An alternative view of knowledge discovery

Christoph Beierle, Gabriele Kern-Isberner
Praktische Informatik VIII - Wissensbasierte Systeme
Fachbereich Informatik, FernUniversität Hagen
58084 Hagen
Germany
E-mail: {christoph.beierle | gabriele.kern-isberner}@fernuni-hagen.de

Abstract

Inductive representation of conditional knowledge means to complete knowledge appropriately and can be looked upon as an instance of quite a general representation problem. The crucial problem of discovering relevant conditional relationships in statistical data can also be addressed in this formal framework. The main point in this paper is to consider knowledge discovery as an operation which is inverse to inductive knowledge representation, giving rise to phrasing the inverse representation problem. This allows us to embed knowledge discovery in a theoretical framework where the vague notion of relevance can be given a precise meaning: relevance here means relevance with respect to an inductive representation method. In order to exemplify our ideas, we present an approach to compute sets of conditionals from statistical data, which are optimal with respect to the information-theoretical principle of maximum entropy.

Keywords: data mining, knowledge discovery, knowledge representation, reasoning under uncertainty, probabilistic conditionals, inverse representation problem

1. Introduction

Commonsense and expert knowledge is most generally expressed by rules, connecting a precondition and a conclusion by an *if-then*-construction. If-then-rules often occur in the form of *probabilistic conditionals*. For instance, such conditionals may express commonsense knowledge like "*Students are young with a probability of (about) 80 %*" and "*Singles (i.e. unmarried people) are young with a probability of (about) 70 %*", the latter knowledge being formally expressed by $\{(young|student)[0.8], (young|single)[0.7]\}$.

In this paper, we address two crucial problems which arise at once when probabilistic conditionals are used for knowledge representation and reasoning:

- How to combine knowledge expressed by conditionals so as to yield expressive answers to queries?
- Where do the (probabilistic) conditionals apt to represent knowledge appropriately come from?

Each of these questions has been investigated for quite a long time. The proper use of rules has been discussed since the days of the first rule-based systems (cf. e.g. [11]) and has also been a topic in philosophical studies [1]. A modern and quite effective approach to represent probabilistic rules is provided, for instance, by Bayesian networks (cf. e.g. [33, 13]), or by the powerful maximum entropy approach [32]. As to the second question above – which actually should be considered in the first place –, one tends to assume that some omniscient expert is able to express his knowledge as (probabilistic) rules. In practice, however, statistical data are often used to (at least) support the building of knowledge bases. Techniques and tools to tackle this problem have been developed in the areas of machine learning, and knowledge discovery and data mining (for an overview, see [17, 16]).

The close relationship between these two problems is obvious: Preferably those conditionals should be discovered the combination of which yields most adequate answers to queries. In spite of this, the questions above have often been dealt with separately. The point of this paper is to provide a formal framework to deal jointly with both problems, and to make the interdependence between them clearly visible. In particular, the often quite vague criterion of *relevance* applied to the rules to be discovered, can be given formally a more precise meaning by referring to a particular knowledge representation method.

The research reported here was partially supported by the DFG – Deutsche Forschungsgemeinschaft within the CONDOR-project under grant BE 1700/5-1.

In detail, our line of argumentation is as follows: For the first problem, we propose a model-based solution by formalizing a general *representation problem*: Given a specification of a theory, select a set of models as its desired representation. The well-known probabilistic principle of maximum entropy (ME-principle) is easily seen to solve this representation problem for probabilistic conditionals in a most satisfying way [32]. Having phrased this general representation problem, we address the second problem: Statistical data may be summarized as a frequency distribution which constitutes a probabilistic model for the rules it represents. From our point of view, the important question how to extract rules from statistical data may thus be viewed as inverting the above mentioned representation problem, in that now a set of probabilistic conditionals (i.e. a probabilistic specification) has to be selected, given a model. So, the task of discovering rules from data can be considered as an instance of this abstract *inverse representation problem*. For the inductive representation of probabilistic conditionals via the ME-principle, we will illustrate the functionality of a KDD (Knowledge Discovery in Data) approach to solve this inverse representation problem, i.e. to compute a concise set of rules which are most relevant with respect to the ME-method.

The rest of this paper is organized as follows: In Section 2, we recall some basic definitions and establish our notation; moreover, we use the notion of institutions for a formalization of probabilitic conditional logic as suggested in [6] and [7]. The general notions of the representation problem and the inverse representation problem are introduced and illustrated in various application areas in Sections 3 and 4, respectively. Section 5 gives an example how KDD can be seen as an instance of the inverse representation problem, while Section 6 gives some conclusions and points out further work.

2 The institution of probabilistic conditional logic

2.1. Theories and their presentations in the framework of institutions

As a general framework for logical systems, Goguen and Burstall introduced the notion of an institution [18]. An institution formalizes the informal notion of a logical system, including syntax, semantics, and the relation of satisfaction between them. The latter poses the major requirement for an institution: that the satisfaction relation is consistent under the change of notation.

Institutions have been used for the general study of logics. For instance, there are widely applicable results about building up larger theories from smaller components. Institution morphisms [20] support the comparison of different logics, they are used for glueing together several logics within one system, and they may permit a theorem prover for one institution to be used on theories from another one. Additionally, institutions have also been used as a basis for specification and development languages; in fact, institutions arose in the context of designing the specification language Clear [12, 18]. For some of the work using institutions see e.g. [19, 34, 9, 37, 20].

Before going into the details of the definition of an institution, we briefly recall some basic facts about category theory which institutions use as a framework.

If C is a category, $|C|$ denotes the objects of C and $/C/$ its morphisms; for both objects $c \in |C|$ and morphisms $\varphi \in /C/$, we also write just $c \in C$ and $\varphi \in C$, respectively. C^{op} is the opposite category of C, with the direction of all morphisms reversed. $\mathcal{SET}$ and $\mathcal{CAT}$ denote the categories of sets and of categories, respectively. (For more information about categories, see e.g. [21] or [30].)

The central definition of an institution [18] is the following (cf. Figure 1 that visualizes the relationships within an institution):

Definition 1 *An institution is a quadruple*

$$Inst \;=\; \langle\, Sig,\; Mod,\; Sen,\; \models \,\rangle$$

with a category Sig of signatures as objects, a functor $Mod : Sig \to \mathcal{CAT}^{op}$ yielding the category of Σ-models for each signature Σ, a functor $Sen : Sig \to \mathcal{SET}$ yielding the sentences over a signature, and a $|Sig|$-indexed relation $\models_\Sigma \;\subseteq\; |Mod(\Sigma)| \times Sen(\Sigma)$ such that for each signature morphism $\varphi : \Sigma \to \Sigma' \in /Sig/$, for each $m' \in |Mod(\Sigma')|$, and for each $f \in Sen(\Sigma)$ the following satisfaction condition *holds:*

$$m' \models_{\Sigma'} Sen(\varphi)(f) \quad \textit{iff} \quad Mod(\varphi)(m') \models_\Sigma f$$

We illustrate this definition by formalizing propositional logic as an institution.

Example 2 The *institution of propositional logic* is denoted by

$$Inst_{\mathcal{B}} \;=\; \langle\, Sig_{\mathcal{B}},\; Mod_{\mathcal{B}},\; Sen_{\mathcal{B}},\; \models_{\mathcal{B}} \,\rangle$$

and its components are as expected:

1. **Signatures:** A propositional signature $\Sigma \in Sig_{\mathcal{B}}$ is a set of propositional variables, $\Sigma = \{a_1, a_2, \ldots\}$.

2. **Models:** $Mod_{\mathcal{B}}(\Sigma)$ contains the set of all propositional interpretations for Σ, i.e.

$$Mod_{\mathcal{B}}(\Sigma) \;=\; \{I \mid I : \Sigma \to Bool\}$$

where $Bool = \{\textit{true}, \textit{false}\}$.

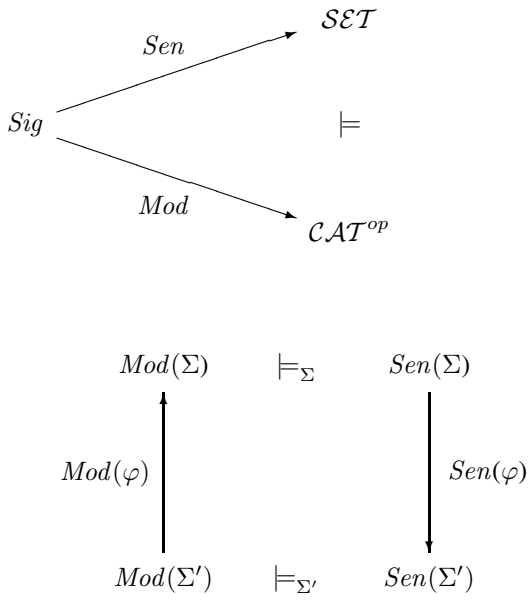

Figure 1. Relationships within an institution $Inst = \langle Sig, Mod, Sen, \models \rangle$ **(cf. [18])**

3. **Sentences:** The set $Sen_{\mathcal{B}}(\Sigma)$ contains the usual propositional formulas constructed from the propositional variables in Σ and the logical connectives $\wedge$ (and), $\vee$ (or), and $\neg$ (not). The symbols $\top$ and $\bot$ denote a tautology (like $a \vee \neg a$) and a contradiction (like $a \wedge \neg a$), respectively. In order to simplify notations, we will often replace conjunction by juxtaposition and indicate negation of a formula by barring it, i.e. $AB = A \wedge B$ and $\overline{A} = \neg A$. As usual, an *atom* is a formula consisting of just a propositional variable, a *literal* is a positive or a negated atom, an *elementary conjunction* is a conjunction of literals, and a *complete conjunction* is an elementary conjunction containing each atom either in positive or in negated form. Ω_Σ denotes the set of all complete conjunctions over a signature Σ; if Σ is clear from the context, we may drop the index Σ.

   Note that there is an obvious bijection between $|\, Mod_{\mathcal{B}}(\Sigma)\,|$ and Ω_Σ, associating with $I \in |\, Mod_{\mathcal{B}}(\Sigma)\,|$ the complete conjunction $\omega_I \in \Omega_\Sigma$ in which an atom $a_i \in \Sigma$ occurs in positive form iff $I(a_i) = \mathit{true}$.

4. **Satisfaction relation:** The satisfaction relation is also defined as expected for propositional logic, e.g. $I \models_{\mathcal{B},\Sigma} a_i$ iff $I(a_i) = \mathit{true}$ and $I \models_{\mathcal{B},\Sigma} A \wedge B$ iff $I \models_{\mathcal{B},\Sigma} A$ and $I \models_{\mathcal{B},\Sigma} B$ for $a_i \in \Sigma$ and $A, B \in Sen_{\mathcal{B}}(\Sigma)$.

□

For sets F, G of Σ-sentences and a Σ-model m we write $m \models_\Sigma F$ iff $m \models_\Sigma f$ for all $f \in F$. The satisfaction relation is lifted to semantical entailment $\models_\Sigma$ between sentences by defining $F \models_\Sigma G$ iff for all Σ-models m with $m \models_\Sigma F$ we have $m \models_\Sigma G$. $F^\bullet = \{f \in Sen(\Sigma) \mid F \models_\Sigma f\}$ is called the *closure* of F, and F is *closed* if $F = F^\bullet$. A Σ-*presentation* is a pair $\langle \Sigma, F \rangle$ with $F \subseteq Sen(\Sigma)$, and a Σ-*theory* is a presentation $\langle \Sigma, F \rangle$ such that F is closed under semantical entailment, i.e. $F = F^\bullet$. $Mod(\langle \Sigma, F \rangle)$ denotes the full subcategory of $Mod(\Sigma)$ of all Σ-models that satisfy F.

2.2. Probabilistic conditional logic

We will first give a very short introduction to probabilistics as far as it is needed here: Let $\Sigma \in |\, Sig_{\mathcal{B}}\,|$ be a propositional signature. A *probability distribution* (or *probability function*) *over* Σ is a function $P : Sen_{\mathcal{B}}(\Sigma) \to [0,1]$ such that $P(\top) = 1$, $P(\bot) = 0$, and $P(A \vee B) = P(A) + P(B)$ for any formulas $A, B \in Sen_{\mathcal{B}}(\Sigma)$ with $AB = \bot$. Each probability distribution P is determined uniquely by its values on the complete conjunctions[1] $\omega \in \Omega_\Sigma$, since

$$P(A) = \sum_{\omega \in \Omega_\Sigma, \omega \models_{\mathcal{B},\Sigma} A} P(\omega) \qquad (1)$$

For two propositional formulas $A, B \in Sen_{\mathcal{B}}(\Sigma)$ with $P(A) > 0$, the *conditional probability of* B *given* A is

$$P(B|A) := \frac{P(AB)}{P(A)}$$

[1] Note that complete conjunctions correspond to elementary events

Based on $Inst_{\mathcal{B}}$, the institution of probabilistic conditional logic is given by

$$Inst_{\mathcal{C}} = \langle Sig_{\mathcal{C}}, Mod_{\mathcal{C}}, Sen_{\mathcal{C}}, \models_{\mathcal{C}} \rangle$$

with components defined as follows:

Signatures: $Sig_{\mathcal{C}}$ is identical to the category of propositional signatures, i.e. $Sig_{\mathcal{C}} = Sig_{\mathcal{B}}$.

Models: For each signature Σ, the objects of $Mod_{\mathcal{C}}(\Sigma)$ are probability distributions over the propositional variables, i.e.

$$Mod_{\mathcal{C}}(\Sigma) = \{P \mid P \text{ is a probability distribution over } \Sigma\}.$$

Example 3 Let $\Sigma = \{a, b, c\}$ be a propositional signature with the atomic propositions *a – being a student*, *b – being young*, *c – being unmarried*. We define a Σ-model P by assigning a probability $P(\omega)$ to every complete conjunction ω over Σ:

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

P can be taken as the representation of some statistical information about a typical population, on the one hand, or as the representation of subjective beliefs concerning the relationships between the variables involved. For instance, the probability $P(abc)$ of *being a student*, *being young*, and *being unmarried* is given by 0.1950, and the probability $P(a\overline{b}c)$ of *being a student*, *not being young*, and *being unmarried* is 0.0408. □

Sentences: For each signature Σ, the set $Sen_{\mathcal{C}}(\Sigma)$ contains *probabilistic conditionals* (sometimes also called *probabilistic rules*) of the form

$$(B|A)[x]$$

where $A, B \in Sen_{\mathcal{B}}(\Sigma)$ are propositional formulas from $Inst_{\mathcal{B}}$. $x \in [0, 1]$ is a probability value indicating the degree of certainty for the occurrence of B under the condition A. Note that a probabilistic fact of the form $B[x]$ can easily be expressed as a conditional $(B|\top)[x]$ with a tautology as trivial antecedent.

Satisfaction relation: The satisfaction relation

$$\models_{\mathcal{C},\Sigma} \subseteq |Mod_{\mathcal{C}}(\Sigma)| \times Sen_{\mathcal{C}}(\Sigma)$$

is defined by

$$P \models_{\mathcal{C},\Sigma} (B|A)[x] \quad \text{iff}$$

$$P(A) > 0 \text{ and } P(B \mid A) = \frac{P(AB)}{P(A)} = x$$

Note that for probabilistic facts we have $P \models_{\mathcal{C},\Sigma} (B|\top)[x]$ iff $P(B) = x$ from the definition of the satisfaction relation since $P(\top) = 1$.

Example 4 Let Σ and P be as in Example 3. Then $P \models_{\mathcal{C},\Sigma} (b|\top)[0.6614]$ since the probability of *being young* is $P(b) = 0.6614$, as can be seen directly from equation (1). Furthermore, $P \models_{\mathcal{C},\Sigma} (b|a)[0.8]$ since the probability of *being young* under the condition of *being a student* is $P(b|a) = 0.8$. □

Presentation of probabilistic theories are pairs $\langle \Sigma, R \rangle$, consisting of a set Σ of propositional variables and a set R of probabilistic conditionals. While Σ lists the attributes which are relevant in the present investigation, R may be thought of as describing probabilistically important relationships between those attributes. In the example above, e.g. the conditional $(b|a)[0.8]$ could be an element of such a set R. In a medical environment, probabilistic conditionals can establish connections between symptoms and diseases, or in economics, they can reflect typical customer behavior.

3. The representation problem and preferred models

The original motivation for institutions was the definition of the semantics of the specification language Clear [12], and institutions have been used for various aproaches to modularized specification and programming development, often involving the notion of an abstract data type (ADT). When trying to use formal methods in software development, one quickly comes across the need for a rigorous method for specifying, refining, and implementing data types at levels that are independent from a specific representation used in e.g. traditional programming languages. Using institutions, a *specification* is a (theory) presentation $\langle \Sigma, F \rangle$ (cf. Sect. 2.1). This certainly meets the requirement of abstractness; but what does $\langle \Sigma, F \rangle$ represent? The general institution framework provides $Mod(\langle \Sigma, F \rangle)$ as a semantics for $\langle \Sigma, F \rangle$, but in many cases we are interested only in specific models. This is what we call the *representation problem*:

> Given a specification $\langle \Sigma, F \rangle$, select a class of (preferred) models $M \subseteq Mod(\langle \Sigma, F \rangle)$ as its desired representation.

In every specification approach based on logic, an answer to the representation problem must be given. In the initial approach to ADT specifications one is interested in models that are initial in the category $Mod(\langle \Sigma, F \rangle)$; other approaches take the terminal models, the finitely generated

models, or even all models as in so-called loose ADT specifications (cf. e.g. [12, 15, 34, 10, 38, 18]).

When it comes to reasoning, the representation problem is crucially relevant for other reasons. Whereas classical logical reasoning is done with respect to *all* models, reasoning with respect to the models selected according to the representation problem requires tailored inference techniques. For instance, when reasoning with respect to equationally defined initial ADTs, induction should be used since the initial models are finitely generated.

While for ADT specifications the selection of models $M \subseteq Mod(\langle \Sigma, F \rangle)$ requires special reasoning techniques, the selection itself is not motivated by these reasoning techniques. On the other hand, this is indeed the case for approaches in defeasible reasoning. Here, the motivation to focus on preferred models is to select models which are most appropriate for yielding plausible conclusions. The set of formulas F in a specification $\langle \Sigma, F \rangle$ is taken to specify incomplete knowledge, and basing entailment upon a relatively small set of models (the most plausible ones) means to extend the knowledge expressed by F, so as to derive more (tentative) conclusions than can be obtained by classical deduction. If there is only one most plausible model, then this model completes the available knowledge, and hence inductively represents F. Thereby, in order to get stronger ('better') inference capabilities, one deliberately accepts leaving the framework of classical logical reasoning and choosing e.g. the *preferential models* approach of nonmonotonic logics (cf. [35, 31, 29]) as appropriate paradigm.

In a probabilistic environment, the problem of plausible inference is even more difficult to be dealt with, since $Mod_{\mathcal{C}}(\langle \Sigma, R \rangle)$ typically contains a huge number of very different distributions, all reflecting the conditional knowledge given by R. Entailment based on all models is quite weak, e.g. it is not possible to derive the probability of a conjunction from the probability of each conjunct. In general, mostly intervals of possible probability values can be obtained which are often inexpressively large.

Example 5 Let R be a set containing two probabilistic conditionals both having C as its conclusion, one under the condition A and the other one under the condition B, i.e. $R = \{(C|A)[x], (C|B)[y]\}$ for some given probabilities $x, y \in (0, 1)$. What does this mean for the occurrence of C under the condition of both A and B? One can show that for *any* probability $z \in (0, 1)$ the set R is compatible with $(C|AB)[z]$, that is, for all (non-trivial) $x, y, z \in (0, 1)$, there is a probability distribution P such that both $P \models_{\mathcal{C}} R$ and $P \models_{\mathcal{C}} (C|AB)[z]$. So, actually nothing can be derived about the probability of $(C|AB)$ from R. □

The problem of yielding plausible inferences from a set R of probabilistic conditionals can be solved by the *principle of maximum entropy (ME-principle)* in the following way: This information-theoretical principle selects a distribution P^* from $Mod_{\mathcal{C}}(\langle \Sigma, R \rangle)$ whose entropy

$$H(P^*) = -\sum_{\omega \in \Omega} P^*(\omega) \log_2 P^*(\omega)$$

is maximal. Similar as for initial ADT specifications (initial objects are unique up to isomorphisms), this selects a *unique* model (cf. [14]), denoted by $P^* = \mathit{ME}(\langle \Sigma, R \rangle)$, as the desired representation for the specification $\langle \Sigma, R \rangle$ which can be used for inferences.

The rationale behind the ME-approach can be described informally as follows: Maximizing entropy in $Mod_{\mathcal{C}}(\langle \Sigma, R \rangle)$ means to permit as much indeterminateness as possible, so that R be represented most faithfully, without external knowledge being added. In this way, the (incomplete) knowledge given by R is completed in an information-theoretically optimal way. More formal justifications of the ME-method are also available: Paris [32] investigates several inductive representation techniques and proves that the ME-principle yields the only method to represent incomplete knowledge in an unbiased way, satisfying a set of postulates describing sound commonsense reasoning. Shore and Johnson [36] also state a list of axioms characterizing the ME-principle as an optimal method to process probabilistic knowledge. Kern-Isberner [22] proves it to be most adequate to handle complex conditional interactions.

4 The inverse representation problem

In system and software development, one generally starts with a set of requirements that have to be specified, further refined, revised, implemented, etc., until one arrives at a model that (hopefully) meets all the requirements. Using specifications along this way, various instances of the representation problem will arise.

For probabilistic conditionals, there is also another line of development. Given (possibly large) sets of data, statistical information can be generated from it, giving us a probability distribution and thus a model. In the area of knowledge discovery in data (KDD) a principal aim is to find a set of peculiarly interesting or *relevant* rules that hold in the given data and may thus be taken as a representation for it (see e.g. [3]). In classical KDD tasks, the relevance of rules is usually measured by some statistical criteria, like e.g. high probabilities, support, or chi-squared values (cf. [4]). In mathematical contexts, one often seeks for a minimal set of sentences (or *axioms*) describing a (set of) preferred model(s). When axiomatizing the essential properties of a concrete data type in an ADT specification, one might aim at a set of equations that is confluent and terminating when interpreted as a set of rewrite rules, while minimality of the set of equations is not important. A further characterization of relevant sentences is to require syntactical sim-

plicity: a single-headed conditional with only one literal in its conclusion is likely to be more interesting than a conditional containing a complex formula in its conclusion. So, in general, the notion of relevance depends heavily on the corresponding application and is – at least in KDD – often frequency-based.

From a more abstract point of view, however, the problem of discovering relevant relationships (in data or in models) can be seen as the problem to compute a set of formulas which represents a given model (or a given set of models) according to some (inductive) representation method. Therefore, we propose to call this the *inverse representation problem*:

> Given a set of Σ-models $M \subseteq Mod(\langle \Sigma, F \rangle)$, find a set of (relevant) sentences F such that the specification $\langle \Sigma, F \rangle$ has M as its desired representation.

As before, M may be a singleton or an arbitrary subset of $Mod(\langle \Sigma, F \rangle)$.

Note that by viewing knowledge discovery as an inverse representation problem, the notion of *relevance* is given a meaning which abstracts from purely statistical aspects: relevance here means relevance with respect to a particular representation method. It can be sharpened by combining it with a demand for minimality, in order to find a kind of a *base* for the given model (or given models, respectively), as in mathematical contexts. Alternatively, one can focus on computing rules with a simple syntax to make the discovered knowledge most expressive and clear. So, although the inverse representation problem provides a clear formal frame for knowledge discovering, context-dependent aspects can also be taken into regard.

After having presented the inverse representation problem as a general and formal framework to deal with knowledge discovery tasks, let us now consider again the logic $Inst_{\mathcal{C}}$ of probabilistic conditionals. Here, as described in Section 3, the ME-principle provides an excellent inductive representation method. We now rephrase the inverse representation problem within the ME-framework as follows:

> Given a probability distribution $P \in Mod_{\mathcal{C}}(\Sigma)$, find a set of rules R such that $P \in Mod_{\mathcal{C}}(\langle \Sigma, R \rangle)$ and such that the entropy of P is maximal in $Mod_{\mathcal{C}}(\langle \Sigma, R \rangle)$, i.e. $P = \mathit{ME}(\langle \Sigma, R \rangle)$.

Whereas previously no tool had been known that helps one to find such an ME-optimal set of rules, in [23] a general approach to solve the inverse representation problem was presented which works for ME-representation (see [25]) and related methods. In the following, we will demonstrate how KDD can be seen as an instance of the inverse representation problem in the ME-framework.

5. KDD as an instance of the inverse representation problem

In this section, we briefly sketch the method described in [25], and illustrate its functionality. This method can be used to compute a concise set of probabilistic conditionals R from a given distribution P over a signature Σ such that $P = \mathit{ME}(\langle \Sigma, R \rangle)$. The approach differs from usual knowledge discovery and data mining methods in that it takes explicitly inductive representation into consideration. It is not based on observing conditional independencies, but aims at learning relevant conditional dependencies in a non-heuristic way. As a further novelty, the method does not compute single, isolated rules, but yields as a result a set of rules while taking into account highly complex interactions of rules.

As a first step to ensure that most informative rules are found, we will make a syntactic restriction and concentrate on *single-elementary conditionals*, i.e. conditionals whose antecedents are conjunctions of literals, and whose consequents consist of a single literal. These conditionals – sometimes also called *association rules* – are often found to be particularly interesting and informative [2].

The basic idea is to exploit numerical relationships found in P as manifestations of interactions of underlying conditional knowledge. To make these interactions transparent and computable, so-called *conditional structures* are calculated for each complete conjunction, reflecting the effect (positive, negative, or neutral) each conditional has on the respective complete conjunction. Further on, a group theoretical framework is developed to be able to match products of probabilities with products of conditional structures. In this way, a crucial connection between numerical and structural information is established which can be used to extract sets of relevant rules from statistical data. Roughly, the corresponding algorithm takes the following steps (for full details of the algorithm, we refer to [24]):

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

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

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

Now, the point that makes this approach applicable to the ME-methodology is that the ME-principle complies with the algebraic theory of conditional structures – it satisfies the so-called *principle of conditional preservation* [22], which proves to be a powerful and effective guideline for inductive knowledge representation and belief revision (cf. [27]). Therefore, it can be assumed that numerical relationships observed in P actually correspond to interactions of conditionals which are ME-represented in P.

In the following example, we will illustrate this abstract description by computing ME-generating rules from the probability distribution shown above.

Example 6 Let $\Sigma = \{a, b, c\}$ be the propositional signature introduced in Example 3, i.e. with the three propositional variables a - *being a student*, b - *being young*, and c - *being unmarried*. Further, let $P \in Mod_{\mathcal{C}}(\Sigma)$ be the distribution given in Example 3 which we repeat here for convenience:

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

Starting with observing relationships between probabilities like

$$P(\overline{a}b\overline{c}) = P(\overline{a}\overline{b}\overline{c})$$
$$\frac{P(abc)}{P(ab\overline{c})} = \frac{P(\overline{a}bc)}{P(\overline{a}b\overline{c})}$$
$$\frac{P(a\overline{b}c)}{P(a\overline{b}\overline{c})} = \frac{P(\overline{a}\overline{b}c)}{P(\overline{a}\overline{b}\overline{c})}$$

the procedure described in [25] yields the set

$$\mathcal{S} = \{(a|\top), (c|\top), (b|a), (b|c)\}$$

of *structural conditionals* not yet having assigned any probabilities to them. Associating the proper probabilities (which are directly computable from P) with these structural conditionals, we obtain

$$\begin{aligned} \mathcal{S}^* \quad = \quad & \{\, (a|\top)[0.4635], \\ & (c|\top)[0.4967], \\ & (b|a)[0.8], \\ & (b|c)[0.7] \,\} \end{aligned}$$

as an ME-generating set for P, i.e. $P = \mathit{ME}(\langle \Sigma, \mathcal{S}^* \rangle)$.

That means, that these four probabilistic conditionals represent P with respect to the *ME*-method. In other words, the probabilistic conditionals

$$\begin{aligned} & (\textit{student}|\top)[0.4635], \\ & (\textit{unmarried}|\top)[0.4967], \\ & (\textit{young}|\textit{student})[0.8], \\ & (\textit{young}|\textit{unmarried})[0.7] \end{aligned}$$

that have been generated from P fully automatically, constitute a consise set of uncertain rules that faithfully represent the complete distribution P in an information-theoretically optimal way. □

This academic example may suffice to illustrate the functionality of the method. In real life applications, several modifications have to be made to cope with the complexity of large data bases. For instance, in order to obtain compact sets of expressive rules, one would prefer to take orders of magnitudes of probabilities into account, instead of working on exact probability values. An interesting application of the theory of conditional structures is that it provides an elegant way to handle the countless number of empty cells in sparse contingency tables appropriately. Basically, the problem with these empty cells is that they should not be interpreted as elementary events with probability 0, since this would amount to introduce knowledge where there is none. In our approach, it is possible to take them as what they are – lacking information. Moreover, it is possible to further tune the general approach to knowledge discovery sketched at the beginning of this section, to the ME-methodology in particular. Doing so, we would find that in the above example, the first two of the discovered rules are redundant – actually, the probabilistic rules

$$\begin{aligned} & (\textit{young}|\textit{student})[0.8], \\ & (\textit{young}|\textit{unmarried})[0.7] \end{aligned}$$

are enough to represent P via the ME-principle.

These techniques, as well as practical details and experiences with applications will be described in more detail in a forthcoming paper.

6 Conclusions and further work

In this paper, by using the abstract logical concept of institutions we described how inductive representation of and reasoning with conditional knowledge can be looked upon as an instance of quite a general representation problem. Moreover, we showed that the crucial problem of discovering relevant conditional relationships in statistical data can also be addressed in this formal framework, namely, by considering knowledge discovery as an operation which is inverse to inductive knowledge representation. This gave rise to phrasing the inverse representation problem.

In order to exemplify our ideas, we illustrated the functionality of an approach to compute sets of conditionals from statistical data, which are optimal with respect to the information-theoretical principle of maximum entropy. This connection between formal logical work and practical uncertain reasoning is part of our CONDOR project (supported by the German science foundation DFG) (cf. [8]) and will be continued there.

Some theoretical aspects only touched on in this paper are discussed in more detail in [5]. Investigations concerning the connection between the institutions of propositional and probabilistic (conditional) logic can be found in [6]. In [7], formal relationships between probabilistic and qualitative conditionals are dealt with. The elaboration of our approach in ordinal frameworks, such as provided by, e.g., possibility theory [28], has been begun in [26], and is a topic of our ongoing research.

Acknowledgements: We thank the anonymous referees of this paper for their helpful comments. The research reported here was partially supported by the DFG – Deutsche Forschungsgemeinschaft within the CONDOR-project under grant BE 1700/5-1.

References

[1] E. Adams. Probability and the logic of conditionals. In J. Hintikka and P. Suppes, editors, *Aspects of inductive logic*, pages 265–316. North-Holland, Amsterdam, 1966.

[2] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, pages 207–216, Washington, DC., 1993.

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

[4] R. J. Bayardo and R. Agrawal. Mining the most interesting rules. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 145–154, San Diego, CA, August 1999. ACM.

[5] C. Beierle and G. Kern-Isberner. Footprints of conditionals. In D. Hutter and W. Stephan, editors, *Festschrift in Honor of Jörg H. Siekmann*. Springer-Verlag, Berlin, Heidelberg, New York, 2002. (to appear).

[6] C. Beierle and G. Kern-Isberner. Looking at probabilistic conditionals from an institutional point of view. In *Proceedings Workshop on Conditionals, Information, and Inference*. Hagen, May 2002.

[7] C. Beierle and G. Kern-Isberner. Using institutions for the study of qualitative and quantitative conditional logics. In *Proceedings of the 8th European Conference on Logics in Artificial Intelligence, JELIA'02*. Springer, LNCS Vol. 2424, 2002.

[8] C. Beierle and G. Kern-Isberner. Modelling conditional knowledge discovery and belief revision by Abstract State Machines. In *International Workshop on Abstract State Machines - ASM2003*, Lecture Notes in Computer Science, Berlin, Heidelberg, New York, 2003. Springer-Verlag. (to appear).

[9] C. Beierle and A. Voss. Viewing implementations as an institution. In D. Pitt, A. Poigné, and D. Rydeheard, editors, *Category Theory and Computer Science*, volume 283 of *Lecture Notes in Computer Science*, Berlin, Heidelberg, New York, 1987. Springer-Verlag.

[10] C. Beierle and A. Voss. Stepwise software development: Combining axiomatic and algorithmic approaches in algebraic specifications. *Technology and Science of Informatics*, 10(1):35–51, January 1991.

[11] B. Buchanan and E. Shortliffe. *Rule-based expert systems. The MYCIN experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, MA, 1984.

[12] R. Burstall and J. Goguen. The semantics of Clear, a specification language. In *Proceedings of the 1979 Copenhagen Winterschool on Abstract Software Specification*, volume 86 of *LNCS*, pages 292–332, Berlin, Heidelberg, New York, 1980. Springer-Verlag.

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

[14] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Prob.*, 3:146–158, 1975.

[15] H. Ehrig and B. Mahr. *Fundamentals of Algebraic Specification 1 - Equations and Initial Semantics*. EATCS Monographs on Theoretical Computer Science. Volume 6, Springer-Verlag, Berlin, Heidelberg, New York, 1985.

[16] U. Fayyad and R. Uthurusamy. Evolving data mining into solutions for insights. *Communications of the ACM*, 45(8):28–61, 2002.

[17] U. Fayyad, R. Uthurusamy, et al. Data mining and knowledge discovery in databases. *Communications of the ACM*, 39(11):24–64, 1996.

[18] J. Goguen and R. Burstall. Institutions: Abstract model theory for specification and programming. *Journal of the ACM*, 39(1):95–146, January 1992.

[19] J. Goguen and W. Tracz. An implementation-oriented semantics for module composition. In G. Leavens and M. Sitaraman, editors, *Foundations of Component-based Systems*, pages 231–263. Cambridge, 2000.

[20] J. A. Goguen and G. Rosu. Institution morphisms. *Formal Aspects of Computing*, 13(3–5):274–307, 2002.
[21] H. Herrlich and G. E. Strecker. *Category theory*. Allyn and Bacon, Boston, 1973.
[22] G. Kern-Isberner. Characterizing the principle of minimum cross-entropy within a conditional-logical framework. *Artificial Intelligence*, 98:169–208, 1998.
[23] G. Kern-Isberner. Solving the inverse representation problem. In *Proceedings 14th European Conference on Artificial Intelligence, ECAI'2000*, pages 581–585, Berlin, 2000. IOS Press.
[24] G. Kern-Isberner. *Conditionals in nonmonotonic reasoning and belief revision*. Springer, Lecture Notes in Artificial Intelligence LNAI 2087, 2001.
[25] G. Kern-Isberner. Discovering most informative rules from data. In *Proceedings International Conference on Intelligent Agents, Web Technologies and Internet Commerce, IAWTIC'2001*, 2001.
[26] G. Kern-Isberner. Representing and learning conditional information in possibility theory. In *Proceedings 7th Fuzzy Days, Dortmund, Germany*, pages 194–217. Springer LNCS 2206, 2001.
[27] G. Kern-Isberner. The principle of conditional preservation in belief revision. In *Proceedings of the Second International Symposium on Foundations of Information and Knowledge Systems, FoIKS 2002*, pages 105–129. Springer LNCS 2284, 2002.
[28] G. Kern-Isberner. A structural approach to default reasoning. In *Proceedings of the Eighth International Conference on Principles of Knowledge Representation and Reasoning, KR'2002*, pages 147–157, San Francisco, Ca., 2002. Morgan Kaufmann.
[29] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.
[30] S. Mac Lane. *Categories for the Working Mathematician*. Springer-Verlag, New York, 1972.
[31] D. Makinson. General theory of cumulative inference. In M. Reinfrank et al., editors, *Non-monotonic Reasoning*, pages 1–18. Springer Lecture Notes on Artificial Intelligence 346, Berlin, 1989.
[32] J. Paris. *The uncertain reasoner's companion – A mathematical perspective*. Cambridge University Press, 1994.
[33] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, Ca., 1988.
[34] D. Sannella and A. Tarlecki. Essential comcepts for algebraic specification and program development. *Formal Aspects of Computing*, 9:229–269, 1997.
[35] Y. Shoham. A semantical approach to non-monotonic logics. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence, IJCAI'87*, 1987.
[36] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, IT-26:26–37, 1980.
[37] A. Tarlecki. Moving between logical systems. In M. Haveraaen, O. Owe, and O.-J. Dahl, editors, *Recent Trends in Data Type Specifications*, volume 1130 of *Lecture Notes in Computer Science*, pages 478–502, Berlin, Heidelberg, New York, 1996. Springer-Verlag.
[38] M. Wirsing. Algebraic specification. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, volume B, pages 675–788. Elsevier Science Publishers B.V., 1990.