Qualitative Knowledge Discovery

Gabriele Kern-Isberner[1], Matthias Thimm[1], and Marc Finthammer[2]

[1] Faculty of Computer Science, Technische Universität Dortmund, Germany
[2] Department of Computer Science, FernUniversität in Hagen, Germany

Abstract. Knowledge discovery and data mining deal with the task of finding useful information and especially rules in unstructured data. Most knowledge discovery approaches associate conditional probabilities to discovered rules in order to specify their strength. In this paper, we propose a qualitative approach to knowledge discovery. We do so by abstracting from actual probabilities to qualitative information and in particular, by developing a method for the computation of an ordinal conditional function from a possibly noisy probability distribution. The link between structural and numerical knowledge is established by a powerful algebraic theory of conditionals. By applying this theory, we develop an algorithm that computes sets of default rules from the qualitative abstraction of the input distribution. In particular, we show how sparse information can be dealt with appropriately in our framework. By making use of the duality between inductive reasoning and knowledge discovery within the algebraic theory of conditionals, we can ensure that the discovered rules can be considered as being most informative in a strict, formal sense.

1 Introduction

Knowledge discovery is the overall process to extract new and useful information from statistical data, with a focus on finding patterns and relationships that reveal generic knowledge, i. e., knowledge that is not specific to a certain situation. Moreover, these relationships should be presented to the user in an intelligible manner. This makes rules appropriate candidates to encode knowledge that is searched for, as they establish (often generic) relationships between isolated facts and are easily comprehensible for human beings. Usually, a conditional probability is associated with each rule by the knowledge discovery process to specify the strength, or the confidence of the rule.

However, while probabilities are a really expressive means to represent knowledge, they are often of only limited use when it comes to commonsense reasoning. First, there is no straightforward way to process probabilistic information. For instance, if the rules “If symptom A then disease D with probability 0.632” and “If symptom B then

The research reported here was partially supported by the Deutsche Forschungsgemeinschaft (grants BE 1700/7-1 and KE 1413/2-1).

disease D with probability 0.715" are shown to the user, what should he believe if the patient he is facing has symptoms A *and* B? Second, while probabilities are appreciated for their (seemingly objective) preciseness, users would not feel comfortable if they had to distinguish sharply between, say, 0.715 and 0.721. Moreover, statistical data are often noisy and may show particularities of the population they were taken from, which does not match the aim of discovering generic, context-independent knowledge. This suggests that precise probabilistic information is neither completely satisfactory nor useful for knowledge discovery.

In this paper, we propose to solve such problems by extracting more coarse-grained rules from data which are only equipped with an order of magnitude of the corresponding probability. Such qualitative rules could be used to reveal plausible relationships to the user, or even as default rules for commonsense reasoning, by applying one of the well-known nonmonotonic inference formalisms (cf. e.g. [1–3]). This perspective of discovering rules from data and feeding them into an inference engine to make inductive reasoning possible will play a decisive part for the methodology to be presented in this paper. More precisely, we will consider knowledge discovery and inductive reasoning as reverse processes (illustrated in Figure 1) – knowledge discovery extracts most relevant partial knowledge that may serve as a basis for further reasoning from frequency distributions representing complete probabilistic information, while inductive model-based reasoning builds up a complete epistemic model from partial knowledge in a knowledge base.

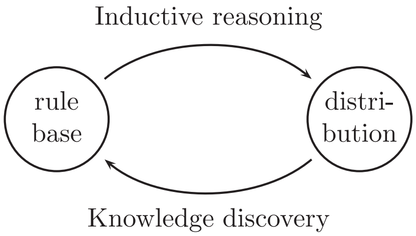

Fig. 1. Knowledge discovery and inductive reasoning as reverse processes

We build upon previous work. In [4, 5], these ideas have been developed and implemented in a fully probabilistic framework. But the core methodology used in these papers is based on structural, algebraic techniques for abstract conditionals and can also be applied in a qualitative framework. However, we first have to transform probabilistic information obtained from data to qualitative rankings. For this, we modify the well-known approach for infinitesimal probabilities [6, 1] to obtain a so-called *ordinal conditional function* [7] which assigns qualitative degrees of disbelief, or rankings, respectively, to propositions and conditionals. The level of qualitative abstraction of probabilities is determined by a parameter ε that specifies a measure of similarity between probabilistic values, according to the needs of the user.

Our approach offers a couple of nice advantages. First, the same methodology is used both for learning and reasoning, handling structural knowledge in a profound algebraic way. Second, the notion of relevance which is crucial for knowledge discovery can be given a precise meaning – rules are relevant wrt. a given set of (already discovered) rules, if they provide additional information for the inductively built model. Third, the qualitative information derived from data reflects an intuitive similarity of the probabilities, different from the approach in [8] in which sums of probabilities have to be used.

The outline of the paper is as follows. In the next section, we will recall basic facts on probabilistic reasoning and ordinal conditional functions. In section 3, we present our approach to extract qualitative information from statistical data. We also propose a heuristic how to find a proper abstraction parameter ε. Section 4 describes the core methodology which can be used for inductive representation and knowledge discovery and that is applied in section 5 for the knowledge discovery task. Based on this theoretical work, an algorithm for discovering default rules in statistical data is represented in section 6. Section 7 concludes the paper with a summary and an outlook on further work.

2 Inductive reasoning with probabilities and rankings

We consider a propositional framework over a finite set $\mathcal{V} = \{V_1, V_2, \ldots\}$ of (multi-valued) 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 overlining, 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 instantiated 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 uncertain rules of the form *if A then B*. Let $(\mathcal{L}|\mathcal{L})$ 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. To provide semantics for conditionals, a richer epistemic framework is needed than a plain bivalued semantics. Basically, for a conditional $(B|A)$ to be accepted, its confirmation, AB, must be more probable, plausible etc. than its refutation, $A\overline{B}$. Moreover, numerical degrees of probability, plausibility and the like can be assigned to conditionals to specify the strength with which they are believed, according to the chosen epistemic framework. In this paper, we will use probabilities to model a fully quantitative frame, and so-called *ordinal conditional functions, OCFs,* (or simply *ranking functions*) to model a qualitative, respectively semi-quantitative frame. We will briefly summarize basic facts on both modelling frames in the following. We will also address the problem which is crucial to this paper: Given partial information in form of a conditional knowledge base, how to obtain an adequate complete model that can be used for inductive reasoning?

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 $x \in [0,1]$. A *conditional probabilistic knowledge base* is a set $\mathcal{R}^{prob} = \{(B_1|A_1)[x_1], \ldots, (B_n|A_n)[x_n]\}$ of probabilistic conditionals.

Suppose such a conditional probabilistic knowledge base $\mathcal{R}^{prob}$ is given. For instance, $\mathcal{R}^{prob}$ may describe the knowledge available to a physician when he has to make a diagnosis. Or, $\mathcal{R}^{prob}$ 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 $\mathcal{R}^{prob} = \{(\text{young}\,|\text{student})[0.8], (\text{young}\,|\text{single})[0.7]\}$. 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 [9] 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 $P^* = \mathsf{ME}(\mathcal{R}^{prob})$ that maximizes $H(P)$ subject to $P \models \mathcal{R}^{prob}$. In this way, the ME-method ensures that no further information is added, so that the knowledge $\mathcal{R}^{prob}$ is represented most faithfully. $\mathsf{ME}(\mathcal{R}^{prob})$ can be written in the form

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

with the α_i's being chosen appropriately so as to satisfy all of the conditional constraints in $\mathcal{R}^{prob}$ (cf. [10]); $\mathsf{ME}(\mathcal{R}^{prob})$ is called the ME-*representation of* $\mathcal{R}^{prob}$. The ME-principle provides a most convenient and theoretically sound method to represent incomplete probabilistic knowledge[3] and for high-quality probabilistic reasoning (cf. [11]).

[3] Efficient implementations of ME-systems can be found via `www.informatik.fernuni-hagen.de/pi8/research/projects.html` and `www.pit-systems.de`

A purely probabilistic representation gives precise numerical values to all propositions and conditionals of the underlying language. This can be problematic with respect to two points: First, when the aim is to model subjective beliefs of an expert or an agent, precise probabilities are hard to specify. Subjective probabilities are more or less rough guidelines that are based on an agent's experience. Second, even objective probabilities derived from statistical data may not represent a completely accurate picture of the world. Statistical data can be noisy and only reflect a snapshot of the world, which can be quite accidental. Therefore, in this paper, we are interested in the qualitative knowledge that underlies some given probabilistic information. To represent such qualitative structures, we use *ordinal conditional functions, OCFs,* as introduced by Spohn [7] as a qualitative abstraction of probability functions.

Definition 1. *An* ordinal conditional function *(or ranking function) κ is a function $\kappa : \Omega \to \mathbb{N} \cup \{\infty\}$ with $\kappa^{-1}(0) \neq \emptyset$.*

An OCF κ assigns a *degree of implausibility* (or *ranking value*) to each world ω: The higher $\kappa(\omega)$, the less plausible is ω. A world ω with $\kappa(\omega) = 0$ is regarded as being completely normal (most plausible), and for a consistent modelling, there has to be at least one such world. For formulas $A \in \mathcal{L}$, a ranking is computed via

$$\kappa(A) = \begin{cases} \min\{\kappa(\omega) \mid \omega \models A\} & \text{if } A \text{ is satisfiable} \\ \infty & \text{otherwise} \end{cases} .$$

So we have $\kappa(A \vee B) = \min\{\kappa(A), \kappa(B)\}$ and in particular, $\kappa(A \vee \overline{A}) = 0$. The *belief* in (or *acceptance* of) a formula A is defined as

$$\kappa \models A \quad \text{iff} \quad \kappa(\overline{A}) > 0 \quad ,$$

i. e., $\kappa(A) = 0$ is necessary but not sufficient to believe A, because $\kappa(\overline{A})$ might be 0 as well; but $\kappa(\overline{A}) > 0$ is sufficient, since it implies $\kappa(A) = 0$.

Similar to the probabilistic framework, conditionals can be quantified. An OCF κ is extended to conditionals by setting

$$\kappa(B|A) = \begin{cases} \kappa(AB) - \kappa(A) & \text{if } \kappa(A) \neq \infty \\ \infty & \text{otherwise} \end{cases} ,$$

and a conditional is *accepted* by κ,

$$\kappa \models (B|A) \quad \text{iff} \quad \kappa(AB) < \kappa(A\overline{B}) \quad \text{iff} \quad \kappa(\overline{B}|A) > 0.$$

As usual, a proposition A is identified with the conditional $(A|\top)$, hence $\kappa \models (A|\top)$ iff $\kappa(\overline{A}) > \kappa(A) = 0$, in accordance with what was said above.

The acceptance relation for quantified *OCF-conditionals* $(B|A)[m]$ is defined by using the difference between $\kappa(AB)$ and $\kappa(A\overline{B})$:

$$\kappa \models (B|A)[m] \text{ iff } \kappa(AB) + m = \kappa(A\overline{B}) \text{ iff } \kappa(\overline{B}|A) = m, \ m \in \mathbb{N}, m \geq 1. \qquad (2)$$

Thus, if $(B|A)$ is believed with a *degree of belief* m then verifying the conditional is m degrees more plausible than falsifying it. So, $\kappa \models (B|A)[1]$ expresses belief in $(B|A)$, but only to the smallest possible degree. For a propositional fact A, this yields

$$\kappa \models A[m] \quad \text{iff} \quad \kappa(\overline{A}) = m.$$

Both qualitative and quantitative OCF-conditionals can be used as default rules for commonsense reasoning [1, 11].

Ranking functions provide a perfect framework for qualitative reasoning, as they allow us to handle conditionals in a purely qualitative manner, but also leave room to take more precise, quantitative information into account. However, even numerical information merely expresses an order of magnitude of probabilities; this will be made more precise in the following section.

Moreover, in a qualitative framework with ordinal conditional functions, a similar concept as an ME-representation can be defined in order to express a certain "well-behavedness" of an OCF with respect to a set of OCF-conditionals [11]. We will come back to these issues and present said concept in section 4. In the next section we will first have a look on how to derive an OCF from an empirically obtained probability distribution.

3 Deriving qualitative information from statistical data

Let P be a probability distribution over $\mathcal{V}$ that could have been collected via a statistical survey. In this paper we are interested in the qualitative structure that underlies the probabilities in P. So we represent P by qualitative probabilities yielding an ordinal conditional function that approximates the quantitative structure in P.

For this reason we start by representing a probability of a specific world ω as polynomial in a fixed base value ε in the spirit of [1]. Using this base representation, the order of magnitude of a probability can be represented only by the corresponding exponents and different probabilities can be compared by these exponents yielding a qualitative abstraction of the original values.

Definition 2. *Let $\varepsilon \in (0,1)$ be a base value to parameterize probabilities. Then a probability value $P(\omega)$ can be expressed as a* polynomial in ε *,*

$$P_\varepsilon(\omega) = a_0\varepsilon^0 + a_1\varepsilon^1 + a_2\varepsilon^2 + \ldots \quad ,$$

with appropriate coefficients $a_i \in \mathbb{N}$ respecting $0 \le a_i < \varepsilon^{-1}$ for all i to match the value $P(\omega)$.

Due to the restriction $0 \le a_i < \varepsilon^{-1}$ the above definition is sound and uniquely determines a base representation $P_\varepsilon(\omega)$ for given $P(\omega)$ and ε with $P_\varepsilon(\omega) = P(\omega)$.

Example 1. Let $\varepsilon = 0.3$. Then the probability $P(\omega_1) = 0.171$ is written as a polynomial $P_\varepsilon(\omega_1) = 0 \cdot 0.3^0 + 0 \cdot 0.3^1 + 1 \cdot 0.3^{\mathbf{2}} + 3 \cdot 0.3^3$ in ε.

Observe that in the above approach the value of a_0 is always zero, except for the case that the world ω has a probability of 1, which is unlikely the case in real world scenarios. Furthermore the above definition differs from the definition of polynomial base representations in [1] in the sense, that Goldszmidt and Pearl implicitly use negative coefficients for their base representation, representing probabilities as polynomials of the form $P'_\varepsilon(\omega) = 1 - a\varepsilon$ or $P'_\varepsilon(\omega) = a\varepsilon^2 - b\varepsilon^4$. However, an additive representation of positive values like probabilities seems more appropriate for our intentions.

Nonetheless, Goldsmizdt and Pearl restrict their attention on qualitative abstractions of probabilities to the case of infinitesimal bases yielding the following definition of a complete translation of all probability values into rankings.

Definition 3 (see [1]). *Let P be a probability distribution and let the probability $P(\omega)$ be written as a polynomial $P'_\varepsilon(\omega)$ in ε with an infinitesimal ε. A ranking function $\kappa_0^P(\omega)$ is defined as follows*

$$\kappa_0^P(\omega) = \begin{cases} \min\{n \in \mathbb{N} \mid \lim_{\varepsilon \to 0} \frac{P'_\varepsilon(\omega)}{\varepsilon^n} \neq 0\} & \textit{if } P'_\varepsilon(\omega) > 0 \\ \infty & \textit{if } P'_\varepsilon(\omega) = 0 \end{cases}$$

The general idea of the above definition is to capture the most significant term of the base representation of a probability of a world ω, i. e., the first coefficient a_i that differs from zero, and use this value as the rank of ω

$$\kappa_0^P(\omega) = \min\{i \mid a_i \neq 0\}, \quad P_\varepsilon(\omega) = a_0\varepsilon^0 + a_1\varepsilon^1 + \ldots \tag{3}$$

In this paper, we use this idea for a fixed value ε for the base representation and take this value throughout the process of qualitative knowledge discovery as an indicator for the granularity of the qualitative probabilities. Given a fixed base value ε, we determine the most significant term of a base representation with respect to ε and use this value as a rank value for an OCF $\tilde{\kappa}_\varepsilon^P$ as in equation (3). More specifically, let ω be a world and $P(\omega)$ its (empirical) probability. From now on let $\varepsilon \in (0, 1)$ be a fixed base value and let

$$P_\varepsilon(\omega) = a_0\varepsilon^0 + a_1\varepsilon^1 + a_2\varepsilon^2 + \ldots$$

be the base representation of $P(\omega)$ according to Definition 2. We are looking for the first a_i that differs from zero to define the rank of ω:

$$\tilde{\kappa}_\varepsilon^P(\omega) = \min\{i \mid a_i \neq 0\} \quad .$$

Let i satisfy $a_i \neq 0$. Then it holds that

$$P(\omega) \geq a_i\varepsilon^i \geq \varepsilon^i$$

because a_i is a natural number and $a_i > 0$. From this observation, it follows immediately

$$\begin{aligned} & P(\omega) \geq \varepsilon^i \\ \Leftrightarrow\ & \log P(\omega) \geq i \log \varepsilon \\ \Leftrightarrow\ & \frac{\log P(\omega)}{\log \varepsilon} \leq i \quad . \end{aligned}$$

Therefore for the minimal i satisfying $a_i \neq 0$ and so for the rank assigned to ω it follows

$$\tilde{\kappa}_\varepsilon^P(\omega) = \left\lceil \frac{\log P(\omega)}{\log \varepsilon} \right\rceil \tag{4}$$

In general, the function $\tilde{\kappa}_\varepsilon^P$ defined using equation (4) does not satisfy $(\tilde{\kappa}_\varepsilon^P)^{-1}(0) \neq \emptyset$. Therefore, we normalize $\tilde{\kappa}_\varepsilon^P$ by shifting all ranking values appropriately, i. e., by defining $\kappa_\varepsilon^P(\omega) := \tilde{\kappa}_\varepsilon^P(\omega) - c$ with $c = \min\{\tilde{\kappa}_\varepsilon^P(\omega) \mid \omega \in \Omega\}$. Then κ_ε^P defines an ordinal conditional function according to Definition 1. As κ_ε^P is the only ordinal conditional function we are dealing with, we will write just κ for κ_ε^P, when P and ε are clear from context.

Example 2. (Continuing Example 1)
With ε being 0.3, the probability $P(\omega_1) = 0.171$ is written as a polynomial $P_\varepsilon(\omega_1) = 0 \cdot 0.3^0 + 0 \cdot 0.3^1 + 1 \cdot 0.3^{\mathbf{2}} + 3 \cdot 0.3^3$ in ε and therefore $\kappa(\omega_1) = 2$. The probabilities $P(\omega_2) = 0.39$ and $P(\omega_3) = 0.48$ are written as $P_\varepsilon(\omega_2) = 0 \cdot 0.3^0 + 1 \cdot 0.3^{\mathbf{1}} + 1 \cdot 0.3^2$ and $P_\varepsilon(\omega_3) = 0 \cdot 0.3^0 + 1 \cdot 0.3^{\mathbf{1}} + 2 \cdot 0.3^2$, respectively, and so they are both projected to the same ranking value $\kappa(\omega_2) = \kappa(\omega_3) = 1$.

A process of transforming a given probability distribution into a qualitative representation (according to equation (4)) is crucially influenced by the chosen base value ε. It depends on ε how similar some probabilities must be to be projected to the same ranking value. Thus, ε is the parameter that controls the qualitative smoothing of the probabilities. For this reason, an appropriate choice for ε is important for the qualitative modeling since it determines the variation in the resulting ranking values and this way it heavily influences all following calculations based on this values. If the value for ε is close to 1, then even quite similar probabilities will still be projected to different ranking values.

However, a too small value of ε will have the effect that even quite different probabilities will be assigned an identical ranking value. Thus, an unacceptable large amount of information contained in the probabilities will be lost, i. e., the probabilities are smoothed so much that the resulting ranking values do not carry enough information to be useful as a qualitative abstraction.

The following example will illustrate to what degree the choice of ε influences the resulting ranking values.

Example 3. Suppose in our universe are *animals* (A), *fish* (B), *aquatic beings* (C), *objects with gills* (D) and *objects with scales* (E). Table 1 may reflect our observations. Table 2 shows the ranking values that result from different choices of ε. Choosing

ω	*object*	*frequency*	*probability*
ω_1	$abcde$	59	0.5463
ω_2	$abcd\overline{e}$	21	0.1944
ω_3	$a\overline{b}cde$	11	0.1019
ω_4	$a\overline{b}cd\overline{e}$	9	0.0833
ω_5	$abc\overline{d}e$	6	0.0556
ω_6	$abc\overline{d}\overline{e}$	2	0.0185

Table 1. Empirical probabilities for Example 3

	ranking value		
ω	ε *=0.1*	ε *=0.6*	ε *=0.9*
ω_1	1	2	6
ω_2	1	4	16
ω_3	1	5	22
ω_4	2	5	24
ω_5	2	6	28
ω_6	2	8	38

Table 2. Ranking values resulting from different choices of ε

$\varepsilon = 0.1$ assigns identical ranking values to ω_1, ω_2 and ω_3 and to ω_4, ω_5 and ω_6, respectively. Mapping the latter ones to the same rank could be acceptable, but mapping the former ones to a common rank is inappropriate, since the probabilities of these worlds cover a (comparative) large range between 0.5463 and 0.1019. Hence, this choice for ε smoothes the probabilities too much, leading to a qualitative abstraction that is so coarse that almost all information of the observed distribution is lost. Choosing $\varepsilon = 0.9$ leads to different ranking value for all ω, although some of the probabilities are quite similar and therefore should not be distinguished in a qualitative setting. Hence, this choice for ε does not seem very appropriate as well because it does not smooth the probabilities effectively. Choosing $\varepsilon = 0.6$ results in a common ranking value for the (comparative) similar probabilities of ω_3 and ω_4. This choice for ε seems to be appropriate to obtain ranking values that form a qualitative representation of the observed probabilities.

In this very small example, the worlds ω offer quite high probabilities. For this reason, the appropriate value for ε is quite big, too. In a more realistic setting with considerably smaller probabilities, a much smaller value for ε would be chosen.

The parameter ε defines a measure of similarity that is to make probabilities indistinguishable. In principle, it is up to the user to set ε, depending on his point of view, but clustering techniques applied to the logarithmic probabilities may help to find an appropriate ε. A useful heuristic may be to fix a logarithmic similarity α, i.e. probabilities should not be distinguished if their logarithmic distance does not exceed α. Then clusters of logarithmic probabilities with maximal width α are built. A dendrogram computed by, e. g., a complete link clustering procedure (cf. [12]) may provide helpful information for this. Moreover, α should be chosen in such a way that no multiple $k\alpha$ of α falls within one of the clusters. Finally, $\varepsilon = e^{-\alpha}$ may serve to extract ranking infomation from the empirical probabilities. We will illustrate this in our Example 3.

Example 4. Table 3 shows the logarithmic probabilities $\log_e P(\omega)$ of our example. If we use a logarithmic similarity $\alpha = 0.5$, then only $P(\omega_3)$ and $P(\omega_4)$ are close enough to be identified, and all multiples of 0.5 discriminate the clusters clearly. Hence $\varepsilon = e^{-0.5} \approx 0.6$ yields an adequate ranking function.

ω	*object*	*frequency*	*probability*	*log. probability*
ω_1	$abcde$	59	0.5463	-0.60
ω_2	$abcd\overline{e}$	21	0.1944	-1.64
ω_3	$a\overline{b}cde$	11	0.1019	-2.28
ω_4	$a\overline{b}cd\overline{e}$	9	0.0833	-2.49
ω_5	$abc\overline{d}e$	6	0.0556	-2.89
ω_6	$abc\overline{d}\overline{e}$	2	0.0185	-3.99

Table 3. Logarithmic probabilities

In the next section, we develop an algebraic theory of conditionals, that is used to obtain structural information from such ordinal conditional functions like the one derived above.

4 Conditional structures and c-representations

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 [13–16]. Some of these algorithms also make use of optimization criteria which are based on entropy [17, 18]. 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 ($P(\text{sick}\,|\text{flu}) = 1$), and they often will complain about headache ($P(\text{headcache}\,|\text{flu}) = 0.9$). Then we have

$$P(\text{headcache}\,|\text{flu}) = P(\text{headcache}\,|\text{flu} \wedge \text{sick}),$$

but we would surely expect

$$P(\text{headcache}\,|\neg\text{flu}) \neq P(\text{headcache}\,|\neg\text{flu} \wedge \text{sick})!$$

Although, from a naïve point of view, the (first) equality suggests a conditional independence between sick and headcache, due to the causal dependency between headcache 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?

A well-known approach to answer this question is *system* Z [1] that builds up a completely specified ranking function from a set of conditionals $\{(B_1|A_1), \ldots, (B_n|A_n)\}$ and yields an inductive reasoning method that satisfies basic properties of default reasoning. In this paper, however, we use *c-representations* for qualitative inductive reasoning that have been developed in [4, 11]; all proofs and lots of examples can be found in [11]. This approach follows the same structural lines as ME-reasoning and provides the techniques for model-based inductive reasoning in a qualitative environment the quality of which outperforms *system Z* clearly [19, 20].

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 unquantified conditionals.

In order to model its non-classical uncertainty, we represent a conditional $(B|A)$ as a three-valued indicator function on worlds

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

where u stands for *unknown*, following an idea of de Finetti (cf., e. g., [21, 22]). 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 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:

$$\sigma_i(\omega) = \begin{cases} \mathbf{a}_i^+ & \text{if} \quad \omega \models A_i B_i \\ \mathbf{a}_i^- & \text{if} \quad \omega \models A_i \overline{B_i} \\ 1 & \text{if} \quad \omega \models \overline{A_i} \end{cases} \tag{5}$$

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

$$\sigma_\mathcal{R}(\omega) = \prod_{1 \le i \le n} \sigma_i(\omega) = \prod_{\substack{1 \le i \le n \\ \omega \models A_i B_i}} \mathbf{a}_i^+ \prod_{\substack{1 \le i \le n \\ \omega \models A_i \overline{B}_i}} \mathbf{a}_i^- \tag{6}$$

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 5. Let $\mathcal{R} = \{(c|a), (c|\overline{b})\}$, where A, B, C are bivalued propositional variables with outcomes $\{a, \overline{a}\}, \{b, \overline{b}\}$ and $\{c, \overline{c}\}$, respectively, and let $\mathfrak{F}_\mathcal{R} = \langle \mathbf{a}_1^+, \mathbf{a}_1^-, \mathbf{a}_2^+, \mathbf{a}_2^- \rangle$. 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 $\sigma_\mathcal{R}(\overline{a}b\overline{c}) = \mathbf{a}_2^-$.

Let $\hat{\mathit{\Omega}} := \langle \hat{\omega} \mid \omega \in \mathit{\Omega} \rangle$ be the free abelian group generated by all $\omega \in \mathit{\Omega}$, and consisting of all products $\hat{\omega} = {\omega_1}^{r_1} \ldots {\omega_m}^{r_m}$ with $\omega_1, \ldots, \omega_m \in \mathit{\Omega}$ 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* $\hat{\omega} \in \hat{\mathit{\Omega}}$ in which only positive exponents occur simply corresponds to a multi-set of worlds. We will often use fractional representations for the elements of $\hat{\mathit{\Omega}}$, 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 $\hat{\mathit{\Omega}}$ 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* $\sigma_{\mathcal{R}} : \hat{\Omega} \rightarrow \mathfrak{F}_{\mathcal{R}}$.

Having the same conditional structure defines an equivalence relation $\equiv_{\mathcal{R}}$ on $\hat{\Omega}$: $\hat{\omega}_1 \equiv_{\mathcal{R}} \hat{\omega}_2$ iff $\sigma_{\mathcal{R}}(\hat{\omega}_1) = \sigma_{\mathcal{R}}(\hat{\omega}_2)$, i. e. iff $\hat{\omega}_1 \hat{\omega}_2^{-1} \in \mathit{ker}\ \sigma_{\mathcal{R}} := \{\hat{\omega} \in \hat{\Omega} \mid \sigma_{\mathcal{R}}(\hat{\omega}) = 1\}$. Thus the kernel of $\sigma_{\mathcal{R}}$ plays an important part in identifying the conditional structure of elements $\hat{\omega} \in \hat{\Omega}$. $\mathit{ker}\ \sigma_{\mathcal{R}}$ contains exactly all group elements $\hat{\omega} \in \hat{\Omega}$ with a balanced conditional structure, that means, where all effects of conditionals in $\mathcal{R}$ on worlds occurring in $\hat{\omega}$ are completely cancelled. Since $\mathfrak{F}_{\mathcal{R}}$ 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 $\mathfrak{F}_{\mathcal{R}}$, so we have $\sigma_{\mathcal{R}}(\hat{\omega}) = 1$ iff $\sigma_i(\hat{\omega}) = 1$ for all i, $1 \leq i \leq n$, and this means

$$\mathit{ker}\ \sigma_{\mathcal{R}} = \bigcap_{i=1}^{n} \mathit{ker}\ \sigma_i \quad .$$

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

Besides the explicit representation of knowledge by $\mathcal{R}$, also the implicit normalizing constraint $\kappa(\top|\top) = 0$ for ordinal conditional functions has to be taken into account. It is easy to check that $\mathit{ker}\ \sigma_{(\top|\top)} = \hat{\Omega}_0$, with

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

Two elements $\hat{\omega}_1 = \omega_1^{r_1} \ldots \omega_m^{r_m}$, $\hat{\omega}_2 = \nu_1^{s_1} \ldots \nu_p^{s_p} \in \hat{\Omega}$ are equivalent modulo $\hat{\Omega}_0$, $\hat{\omega}_1 \equiv_{\top} \hat{\omega}_2$, iff $\hat{\omega}_1 \hat{\Omega}_0 = \hat{\omega}_2 \hat{\Omega}_0$, i.e. iff $\sum_{1 \leq j \leq m} r_j = \sum_{1 \leq k \leq p} s_k$. This means that $\hat{\omega}_1$ and $\hat{\omega}_2$ are equivalent modulo $\hat{\Omega}_0$ iff they both are a (cancelled) product of the same number of generators, each generator being counted with its corresponding exponent. Set

$$\mathit{ker}_0\ \sigma_{\mathcal{R}} := \mathit{ker}\ \sigma_{\mathcal{R}} \cap \hat{\Omega}_0 = \mathit{ker}\ \sigma_{\mathcal{R} \cup \{(\top|\top)\}} \quad .$$

In the following, if not stated otherwise, we will assume that all ordinal conditional functions are finite, i. e., it is $\kappa(A) \neq \infty$ for every A. 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. [11]). Moreover, in section 5 we will see that we can get rid of all infinite ranking values (which correspond to zero probabilities in the empirical distribution) right from the beginning.

Finite ranking functions κ may be extended easily to homomorphisms $\kappa : \hat{\Omega} \rightarrow (\mathbb{Z}, +)$ from $\hat{\Omega}$ into the additive group of integers in a straightforward way by setting

$$\kappa({\omega_1}^{r_1} \ldots {\omega_m}^{r_m}) = r_1 \kappa(\omega_1) + \ldots + r_m \kappa(\omega_m) \quad .$$

Definition 4 (Conditional indifference). *Suppose κ is a (finite) ordinal conditional function, and let $\mathcal{R} = \{(B_1|A_1), \ldots, (B_n|A_n)\}$ be a set of conditionals. κ is* (conditionally) indifferent *with respect to $\mathcal{R}$ iff $\kappa(\hat{\omega}_1) = \kappa(\hat{\omega}_2)$, whenever both $\hat{\omega}_1 \equiv_{\mathcal{R}} \hat{\omega}_2$ and $\hat{\omega}_1 \equiv_{\top} \hat{\omega}_2$ hold for $\hat{\omega}_1, \hat{\omega}_2 \in \hat{\Omega}$.*

If κ is indifferent with respect to $\mathcal{R}$, then it does not distinguish between elements $\hat{\omega}_1 \equiv_{\top} \hat{\omega}_2$ with the same conditional structure with respect to $\mathcal{R}$. Conversely, any deviation $\kappa(\hat{\omega}) \neq 0$ can be explained by the conditionals in $\mathcal{R}$ acting on $\hat{\omega}$ in a non-balanced way. Note that the notion of indifference only aims at observing conditional structures, without making use of any degrees of belief that are associated with the conditionals.

The following proposition shows, that conditional indifference establishes a connection between the kernels $ker_0\ \sigma_{\mathcal{R}}$ and

$$ker_0\ \kappa := \{\hat{\omega} \in \hat{\Omega}_0 \mid \kappa(\hat{\omega}) = 0\}$$

which will be crucial to elaborate conditional structures:

Proposition 1. *An ordinal conditional function κ is indifferent with respect to a set $\mathcal{R} \subseteq (\mathcal{L}|\mathcal{L})$ of conditionals iff $ker_0\ \sigma_{\mathcal{R}} \subseteq ker_0\ \kappa$.*

If $ker_0\ \sigma_{\mathcal{R}} = ker_0\ \kappa$, then $\kappa(\hat{\omega}_1) = \kappa(\hat{\omega}_2)$ iff $\sigma_{\mathcal{R}}(\hat{\omega}_1) = \sigma_{\mathcal{R}}(\hat{\omega}_2)$, for $\hat{\omega}_1 \equiv_{\top} \hat{\omega}_2$. In this case, κ completely follows the conditional structures imposed by $\mathcal{R}$ – it observes $\mathcal{R}$ *faithfully*.

The next theorem characterizes indifferent ordinal conditional functions:

Theorem 1. *An ordinal conditional function κ is indifferent with respect to a set $\mathcal{R} = \{(B_1|A_1), \ldots, (B_n|A_n)\} \subseteq (\mathcal{L}|\mathcal{L})$ iff $\kappa(A_i) \neq \infty$ for all i, $1 \leq i \leq n$ and there are rational numbers $\kappa_0, \kappa_1^+, \kappa_1^-, \ldots, \kappa_n^+, \kappa_n^- \in \mathbb{Q}$, such that*

$$\kappa(\omega) = \kappa_0 + \sum_{\substack{1 \leq i \leq n \\ \omega \models A_i B_i}} \kappa_i^+ + \sum_{\substack{1 \leq i \leq n \\ \omega \models A_i \overline{B_i}}} \kappa_i^-, \tag{7}$$

for all $\omega \in \Omega$.

There are striking similarities between (1), (6), and (7). The equations (1) and (7) are both implementations of (6): while in (1) multiplication is used for combining the operands, in (7) it is addition. Furthermore, in (1), the abstract symbols $\mathbf{a}_i^+, \mathbf{a}_i^-$ of (6) 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). In (7), additive constants κ_i^+, κ_i^- realize the structural effects of conditionals. Both the α_i's and the κ_i's bear crucial conditional information, leaving "footprints" on probabilities resp. ranking values when inductively representing conditionals (also cf. [10]). In [11] it is shown that ordinal conditional functions and probability distributions can be subsumed by the general concept of *conditional valuation functions*.

Example 6. We continue Example 5. 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, $\dfrac{abc \cdot \overline{a}\overline{b}\overline{c}}{a\overline{b}c \cdot \overline{a}bc} \in ker_0\ \sigma_{\mathcal{R}}$. Then any ordinal conditional function κ that is indifferent with respect $\mathcal{R}$ will fulfill $\kappa\left(\dfrac{abc \cdot \overline{a}\overline{b}\overline{c}}{a\overline{b}c \cdot \overline{a}bc}\right) = 0$, i. e., $\kappa(abc)+\kappa(\overline{a}\overline{b}\overline{c}) = \kappa(a\overline{b}c)+\kappa(\overline{a}bc)$.

In [23], we investigate the exact relationship between *conditional indifference* and *conditional independence* and show that conditional indifference is the strictly more general concept.

Now, in order to obtain a proper representation of a set of conditionals $\mathcal{R}$, we can use the schema (7) and impose the constraints induced by the conditionals in $\mathcal{R}$.

Definition 5 (C-representation 1). *An ordinal conditional function κ is a* c-representation *of a set $\mathcal{R} = \{(B_1|A_1), \ldots, (B_n|A_n)\}$ of conditionals iff κ is indifferent with respect to $\mathcal{R}$ and accepts all conditionals in $\mathcal{R}$, i. e. $\kappa \models \mathcal{R}$.*

For the constraints $\kappa \models (B_i|A_i)$, $1 \leq i \leq n$, to hold, the additive constants κ_i^+, κ_i^- have to satisfy certain relationships which can be checked easily.

Proposition 2. *An ordinal conditional function κ is a* c-representation *of a set $\mathcal{R} = \{(B_1|A_1), \ldots, (B_n|A_n)\}$ of conditionals, iff κ has the form (7) and the κ_i^+, κ_i^-, $1 \leq i \leq n$, fulfill the following inequality:*

$$\begin{aligned} \kappa_i^- - \kappa_i^+ > \min_{\omega \models A_i B_i} \Big(&\sum_{\substack{j \neq i \\ \omega \models A_j B_j}} \kappa_j^+ + \sum_{\substack{j \neq i \\ \omega \models A_j \overline{B}_j}} \kappa_j^- \Big) \\ - \min_{\omega \models A_i \overline{B}_i} \Big(&\sum_{\substack{j \neq i \\ \omega \models A_j B_j}} \kappa_j^+ + \sum_{\substack{j \neq i \\ \omega \models A_j \overline{B}_j}} \kappa_j^- \Big) \end{aligned} \tag{8}$$

This approach can be generalized in a straightforward manner to handle quantified OCF-conditionals. If $\mathcal{R}^{\mathrm{OCF}} = \{(B_1|A_1)[m_1], \ldots, (B_n|A_n)[m_n]\}$ is a set of quantified OCF-conditionals, then we denote by $\mathcal{R} = \{(B_1|A_1), \ldots, (B_n|A_n)\}$ its corresponding set of purely qualitative conditinals.

Definition 6 (C-representation 2). *An ordinal conditional function κ is a* c-representation *of a set $\mathcal{R}^{\mathrm{OCF}} = \{(B_1|A_1)[m_1], \ldots, (B_n|A_n)[m_n]\}$ of quantified OCF-conditionals iff κ is indifferent with respect to $\mathcal{R}$ and accepts all conditionals in $\mathcal{R}^{\mathrm{OCF}}$, i. e. $\kappa \models \mathcal{R}^{\mathrm{OCF}}$.*

According to (2), the constraints imposed by $\kappa \models (B_i|A_i)[m_i]$ can be handled in a way similar to the purely qualitative case.

Proposition 3. *An ordinal conditional function κ is a* c-representation *of a set $\mathcal{R}^{\mathrm{OCF}} = \{(B_1|A_1)[m_1], \ldots, (B_n|A_n)[m_n]\}$ of quantified OCF-conditionals, iff κ has the form (7) and the κ_i^+, κ_i^-, $1 \leq i \leq n$, fulfill the following inequality:*

$$\begin{aligned} \kappa_i^- - \kappa_i^+ = m_i + \min_{\omega \models A_i B_i} \Big(&\sum_{\substack{j \neq i \\ \omega \models A_j B_j}} \kappa_j^+ + \sum_{\substack{j \neq i \\ \omega \models A_j \overline{B}_j}} \kappa_j^- \Big) \\ - \min_{\omega \models A_i \overline{B}_i} \Big(&\sum_{\substack{j \neq i \\ \omega \models A_j B_j}} \kappa_j^+ + \sum_{\substack{j \neq i \\ \omega \models A_j \overline{B}_j}} \kappa_j^- \Big) \end{aligned} \tag{9}$$

For the sake of informational economy, the difference $\kappa_i^- - \kappa_i^+$ reflecting the amount of distortion imposed by a conditional belief should be minimal. A reasonable approach to obtain “small” c-representations κ is to set $\kappa_i^+ = 0$ and to choose κ_i^- minimal, in accordance with (8) resp. (9). This simplifies the reasoning with c-representations a lot. The schema (7) shrinks to

$$\kappa(\omega) = \sum_{\substack{1 \leq i \leq n \\ \omega \models A_i\overline{B_i}}} \kappa_i^-, \; \omega \in \Omega, \tag{10}$$

as for consistent sets of conditionals the normalizing constant κ_0 is always zero, and the inequalities (8) now read

$$\kappa_i^- > \min_{\omega \models A_i B_i} \Big(\sum_{\substack{j \neq i \\ \omega \models A_j \overline{B}_j}} \kappa_j^- \Big) - \min_{\omega \models A_i \overline{B}_i} \Big(\sum_{\substack{j \neq i \\ \omega \models A_j \overline{B}_j}} \kappa_j^- \Big) \tag{11}$$

However, different from the ME-principle in the probabilistic case, even minimal c-representations are not uniquely determined. It is still an open problem of research to specify conditions for unique c-representations. For the knowledge discovery problem dealt with in this paper, this is not a severe problem, as the ranking function is not searched for, but is derived from the given empirical distribution.

In summary, any ordinal conditional function κ that is indifferent with respect to a set of conditionals $\mathcal{R}^{\mathrm{OCF}}$ follows the conditional structures that are imposed by the conditionals in $\mathcal{R}$ onto the worlds and is thus most adequate to represent ordinal conditional knowledge.

In the following, we will put these ideas in formal, algebraic terms and prepare the theoretical grounds for the data mining techniques to be presented in this paper.

5 Discovering structural information

In this section, we will describe our approach to knowledge discovery which is based on the algebraic theory of conditionals sketched above. More precisely, we will show how to compute sets $\mathcal{R}$, or $\mathcal{R}^{\mathrm{OCF}}$, respectively, of (quantified) default rules that are apt to generate some given (finite) ordinal conditional function κ^P that is indifferent with respect to $\mathcal{R}$, respectively $\mathcal{R}^{\mathrm{OCF}}$. κ^P has been chosen to represent the observed statistical data P, as has been described in Section 3. More details and all proofs can be found in [11]; a generalization to multivalued variables (instead of bivalued variables) is straightforward.

In our scenario, an empirically obtained probability distribution P is given that may simply consist of relative frequencies. Usually, the aim of a data mining task is to compute a set of probabilistic rules $\mathcal{R}^{prob} = \{(B_1|A_1)[x_1], \ldots, (B_n|A_n)[x_n]\}$, such that this set predicts P best. This task was handled in [5] and also uses the algebraic theory of conditionals sketched above. The problem with the approach of [5] is that usually the empirically obtained probability distribution P is noisy and one can not find an appropriate (and particularly compact) set of probabilistic rules $\mathcal{R}^{prob}$ that explains the observed P. The set of computed probabilistic rules tends to be large and the rules are getting too specific to be helpful in a general context. On this account we present an alternative approach to knowledge discovery that also makes use of the algebraic theory of conditionals of [11] but is based on a representation of P by qualitative probabilities, i. e. by rankings.

As a first step, the probability distribution P is qualified in a sense, that we compute its ranking representation κ^P regarding equation (4). By doing this, we fuse several similar probabilities, that should not be distinguished in a qualitative setting, to one rank value of the obtained ranking function. Therefore we minimize the noise, that could be present in the original distribution P, obtaining a qualitative representation. We can now use the formalism of the algebraic theory of conditionals to compute a set $\mathcal{R}^{\mathrm{OCF}} = \{(B_1|A_1)[m_1], \ldots, (B_n|A_n)[m_n]\}$ of OCF-conditionals, that best explains κ^P, i. e., that is (in the best case) a faithful representation of κ^P. More precisely, we are looking for a set $\mathcal{R}$ of (unquantified) conditionals, such that κ^P is indifferent with respect to $\mathcal{R}$, i. e., $ker_0\ \sigma_{\mathcal{R}} \subseteq ker_0\ \kappa^P$ by Proposition 1. Ideally, we would have κ^P to represent $\mathcal{R}$ faithfully, that is,

$$\kappa^P \models \mathcal{R} \text{ and } ker_0\ \kappa^P = ker_0\ \sigma_{\mathcal{R}} \tag{12}$$

This means κ^P is indifferent with respect to $\mathcal{R}$, and no equation $\kappa^P(\hat{\omega}) = 0$ is fulfilled accidentally, but any of these equations is induced by $\mathcal{R}$.

Finally, we can assign rankings to these conditionals, derived immediately from κ^P thus obtaining a set $\mathcal{R}^{\mathrm{OCF}}$ of OCF-conditionals.

Under the assumption of faithfulness, the structures of the conditionals in $\mathcal{R}$ become manifest in the elements of $ker_0\ \kappa^P$, that is, in elements $\hat{\omega} \in \hat{\Omega}$ with $\kappa^P(\hat{\omega}) = 0$. As a further prerequisite, we will assume that this knowledge inherent to κ^P is representable by a set of single-elementary conditionals. This restriction is not too hard, because single-elementary conditionals are expressive enough to represent most commonsense knowledge. As our approach will work for any given ranking function κ, we omit the superscript P in this section.

So assume $\mathcal{R}^{\mathrm{OCF}} = \{(b_1|A_1)[m_1], \ldots, (b_n|A_n)[m_n]\}$ is an existing, but hidden set of single-elementary conditionals, such that (12) holds. Let us further suppose that $ker_0\ \kappa$ (or parts of it) is known from exploiting numerical relationships. Since conditional indifference is a structural notion, the quantifications m_i of the conditionals will not be needed in what follows. Let $\sigma_{\mathcal{R}} : \hat{\Omega} \to \mathfrak{F}_{\mathcal{R}} = \langle \mathbf{a}_1^+, \mathbf{a}_1^-, \ldots, \mathbf{a}_n^+, \mathbf{a}_n^- \rangle$ 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 $(B|A), (D|C) \in (\mathcal{L}|\mathcal{L})$ with $ABC\overline{D} \equiv A\overline{B}CD \equiv \bot$, the supremum $(B|A) \sqcup (D|C)$ in $(\mathcal{L}|\mathcal{L})$ with respect to $\sqsubseteq$ exists and is given by

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

(cf. [11]). 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 1. *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 (5), Definition 4 and Proposition 1, it is clear that in an inductive reasoning process such as a 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 $ker_0\ \kappa$. 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:

$$\psi_{v,l} = (v \mid C_{v,l}) \tag{13}$$

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 $(\mathcal{L}|\mathcal{L})$, and let $\mathfrak{F}_{\mathcal{B}} = \langle \mathbf{b}^{+}_{v,l}, \mathbf{b}^{-}_{v,l}\rangle_{v,l}$ be the free abelian group corresponding to $\mathcal{B}$ with conditional structure homomorphism $\sigma_{\mathcal{B}} : \hat{\Omega} \to \mathfrak{F}_{\mathcal{B}}$. Note that $\sigma_{\mathcal{B}}$ and $\mathfrak{F}_{\mathcal{B}}$ are known, whereas $\sigma_{\mathcal{R}}$ and $\mathfrak{F}_{\mathcal{R}}$ are not. We only know the kernel, $ker_0\ \sigma_{\mathcal{R}}$, of $\sigma_{\mathcal{R}}$, which is, by assuming faithfulness (12), the same as the kernel, $ker_0\ \kappa$, of κ. Now, to establish a connection between what is obvious ($\mathcal{B}$) and what is searched for ($\mathcal{R}$), we define a homomorphism $g : \mathfrak{F}_{\mathcal{B}} \to \mathfrak{F}_{\mathcal{R}}$ via

$$g(\mathbf{b}^{\pm}_{v,l}) := \prod_{\substack{1 \leq i \leq n \\ \psi_{v,l} \sqsubseteq (b_i|A_i)}} \mathbf{a}_i^{\pm} = \prod_{\substack{1 \leq i \leq n \\ b_i = v, C_{v,l} \models A_i}} \mathbf{a}_i^{\pm}, \tag{14}$$

where the second equality holds due to Lemma 1. 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 2. *Let $g : \mathfrak{F}_{\mathcal{B}} \to \mathfrak{F}_{\mathcal{R}}$ be as in (14). Then*

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

In particular, $\hat{\omega} \in ker_0\ \sigma_{\mathcal{R}} = ker_0\ \kappa$ iff $\hat{\omega} \in \hat{\Omega}_0$ and $\sigma_{\mathcal{B}}(\hat{\omega}) \in ker\ g$.

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

Proposition 4. *Let $\hat{\omega} = \omega_1^{r_1} \cdot \ldots \cdot \omega_m^{r_m} \in \hat{\Omega}_0$. Then $\sigma_{\mathcal{B}}(\omega_1^{r_1} \cdot \ldots \cdot \omega_m^{r_m}) \in ker\ g$ iff for all literals v in $\mathcal{L}$,*

$$\prod_{C_{v,l}} \prod_{\substack{1 \le k \le m \\ \omega_k \models C_{v,l} v}} (\mathbf{b}_{v,l}^+)^{r_k}, \prod_{C_{v,l}} \prod_{\substack{1 \le k \le m \\ \omega_k \models C_{v,l} \overline{v}}} (\mathbf{b}_{v,l}^-)^{r_k} \in ker\ g. \tag{15}$$

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 $\mathfrak{F}_{\mathcal{B}}$. Moreover, Proposition 4 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 [11]).

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

$$ker_0\ \sigma_{\mathcal{S}^{(0)}} \subseteq ker_0\ \sigma_{\mathcal{S}^{(1)}} \subseteq \ldots \subseteq ker_0\ \sigma_{\mathcal{R}} = ker_0\ \kappa \tag{16}$$

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

Lemma 2. *$\sigma_{\mathcal{B}}$ is injective, i.e. $ker_0\ \sigma_{\mathcal{B}} = \{1\}$.*

So $\sigma_{\mathcal{B}}$ provides the most finely grained conditional structure on $\hat{\Omega}$: No different elements $\hat{\omega}_1 \neq \hat{\omega}_2$ 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 κ). Each $\mathcal{S}^{(t)}$ is assumed to be a set of conditionals $\phi_{v,j}^{(t)}$ with a single literal v in the conclusion, and the antecedent $D_{v,j}^{(t)}$ of $\phi_{v,j}^{(t)}$ is a disjunction of elementary conjunctions not mentioning the variable V. Let $\mathfrak{F}_{\mathcal{S}^{(t)}} = \langle \mathbf{s}_{v,j}^{(t)+}, \mathbf{s}_{v,j}^{(t)-} \rangle_{v,j}$ be the free abelian group associated with $\mathcal{S}^{(t)}$, and $\sigma_{\mathcal{S}^{(t)}} : \hat{\Omega} \rightarrow \mathfrak{F}_{\mathcal{S}^{(t)}}$ the corresponding structure homomorphism; let $g^{(t)} : \mathfrak{F}_{\mathcal{S}^{(t)}} \rightarrow \mathfrak{F}_{\mathcal{R}}$ be the homomorphism defined by

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

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 $\mathbf{s}_1, \mathbf{s}_2 \in \mathfrak{F}_{\mathcal{S}^{(t)}}$. In the following, for ease of notation, we will omit the $+, -$ superscripts on group generators; this is justified, since, by Proposition 4, 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 $ker_0\ \kappa$ by applying Theorem 2 and the faithfulness assumption (12) is of the form

$$\mathbf{s}_{v,j_0}^{(t)} \equiv_{g^{(t)}} \mathbf{s}_{v,j_1}^{(t)} \ldots \mathbf{s}_{v,j_m}^{(t)} \tag{17}$$

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

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

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

It can be shown (cf. [11]) that

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

and hence

$$ker_0\ \sigma_{\mathcal{S}^{(t)}} \subseteq ker_0\ \sigma_{\mathcal{S}^{(t+1)}} \subseteq ker_0\ \sigma_{\mathcal{R}}$$

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.

6 Learning default rules from data

In this section, we will describe an adjusted version of the *CKD*-algorithm (= *Conditional Knowledge Discovery*) for the determination of default rules from qualitative approximations of statistical data. This algorithm is sketched in Figure 2. The original *CKD*-algorithm for mining probabilistic conditionals from statistical data has been implemented in the CONDOR-system (for an overview, cf. [24]). The resulting set of default rules or OCF-conditionals will reveal relevant relationships and may serve to represent inductively the corresponding ordinal conditional function faithfully.

A problem that has already been mentioned but postponed in section 5 is that the set $\mathcal{B}$ of *all* basic single elementary conditionals is virtually unmanageable. Therefore it cannot be used as an adequate starting set in the algorithm. Another problem emerges from the frequency distributions calculated from a data set. In a realistic setting, these distributions are sparse, i. e., they deliver zero values for many worlds. Hence, the probability value of these worlds is zero as well and according to Definition 3, a world with a zero probability is assigned an infinite ranking value. Besides calculational difficulties, the correct interpretation of such worlds, which have not been observed in the analyzed data and therefore have a frequency of zero, is not clear: On the one hand, these

```
Algorithm CKD for OCFs
(Conditional Knowledge Discovery)

Input   A probability distribution $P$ obtained from statistical data,
        (only explicitly listing those entries with positive probabilities)
        together with information on variables and appertaining values
        and an abstraction parameter $\varepsilon \in (0, 1)$
Output  A set of OCF-conditionals (default rules)

Begin
  % Qualitative representation
  Calculate the ranking value $\tilde{\kappa}_\varepsilon^P(\omega)$ for each input value $P(\omega)$;
  Normalize $\tilde{\kappa}_\varepsilon^P$ to obtain the ordinal conditional function $\kappa_\varepsilon^P$;

  % CKD Initialization
  Compute the basic tree of conjunctions;
  Compute the list NC of null-conjunctions;
  Compute the set $\mathcal{S}^{(0)}$ of basic rules;
  Compute $ker_0 \; \kappa_\varepsilon^P$;
  Compute $ker \; g$;
  Set $\mathcal{K} := ker \; g$;
  Set $\mathcal{S} := \mathcal{S}^{(0)}$;

  % CKD Main loop
  While equations of type (17) are in $\mathcal{K}$ Do
    Choose $gp \in \mathcal{K}$ of type (17);
    Modify (and compactify) $\mathcal{S}$;
    Modify (and reduce) $\mathcal{K}$;

  % Present results
  Calculate the degrees of belief of the conditionals in $\mathcal{S}$;
  Return $\mathcal{S}$ and appertaining degrees of belief;
End.
```

Fig. 2. The CKD-algorithm for OCFs

worlds might just have *not been captured* when recorded the data; perhaps because the amount of recorded data was not large enough and they have merely been missed. In this case, assigning these worlds a zero probability would be misleading. On the other hand, these worlds might *not exist* at all (and could therefore not have been recorded), so a zero probability would be completely justified; but this could never be assured by pure observation. The problem of zero probabilities is addressed more deeply in [5].

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 an observed frequency distribution 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 an unknown but equal probability, we have $P(\omega_1) = P(\omega_2)$; with κ^P being the corresponding ordinal conditional function, this leads to $\kappa^P(\omega_1) = \kappa^P(\omega_2)$ and hence $\frac{\omega_1}{\omega_2} \in ker_0\ \kappa^P$. In this way, all these so-called *null-worlds* contribute to $ker_0\ \kappa^P$, 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 ranking value ∞.

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 4, 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 finite ranking values, 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.

Now, that the missing values in the frequency distribution corresponding to infinite degrees of disbelief have been absorbed by the shortened basic rules, we explore the finite rankings derived from P to set up $ker_0\ \kappa^P$. Usually, numerical relationships $\kappa^P(\hat{\omega}) = 0$ induced by 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 κ^P, the vertices of which are the non-∞-worlds, labelled by their finite ranking values, and with edges connecting any two neighbors. Then any such relationship $\kappa^P(\hat{\omega}) = 0$ corresponds to a cycle of even length (i. e. involving an even number of vertices) in the neighbor graph, such that the alternating sum built from the values associated with the vertices, with alternating coefficients $+1$ and -1 according to the order of vertices in the cycle, amounts to 0. Therefore, the search for numerical relationships holding in κ^P amounts to searching for cycles with sum 0 in the neighbor graph.

At this point, an important advantage of using qualitative probabilities, i. e., ranking values, becomes clear: Because the ranking values are discrete values, we can demand that the vertices of a cycle must sum up to *exactly* zero. In the approach of [5] that uses the empirically obtained probabilities directly, one can only demand that vertices of a cycle must *approximately* fulfill the corresponding equation, because equality can usually not be reached when calculating with the exact probabilities, i. e., with continuous values. So in the approach of [5] the important step of exploring the numerical relationships depends implicitly on the notion of ”approximately”. But by using an (appropriate) explicit parameter ε for the qualitative abstraction of the original probabilities, the search for numerical relationships is defined precisely.

Finally, as the last step of the initialization, $ker\ g$ has to be computed from $ker_0\ \kappa^P$ with respect to the set $\mathcal{S}^{(0)}$ of conditionals, as described in the previous section.

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, $\mathcal{K} = ker\ g$ 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 (17) 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 (17) can be found in $\mathcal{K}$. Note that no ranking values are used in this main loop – only structural information (derived from numerical information) is processed. It is only afterwards, that the ranking values of the conditionals in the final set $\mathcal{S}$ are computed from κ^P, and the OCF-conditionals (default rules) are returned.

Although equations of type (17) 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.

Example 7. (Continuing Example 3)
From the observed probabilities, we calculate qualitative probabilities, using $\varepsilon = 0.6$ as base value. We adjust the calculated qualitative probabilities by subtracting the normalization constant $c = 2$, so that the lowest ranking becomes 0. This gives us the ranking values $\kappa^P(\omega)$ that define the ordinal conditional function κ^P, as can be seen from Table 4.

object	*frequency*	*probability*	*rank*
$abcde$	59	0.5463	0
$abcd\overline{e}$	21	0.1944	2
$a\overline{b}cde$	11	0.1019	3
$a\overline{b}cd\overline{e}$	9	0.0833	3
$abc\overline{d}e$	6	0.0556	4
$abc\overline{d}\overline{e}$	2	0.0185	6

Table 4. Empirical probabilities and corresponding ranking values

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:

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

So, the missing information reflected by the set NC of 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 balanced:

$$\kappa^P(a\overline{b}cde) = \kappa^P(a\overline{b}cd\overline{e}) \quad \text{and} \quad \kappa^P(abcde) - \kappa^P(abcd\overline{e}) = \kappa^P(abc\overline{d}e) - \kappa^P(abc\overline{d}\overline{e})$$

At this point, it becomes clear how crucial an appropriate choice for ε is. If ε had been chosen too high, e. g. $\varepsilon = 0.9$ as in Example 3, then the neighboring worlds ω_3 and ω_4 would have been assigned different ranking values, so the first numerical relationship would not hold. Thus an important piece of structural information would have been missed. On the other hand, if ε had been chosen much too small, e. g. $\varepsilon = 0.01$, then all worlds would have been projected to the same ranking value. Thus relationships between all neighboring worlds would have been established, leading to no useful results.

Continuing the example, the first relationship can be translated into the following structural equations by using $\sigma_{\mathcal{B}}$, according to Theorem 2:

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

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 κ^P, we obtain $\mathbf{b}^\pm_{d,1} \equiv \mathbf{b}^\pm_{d,2}$ and $\mathbf{b}^\pm_{e,1} \equiv \mathbf{b}^\pm_{e,2}$, that is, the corresponding conditionals have to be joined. As a final output, the CKD algorithm returns the set of conditionals that is shown in Table 5.
All these conditionals are accepted in κ^P. For each of them the degree of belief regarding κ^P can be stated as well as the probability regarding the observed distribution P. So all objects in our universe are aquatic animals which are fish or have gills. Aquatic animals with gills are mostly fish (with a degree of belief 3 and a probability of 0.80), aquatic fish usually have gills (with a degree of belief 4 and a probability of 0.91) and scales (with a degree of belief 2 and a probability of 0.74).

Furthermore, an approximated probability based on the ranking values can be calculated for each conditional $(B|A)$. Because the ranking values are determined according to equation (4), each probability $P(\omega)$ is qualitatively approximated by its corresponding ranking value, so we have:

$$P(\omega) \approx \varepsilon^{\kappa^P(\omega)} \tag{18}$$

conditional	*empirical probability*	*degree of belief*
$(a\|\top)$	1	∞
$(b\|\overline{d})$	1	∞
$(b\|acd)$	0.80	3
$(e\|abc)$	0.74	2
$(c\|\top)$	1	∞
$(d\|\overline{b})$	1	∞
$(d\|abc)$	0.91	4

Table 5. Conditionals calculated by the CKD algorithm

By taking into consideration the equation

$$P(B|A) = \frac{1}{\frac{P(A)}{P(AB)}} = \frac{1}{\frac{P(AB)+P(A\overline{B})}{P(AB)}} = \frac{1}{1+\frac{P(A\overline{B})}{P(AB)}} \quad ,$$

we can approximate the probability of a conditional by its degree of belief[4] m:

$$P(B|A) \approx \frac{1}{1+\frac{\varepsilon^{\kappa^P(A\overline{B})}}{\varepsilon^{\kappa^P(AB)}}} = \frac{1}{1+\varepsilon^{\kappa^P(A\overline{B})-\kappa^P(AB)}} = \frac{1}{1+\varepsilon^m} \tag{19}$$

Example 8. (Continuing Example 7)
The application of formula (19) results in approximated conditional probabilities, listed in Table 6. Compared to the exactly calculated empirical probabilities (cf. Table 5),

conditional	*degree of belief*	*approx. probability*
$(a\|\top)$	∞	1
$(b\|\overline{d})$	∞	1
$(b\|acd)$	3	0.82
$(e\|abc)$	2	0.74
$(c\|\top)$	∞	1
$(d\|\overline{b})$	∞	1
$(d\|abc)$	4	0.89

Table 6. Conditionals and their approximated probabilities

[4] At this point, is does not matter whether the ranking values originating directly from equation (4) or the normalized ones are used, because the normalization constant will be cancelled out when considering conditionals.

the deviation of the approximated conditional probabilities is comparatively small. Although the statistical probability values $P(\omega)$ have been abstracted by qualitative values and the approximation in equation (18) might appear somewhat coarse, the results are nevertheless quite accurate. This illustrates that the qualitative abstraction of the original probabilities conserves enough information to be useful in handling questions of structural relationship.

7 Summary and further work

We have proposed an approach to qualitative knowledge discovery that followed the mechanisms of reverse inductive knowledge representation developed in [5] but is based on a qualitative representation of the empirically obtained probability distribution P that serves as input to the data mining process. An ordinal conditional function κ^P based on qualitative probabilities [1] was used to capture the qualitative information inherent to P. With the use of an algebraic theory of conditionals, the approach generates default rules that are apt to compactly represent the information of κ^P. We briefly described the theoretical and methodological background, and also made clear how our method can be implemented by sketching an algorithm.

A problem of open research is the question, of how to determine the abstraction parameter that is needed to represent the probabilities as polynomials in that parameter in an optimal way. As mentioned before, this determination is crucial when computing the qualitative abstractions of the information inherent in the original distribution because the precision of the computed qualitative representation depends particularly on the chosen parameter.

The purely probabilistic version of the described algorithm has been developed and implemented during the CONDOR-project[5]. CONDOR is an integrated system for learning, reasoning and belief revision in a probabilistic environment. For future work, we are planning to implement the algorithm for qualitative knowledge discovery presented in this paper and integrate it into CONDOR to also provide qualitative learning and reasoning facilities. The common methodological grounds based on c-representations which can be used both for probabilistic and default reasoning will establish clear links between quantitative and qualitative frameworks, as was illustrated in the running example of this paper.

References

1. Goldszmidt, M., Pearl, J.: Qualitative probabilities for default reasoning, belief revision, and causal modeling. Artificial Intelligence (1996)
2. Benferhat, S., Dubois, D., Prade, H.: Nonmonotonic reasoning, conditional objects and possibility theory. Artificial Intelligence (92) (1997) 259276
3. Kraus, S., Lehmann, D., Magidor, M.: Nonmonotonic reasoning, preferential models and cumulative logics. Artificial Intelligence **44** (1990) 167–207
4. Kern-Isberner, G.: Solving the inverse representation problem. In: Proceedings 14th European Conference on Artificial Intelligence, ECAI’2000, Berlin, IOS Press (2000) 581–585

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

5. Kern-Isberner, G., Fisseler, J.: Knowledge discovery by reversing inductive knowledge representation. In: Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning, KR-2004, AAAI Press (2004) 34–44
6. Adams, E.: Probability and the logic of conditionals. In Hintikka, J., Suppes, P., eds.: Aspects of inductive logic. North-Holland, Amsterdam (1966) 265–316
7. Spohn, W.: Ordinal conditional functions: a dynamic theory of epistemic states. In Harper, W., Skyrms, B., eds.: Causation in Decision, Belief Change, and Statistics. Volume 2. Kluwer Academic Publishers (1988) 105–134
8. Benferhat, S., Dubois, D., Lagrue, S., Prade, H.: A big-stepped probability approach for discovering default rules. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems(IJUFKS) **11** (2003) 1–14
9. Paris, J.: The uncertain reasoner's companion – A mathematical perspective. Cambridge University Presse (1994)
10. Kern-Isberner, G.: Characterizing the principle of minimum cross-entropy within a conditional-logical framework. Artificial Intelligence **98** (1998) 169–208
11. Kern-Isberner, G.: Conditionals in nonmonotonic reasoning and belief revision. Springer, Lecture Notes in Artificial Intelligence LNAI 2087 (2001)
12. Jain, A., Murty, M., Flynn, P.: Data clustering: A review. ACM Computing Surveys **31**(3) (1999)
13. Cooper, G., Herskovits, E.: A bayesian method for the induction of probabilistic networks from data. Machine Learning **9** (1992) 309–347
14. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction and Search. Number 81 in Lecture Notes in Statistics. Springer (1993)
15. Heckerman, D.: Bayesian networks for knowledge discovery. In Fayyad, U., Piatsky-Shapiro, G., Smyth, P., Uthurusamy, R., eds.: Advances in knowledge discovery and data mining. MIT Press, Cambridge, Mass. (1996)
16. Buntine, W.: A guide to the literature on learning probabilistic networks from data. IEEE Transactions on Knowledge and Data Engineering **8**(2) (1996) 195–210
17. Herskovits, E., Cooper, G.: Kutató: An entropy-driven system for construction of probabilistic expert systems from databases. Technical Report KSL-90-22, Knowledge Systems Laboratory (1990)
18. Geiger, D.: An entropy-based learning algorithm of bayesian conditional trees. In: Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence. (1992) 92–97
19. Kern-Isberner, G.: Following conditional structures of knowledge. In: KI-99: Advances in Artificial Intelligence, Proceedings of the 23rd Annual German Conference on Artificial Intelligence, Springer Lecture Notes in Artificial Intelligence LNAI 1701 (1999) 125–136
20. Kern-Isberner, G.: Handling conditionals adequately in uncertain reasoning. In: Proceedings European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU'01, Springer LNAI 2143 (2001) 604–615
21. DeFinetti, B.: Theory of Probability. Volume 1,2. John Wiley and Sons, New York (1974)
22. Calabrese, P.: Deduction and inference using conditional logic and probability. In Goodman, I., Gupta, M., Nguyen, H., Rogers, G., eds.: Conditional Logic in Expert Systems. Elsevier, North Holland (1991) 71–100
23. Kern-Isberner, G.: A thorough axiomatization of a principle of conditional preservation in belief revision. Annals of Mathematics and Artificial Intelligence **40(1-2)** (2004) 127–164
24. Beierle, C., Kern-Isberner, G.: Modelling conditional knowledge discovery and belief revision by abstract state machines. In Boerger, E., Gargantini, A., Riccobene, E., eds.: Abstract State Machines 2003 – Advances in Theory and Applications, Proceedings 10th International Workshop, ASM2003, Springer, LNCS 2589 (2003) 186–203