search, starting in every vertex of the neighbor graph. During a depth-first search, one keeps track of the vertices already visited. During the modified depth-first search, only the vertices on the path from the start node to the currently visited note are memorised. As soon as a node contained in this set of already visited nodes is visited again, a cycle has been detected. If this node is the start node and the cycle has even length, one has found an element of *ker*$_0$ P. Of course, using this simple algorithm, elements of *ker*$_0$ P are found more than once, at least twice because the graph is undirected. So the set of newly found elements of *ker*$_0$ P must be postprocessed after every depth-first search. The whole function to compute the elements of *ker*$_0$ P consists of 30 lines of code, which is very short for the amount of work done and illustrates how Haskell's syntax supports the user in writing concise and clear programs.

Summary and further work

In this paper, we present and elaborate an approach to knowledge discovery as a process which reverses inductive knowledge representation. Relevant relationships to be discovered from the data are those that are apt to generate the inherent probabilistic information via an inductive representation method as, for instance, the well-known principle of maximum entropy. We briefly describe the theoretical and methodological background, and also make clear how our method can be implemented by sketching an algorithm. In general, the complexity of this algorithm is determined by the number of non-zero entries in the frequency table (and not by the number of possible worlds, which would make the problems intractable).

The CKD-algorithm and the prototype which are reported on in this paper have been developed and implemented during the CONDOR-project[2]. CONDOR is designed as a system for complex knowledge processing. It will be able to deal both with probabilistic and qualitative knowledge, and its components are devised for knowledge discovery, inductive knowledge representation, inferencing, and belief change operations (Beierle & Kern-Isberner 2003). CONDOR is supposed to be applied for modelling and diagnosis in medical and economical domains.

References

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; and Verkamo, A. 1996. Fast discovery of association rules. In

[2] supported by the German Research Society, DFG, under grant BE 1700/5-1

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds., *Advances in knowledge discovery and data mining*. Cambridge, Mass.: MIT Press. 307–328.

Beierle, C., and Kern-Isberner, G. 2003. Modelling conditional knowledge discovery and belief revision by abstract state machines. In Boerger, E.; Gargantini, A.; and Riccobene, E., eds., *Abstract State Machines 2003 – Advances in Theory and Applications, Proceedings 10th International Workshop, ASM2003*, 186–203. Springer, LNCS 2589.

Benferhat, S.; Dubois, D.; and Prade, H. 1997. Nonmonotonic reasoning, conditional objects and possibility theory. *Artificial Intelligence* 92:259–276.

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

Calabrese, P. 1991. Deduction and inference using conditional logic and probability. In Goodman, I.; Gupta, M.; Nguyen, H.; and Rogers, G., eds., *Conditional Logic in Expert Systems*. Elsevier, North Holland. 71–100.

Cheeseman, P., and Oldford, R., eds. 1994. *Selecting models from data*. Number 89 in Lecture Notes in Statistics. New York Berlin Heidelberg: Springer.

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

Cowell, R.; Dawid, A.; Lauritzen, S.; and Spiegelhalter, D. 1999. *Probabilistic networks and expert systems*. New York Berlin Heidelberg: Springer.

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

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

Heckerman, D. 1996. Bayesian networks for knowledge discovery. In Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds., *Advances in knowledge discovery and data mining*. Cambridge, Mass.: MIT Press.

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

Jaynes, E. 1983. *Papers on Probability, Statistics and Statistical Physics*. Dordrecht, Holland: D. Reidel Publishing Company.

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

Kern-Isberner, G. 1999. *A unifying framework for symbolic and numerical approaches to nonmonotonic reasoning and belief revision*. Department of Computer Science, FernUniversität Hagen. Habilitation thesis.

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

Kern-Isberner, G. 2001a. Conditional indifference and conditional preservation. *Journal of Applied Non-Classical Logics* 11(1-2):85–106.

Kern-Isberner, G. 2001b. *Conditionals in nonmonotonic reasoning and belief revision*. Springer, Lecture Notes in Artificial Intelligence LNAI 2087.

Kern-Isberner, G. 2003. A thorough axiomatization of a principle of conditional preservation in belief revision. *Annals of Mathematics and Artificial Intelligence*. (to appear).

Nute, D., and Cross, C. 2002. Conditional logic. In Gabbay, D., and Guenther, F., eds., *Handbook of Philosophical Logic*, volume 4. Kluwer Academic Publishers, second edition. 1–98.

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

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. San Mateo, Ca.: Morgan Kaufmann.

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