Mining Default Rules from Statistical Data

Gabriele Kern-Isberner and **Matthias Thimm**
Department of Computer Science,
Technische Universität Dortmund,
Germany

Marc Finthammer and **Jens Fisseler**
Department of Computer Science,
FernUniversität in Hagen,
Germany

Abstract

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,* (or ranking functions) as a qualitative abstraction of probability functions. The basic idea for transforming probabilities into ordinal rankings is to find well-behaved clusterings of the negative logarithms of the probabilities. We show how popular clustering tools can be used for this, and propose measures for the evaluation of the clustering results in this context. From the so obtained ranking functions, we extract conditionals that may serve as a base for inductive default reasoning.

Introduction

Knowledge discovery and data mining is an area of very active and successful research in the machine learning community. It seems to be common knowledge that this has to be done in a probabilistic context by making use of statistical information obtained from terabytes of data. Hence, there is no well-understood connection to the area of knowledge representation and commonsense reasoning that usually does not make use of data and is more interested in qualitative plausible reasoning. However, to overcome the bottleneck of knowledge acquisition, using present data in an appropriate way seems to be a tempting option; on the other hand, humans do not feel at ease when they have to base important decisions just on numbers, usually, they need some qualitative justification to support a decision convincingly. To date, the links between qualitative and probabilistic information have not been elaborated and used in a systematic way, although some work has been done in a similar direction in the possibility theory community (Benferhat et al. 2003; Borgelt and Kruse 1997).

In this paper, we use *ordinal conditional functions, OCFs,* (or ranking functions) as introduced by Spohn (Spohn 1988), as qualitative abstractions of probability functions. The main contribution of this paper consists of two parts. First, we introduce a mechanism to derive from an empirically obtained probability distribution a more coarse-grained ranking function that subsumes the probability distribution in a more qualitative manner. This abstraction is inspired by the work on infinitesimal probabilities reported in (Adams 1966; Goldszmidt and Pearl 1996). Many different derivations are possible here, so we have to make clear how appropriate rankings can be obtained. We show how clustering techniques (Hartigan 1975) can be used for this purpose. Clusters are to collect probabilities that are "similar enough" so that they can be regarded to be of the same order of magnitude and hence induce the same qualitative information. The obtained ranking functions are used for the application of the CONDORCKD algorithm, which has been developed and implemented in a fully probabilistic framework in (Kern-Isberner and Fisseler 2004) and has recently been adapted for the task of qualitative knowledge discovery on ranking functions (Kern-Isberner, Thimm, and Finthammer 2008). This is possible only due to an inherent connection between ranking functions and probability functions which can both be subsumed by a more general concept (Kern-Isberner 2001a). Second, we conducted several experiments using different settings for the clustering algorithm obtaining different qualitative representations which are used as input for CONDORCKD. The conditionals discovered by CONDORCKD from the data represent plausible relationships and can be used as default rules for commonsense reasoning, by applying one of the well-known nonmonotonic inference formalisms (cf. e.g. (Goldszmidt and Pearl 1996)).

This paper is organized as follows. We start with some preliminaries on conditionals and ranking functions and afterwards introduce the notion of c-representations as a means for default reasoning with ranking functions. Then we present our parametrized approach to obtain qualitative information from a(n empirical) probability distribution. Consequently, we show how clustering techniques can be applied to guide the search for good parameters to obtain suitable ranking functions which are used as input to our mining system CONDORCKD. We continue with reporting on experiments with different parameters and compare the results. Finally, we conclude with a summary and an outlook on further work.

Conditionals and Ranking Functions

We are working with a propositional language $\mathcal{L}$ over a finite set $\mathcal{V} = \{V_1, V_2, \ldots\}$ of propositional variables V_i with finite domains. For each variable $V_i \in \mathcal{V}$, the values are denoted by v_i. Expressions of the form $V_i = v_i$ are called *literals* and are abbreviated by just 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. We will write AB for $A \wedge B$ 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. We give semantics for conditionals using a notion of acceptance of conditionals. Basically, for a conditional $(B|A)$ to be accepted, its confirmation, AB, must be more probable, plausible etc. than its refutation, $A\overline{B}$. Conditionals can be annotated with quantitative values, e. g. probabilities or ranking values, to specify the strength with which they are believed. 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 (Spohn 1988) 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. To keep technical details easy, we will only consider finitely valued OCFs in this paper. For formulas $A \in \mathcal{L}$, a ranking is computed via

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

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

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

Notice, that $\kappa(\overline{A}) > 0$ implies $\kappa(A) = 0$. An OCF κ is extended to conditionals by setting

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

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 $\kappa(A\overline{B})$:

$$\begin{aligned} \kappa \models (B|A)[m] \quad & \text{iff} \quad \kappa(AB) + m = \kappa(A\overline{B}) \\ & \text{iff} \quad \kappa(\overline{B}|A) = m, \; m \in \mathbb{N}, m \geq 1. \end{aligned}$$

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

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

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.

Default Reasoning with OCFs

In this paper, we focus on qualitative inductive reasoning that is based on so-called *c-representations* which have been introduced in (Kern-Isberner 2000; 2001a); many properties, proofs, and lots of examples can be found in (Kern-Isberner 2001a). This approach follows the same structural lines as probabilistic reasoning under maximum entropy and provides the techniques for model-based inductive reasoning in a qualitative environment the quality of which outperforms *system Z* clearly (Kern-Isberner 2001b). Due to space restrictions, we give only a short and mostly informal overview here.

First, an indicator function σ is defined to represent the effects of a conditional on possible worlds by associating to each conditional $(B_i|A_i)$ in $\mathcal{R} = \{(B_1|A_1), \ldots, (B_n|A_n)\} \subseteq (\mathcal{L}|\mathcal{L})$ 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:

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

Here, 1 is 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 $\sigma_{\mathcal{R}} : \Omega \to \mathfrak{F}_{\mathcal{R}}$, defined by

$$\sigma_{\mathcal{R}}(\omega) = \prod_{1 \leq i \leq n} \sigma_i(\omega) = \prod_{\substack{1 \leq i \leq n \\ \omega \models A_i B_i}} \mathbf{a}_i^+ \prod_{\substack{1 \leq i \leq n \\ \omega \models A_i \overline{B}_i}} \mathbf{a}_i^- \tag{2}$$

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

Using this representation we can identify possible worlds, or more generally, combinations of possible worlds on which the conditionals in $\mathcal{R}$ have equal effects. Then an OCF κ that assigns equal values to these worlds, or combinations of worlds, is called *conditionally indifferent with* *respect to* $\mathcal{R}$. The next theorem characterizes indifferent ordinal conditional functions:

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

$$\kappa(\omega) = \kappa_0 + \sum_{\substack{1 \leq i \leq n \\ \omega \models A_i B_i}} \kappa_i^+ + \sum_{\substack{1 \leq i \leq n \\ \omega \models A_i \overline{B_i}}} \kappa_i^- . \qquad (3)$$

Now, in order to obtain a proper representation of a set of conditionals $\mathcal{R}$, we can use the schema (3) and impose the constraints induced by the conditionals in $\mathcal{R}$.

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

In an analogous way, c-representations for quantified OCF-conditionals can be defined. However, different from the maximum entropy principle in the probabilistic case, ordinal 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 will be derived from the given empirical distribution.

Rankings as Qualitative Probabilities

Let P be a probability distribution over $\mathcal{V}$ that could have been collected via a statistical survey. 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. This can be done using the approach discussed in (Kern-Isberner, Thimm, and Finthammer 2008). We start by representing a probability of a specific world ω as a polynomial in a fixed base value ε in the spirit of (Goldszmidt and Pearl 1996). 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 3. 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 ,

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

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

Due to the restriction $0 \leq a_i < \varepsilon^{-1}$ the above definition is sound and uniquely determines a base representation $P_\varepsilon(\omega)$ for given $P(\omega)$ and ε with $P_\varepsilon(\omega) = P(\omega)$. The above definition is quite similar to the one proposed in (Goldszmidt and Pearl 1996) but unlike there we use positive coefficients for the base representation, which seems more natural for our intentions, see (Kern-Isberner, Thimm, and Finthammer 2008).

We use 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 $\tilde{\kappa}_\varepsilon^P$. 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

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

be the base representation of $P(\omega)$ according to Definition 3. 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 that

$$P(\omega) \geq \varepsilon^i \quad \text{iff} \quad \frac{\log P(\omega)}{\log \varepsilon} \leq i \quad .$$

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

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

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.

Furthermore, we can state an approximated probability based on the ranking values 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:

$$P(\omega) \approx \varepsilon^{\kappa^P(\omega)} \qquad (5)$$

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 m:

$$P(B|A) \approx \frac{1}{1 + \frac{\varepsilon^{\kappa^P(A\overline{B})}}{\varepsilon^{\kappa^P(AB)}}} = \frac{1}{1 + \varepsilon^m} \qquad (6)$$

The 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.

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. We will investigate the determination of a suitable value of ε using heuristics and clustering techniques in the next sections.

Mining Default Rules

In order to extract qualitative information from a probability distribution P obtained from statistical data, we have to compute ranking values from the probabilities, as has been described in the previous section. Once a ranking function κ has been derived, a qualitative variation of the CONDORCKD algorithm (Kern-Isberner and Fisseler 2004; Kern-Isberner, Thimm, and Finthammer 2008) can be used to discover a set $\mathcal{R}$ of quantified or unquantified default rules in the form of conditionals such that κ is a c-representation of $\mathcal{R}$ (see Definition 2). Hence, $\mathcal{R}$ may serve as a base for inductive default reasoning via c-representations.

Originally, CONDORCKD was implemented to discover probabilistic conditionals from frequency distributions such that the computed set of rules represents most informative information in the sense that applying the principle of maximum entropy to it will generate a probability distribution which approximates the given frequency distribution (Kern-Isberner 2001a). In particular, the computed rules are as concise as possible. This is due to the exploitation of structural information and conditional indifference (see Theorem 1). Since conditional indifference generalizes conditional independence, it allows the finding of correlations that are interesting and expressive but are not necessarily in accordance with a strictly causal interpretation. Since maximum entropy reasoning and c-representations make use of the same structural foundation, the same machinery can be used for knowledge discovery in both frameworks. Qualitative CONDORCKD just has to replace locally probabilities by ranking values and respect some basic differences between OCF and probabilistic reasoning. Note that the core of CONDORCKD – even for probabilities – works on abstract algebraic information and returns a set of unquantified conditionals which can easily be given the proper semantics by computing the respective values directly from the given data, be they statistical or qualitative in nature.

So, the road to extract default rules from ranking functions has already been paved, thanks to work previously done. What still has to be done is to transform statistical information into the ranking values of an appropriate ordinal conditional function. As has been discussed at the end of the previous section, this comes down to finding a proper parameter ε which defines a measure of similarity that is to make probabilities indistinguishable. From equation (4) we see that rankings correspond to equidistant intervals of (negative) logarithmic probabilities, each of which has equal length $\alpha = -\ln \varepsilon$. The logarithmic probabilities inside each interval should be similar enough to give rise to the same qualitative abstraction, while logarithmic probabilities from different intervals should be different enough to be clearly distinguished. This amounts to finding an optimal clustering of the probabilities. However, neither the maximal cluster width α is known in advance, nor is the number of clusters. Moreover, we expect the clusters to fit into an equidistant partitioning of the interval defined by minimal and maximal (negative logarithmic) probability found in the data, and we also would like to have empty clusters (corresponding to empty ordinal layers) represented. We are not aware of any clustering method that satisfies perfectly all these requirements, so we had to use an existing algorithm and modify it according to our needs.

We chose *k-means* (Hartigan 1975) as a suitable clustering technique for our experiments. The k-means algorithm (or Lloyd's algorithm) starts by randomly distributing k so called *centroids* on the event space and assigning each point to its nearest centroid, forming a partitioning. Then, for each partition a new centroid is calculated and the procedure is repeated until a convergence criterium is reached, e. g., the partitioning does not change any more, and the partitions represent clusters. Here, the event space is an interval of the real numbers, so calculating the new centroid of a partition is equal to calculating the mean of the points in the partition.

The idea is to use k-means with equidistant starting centroids for various numbers k and check how well the resulting clustering fits into an equidistant partitioning of the interval $[(\ln p)_{min}, (\ln p)_{max}]$ needed to cover all negative logarithmic probabilities/frequencies, i.e. $(\ln p)_{min} = \min_{\omega: P(\omega) \neq 0} - \ln P(\omega)$ and $(\ln p)_{max} = \max_{\omega: P(\omega) \neq 0} - \ln P(\omega)$. Since we expect the clusters to be roughly of the same size, the estimated width of each cluster correspond to the logarithmic similarity α that is searched for: $\frac{(\ln p)_{max} - (\ln p)_{min}}{k} = \alpha$. In this way, numbers k that lead to clustering results that are close enough to an equidistant partitioning, yield candidates for α. Then, we use such α's to compute (normalized) ranking functions κ_ε^P with $\varepsilon = e^{-\alpha}$, according to (4), and evaluated the quality of the ranking by measuring similarity within each non-empty ranking layer and dissimilarity between each two neighbouring non-empty layers.

To be more precise, we use the following measurement for the evaluation of the ranking derived from the (logarithmic) probabilities: Let C_λ be the cluster of all negative logarithmic probabilities (different from 0) that are mapped to (finite) λ by equation (4) (plus normalization):

$$C_\lambda = \{-\ln P(\omega) \mid \kappa_\varepsilon^P(\omega) = \lambda\}.$$

Let $C_{\lambda_0}, \ldots, C_{\lambda_s}$ with $\lambda_0 = 0 < \lambda_1 < \ldots \lambda_s$ be the non-empty clusters among these clusters. We define the similarity[1] associated with C_{λ_i} as the average distance between two neighbouring members of C_{λ_i}:

$$sim_i = \begin{cases} (\max_{C_{\lambda_i}} - \min_{C_{\lambda_i}})/\#C_{\lambda_i} & \text{if } \#C_{\lambda_i} > 1 \\ \alpha/2 & \text{if } \#C_{\lambda_i} = 1, \end{cases}$$

where $\#C_{\lambda_i} (> 0)$ denotes the cardinality of C_{λ_i}. If C_{λ_i} contains only one value, then sim_i is set to $\alpha/2$, in order to take the estimated width α of the cluster into account. To measure dissimilarity between clusters, let $dist_i^-$ ($dist_i^+$) be the distance between $\min_{C_{\lambda_i}}$ ($\max_{C_{\lambda_i}}$) and the next smaller (greater) negative logarithmic probability. $dist_i^-$ resp. $dist_i^+$ can be interpreted as the distance of C_{λ_i} to its left resp. right neighbour cluster. Let $dissim_i = \min\{dist_i^-, dist_i^+\}$. Then $q_i^{sim} = dissim_i / sim_i$ reflects how well C_{λ_i} is discriminated from its neighbours, compared to its inner structure. The quality of the whole clustering or the ranking, respectively, is measured by $Q^{sim} = \min_i q_i^{sim}$.

Afterwards, default rules are extracted from high quality rankings with the aid of qualitative CONDORCKD. Ranking values to express the strengths of these rules can be computed directly from the ranking if needed. We will illustrate this procedure by describing some experiments in the next section.

Experiments

As statistical input to our experiments, we use a probability distribution on the binary variables *young, student* and *parent*, and the three-valued variable *marital_status* with outcomes s = *single*, m = *married*, and c = *cohabiting*. The distribution is depicted in Figure 1 and can be considered as empirically obtained statistical data for the well-known "Lea Sombe" example from (Sombé 1990).

The experimental setup follows the plan that we described in the previous section. First, we computed the negative logarithms of the probabilities of the distribution, and then applied k-means to these one-dimensional data. We used the YALE machine learner (Mierswa et al. 2006) for this, with various k-values, but each time with equidistant starting centroids. For selected values of k, we calculated normalized ranking functions κ_ε^P with $\varepsilon = e^{-\alpha}$ and $\alpha \approx k^{-1} * (\max_{\omega:P(\omega)\neq 0} - \log P(\omega) - \min_{\omega:P(\omega)\neq 0} - \log P(\omega))$. The YALE output for $k = 14$ and $k = 34$ is shown in Figure 2 where each logarithmic probability is assigned the number of the cluster it belongs to.

Figure 1 shows the resulting rankings for $\alpha = 0.2$ (corresponding to $k = 34$), and $\alpha = 0.5$ (corresponding to $k = 14$). Note that in general the clustering computed by k-means will only approximate the sharply calculated ranks. Both rankings are evaluated by the similarity-dissimilarity-measure $Q^{sim} = \min_i q_i^{sim}$, the respective values q_i^{sim} can

			$\alpha = 0.2$		$\alpha = 0.5$	
ω	P	$-\ln P$	κ	q_i^{sim}	κ	q_i^{sim}
$y\overline{s}\overline{p}\mathtt{s}$	0.1757	1.74	0	3.9	0	1.5
$y\overline{s}p\mathtt{s}$	0.1757	1.74	0		0	
$ys\overline{p}\mathtt{s}$	0.1196	2.12	2	1.9	1	5.9
$\overline{y}\overline{s}\overline{p}\mathtt{s}$	0.0986	2.32	3	1.3	1	
$\overline{y}\overline{s}p\mathtt{s}$	0.0986	2.32	3		1	
$\overline{y}\overline{s}\overline{p}\mathtt{m}$	0.0865	2.45	4	1.3	1	
$\overline{y}\overline{s}p\mathtt{m}$	0.0865	2.45	4		1	
$y\overline{s}\overline{p}\mathtt{c}$	0.0206	3.88	11	103.4	4	103.4
$y\overline{s}p\mathtt{c}$	0.0206	3.88	11		4	
$ysp\mathtt{c}$	0.0204	3.89	11		4	
$ys\overline{p}\mathtt{c}$	0.0141	4.26	13	1.9	5	6.8
$y\overline{s}\overline{p}\mathtt{m}$	0.0117	4.45	14	52.6	5	
$y\overline{s}p\mathtt{m}$	0.0117	4.45	14		5	
$ysp\mathtt{m}$	0.0115	4.46	14		5	
$\overline{y}\overline{s}\overline{p}\mathtt{c}$	0.0082	4.80	16	9.1	6	9.1
$\overline{y}\overline{s}p\mathtt{c}$	0.0082	4.80	16		6	
$ys\overline{p}\mathtt{m}$	0.0080	4.83	16		6	
$\overline{y}sp\mathtt{m}$	0.0074	4.91	16		6	
$\overline{y}s\overline{p}\mathtt{s}$	0.0058	5.15	17	1.3	7	2.3
$\overline{y}s\overline{p}\mathtt{m}$	0.0051	5.28	18	1.3	7	
$ysp\mathtt{s}$	0.0042	5.46	19	1.8	7	
$\overline{y}sp\mathtt{c}$	0.0008	7.19	27	3.7	11	1.5
$\overline{y}s\overline{p}\mathtt{c}$	0.0005	7.56	29	3.7	12	1.5
$\overline{y}sp\mathtt{s}$	0.0002	8.49	34	9.3	13	3.7
minimum quality Q^{sim}				1.3		1.5

Figure 1: Distribution and rankings for $\alpha = 0.2$ and $\alpha = 0.5$

ω	P	$-\ln P$	$k = 14$	$k = 34$
$y\overline{s}\overline{p}\mathtt{s}$	0.1757	1.74	0	0
$y\overline{s}p\mathtt{s}$	0.1757	1.74	0	0
$ys\overline{p}\mathtt{s}$	0.1196	2.12	1	2
$\overline{y}\overline{s}\overline{p}\mathtt{s}$	0.0986	2.32	1	3
$\overline{y}\overline{s}p\mathtt{s}$	0.0986	2.32	1	3
$\overline{y}\overline{s}\overline{p}\mathtt{m}$	0.0865	2.45	1	3
$\overline{y}\overline{s}p\mathtt{m}$	0.0865	2.45	1	3
$y\overline{s}\overline{p}\mathtt{c}$	0.0206	3.88	4	10
$y\overline{s}p\mathtt{c}$	0.0206	3.88	4	10
$ysp\mathtt{c}$	0.0204	3.89	4	11
$ys\overline{p}\mathtt{c}$	0.0141	4.26	5	12
$y\overline{s}\overline{p}\mathtt{m}$	0.0117	4.45	5	13
$y\overline{s}p\mathtt{m}$	0.0117	4.45	5	13
$ysp\mathtt{m}$	0.0115	4.46	5	13
$\overline{y}\overline{s}\overline{p}\mathtt{c}$	0.0082	4.80	6	15
$\overline{y}\overline{s}p\mathtt{c}$	0.0082	4.80	6	15
$ys\overline{p}\mathtt{m}$	0.0080	4.83	6	15
$\overline{y}sp\mathtt{m}$	0.0074	4.91	6	15
$\overline{y}s\overline{p}\mathtt{s}$	0.0058	5.15	7	17
$\overline{y}s\overline{p}\mathtt{m}$	0.0051	5.28	7	17
$ysp\mathtt{s}$	0.0042	5.46	7	18
$\overline{y}sp\mathtt{c}$	0.0008	7.19	11	27
$\overline{y}s\overline{p}\mathtt{c}$	0.0005	7.56	11	28
$\overline{y}sp\mathtt{s}$	0.0002	8.49	13	33

Figure 2: YALE: k-means for $k = 14$ and $k = 34$

[1] Strictly spoken, sim_i is not a similarity but a distance measure, but it reflects similarity and serves our needs better in this form.

$r_1:\quad (\text{young} \mid \text{student} \wedge \text{parent} \wedge \text{marStat}=\texttt{s})$
$r_2:\quad (\text{young} \mid \top)$
$r_3:\quad (\text{young} \mid \neg\text{student})$
$r_4:\quad (\neg\text{student} \mid \neg\text{young})$
$r_5:\quad (\text{marStat}=\texttt{s} \vee \text{marStat}=\texttt{m}) \mid \top$
$r_6:\quad (\text{marStat}=\texttt{c} \vee \text{marStat}=\texttt{m}$
$\qquad\quad \mid \text{young} \wedge \text{student} \wedge \text{parent})$
$r_7:\quad (\text{marStat}=\texttt{c} \vee \text{marStat}=\texttt{m}$
$\qquad\quad \mid \neg\text{young} \wedge \text{student} \wedge \text{parent})$

Figure 3: Discovered default rules $\Delta_{0.5}$ for $\alpha = 0.5$

		$\alpha = 0.2$		$\alpha = 0.5$	
rule	P exact	Rank	P approx.	Rank	P approx.
r_1 :	0.94	15	0.95	6	0.95
r_2 :	0.59	3	0.65	1	0.62
r_3 :	0.52	3	0.65	1	0.62
r_4 :	0.95	13	0.93	5	0.92
r_5 :	0.91	11	0.90	4	0.88
r_6 :	0.89	8	0.83	3	0.82
r_7 :	0.97	18	0.97	7	0.97

Figure 4: Ranks, exact and approx. probabilities for $\Delta_{0.5}$

also be seen from the figure. The assignment $\alpha = 0.5$ induced the best ranking function in our tests, so we used it for qualitative knowledge discovery in the first place. The computed default rules $\Delta_{0.5}$ are listed in Figure 3.

Figure 4 shows the qualitative ranks of the rules, as well as a comparison between approximated probability (via equation (6)) and exact probability. It is obvious that both $\alpha = 0.5$ and $\alpha = 0.2$ yield good approximations of the exact probabilities.

Summary and Conclusion

In this paper, we presented a method to extract qualitative default rules from probability distributions. The set of extracted rules can be used as a base for inductive default reasoning via c-representations and hence is expected to represent core dependencies between variables.

We first transformed the probabilities into ordinal rankings by applying clustering techniques to the logarithmic probabilities. By modifying the probabilistic knowledge discovery tool CONDORCKD appropriately, we computed conditionals from these rankings by elaborating structural information underlying the rankings. The resulting conditionals can be equipped with qualitative ranks in order to show their strength. We also showed how well approximated probabilities can be computed from the ordinal ranks.

The methods and the experimental results described in this paper show a tight connection between qualitative and probabilistic knowledge representation from which either framework can take profit. Ordinal conditional functions once again proved to allow high quality default reasoning very close to the full probabilistic framework; probabilistic results, on the other hand, can be enriched by qualitative justification abstracting from numerical subtleties.

We will pursue this work on the borderline between qualitative and probabilistic knowledge representation by developing a full qualitative version of CONDORCKD and by further optimizing the transformation of probabilities into rankings via clustering.

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

References

Adams, E. 1966. Probability and the logic of conditionals. In Hintikka, J., and Suppes, P., eds., *Aspects of inductive logic*. Amsterdam: North-Holland. 265–316.

Benferhat, S.; Dubois, D.; Lagrue, S.; and Prade, H. 2003. A big-stepped probability approach for discovering default rules. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems(IJUFKS)* 11:1–14.

Borgelt, C., and Kruse, R. 1997. Some experimental results on learning probabilistic and possibilistic networks with different evaluation measures. In *Proceedings First International Joint Conference on Qualitative and Quantitative Practical Reasoning, ECSQARU-FAPR'97*, 71–85.

Goldszmidt, M., and Pearl, J. 1996. Qualitative probabilities for default reasoning, belief revision, and causal modeling. *Artificial Intelligence*.

Hartigan, J. A. 1975. *Clustering Algorithms*. New York, NY, USA: John Wiley & Sons, Inc.

Kern-Isberner, G., and Fisseler, J. 2004. Knowledge discovery by reversing inductive knowledge representation. In *Proc. of the Ninth Int. Conf. on the Principles of Knowledge Representation and Reasoning*, 34–44. AAAI Press.

Kern-Isberner, G.; Thimm, M.; and Finthammer, M. 2008. Qualitative knowledge discovery. In Schewe, K.-D., and Thalheim, B., eds., *3rd International Workshop on Semantics in Data and Knowledge Bases (SDKB)*, volume 4925 of *Lecture Notes in Computer Science*. Springer. 88–113.

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

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

Kern-Isberner, G. 2001b. Handling conditionals adequately in uncertain reasoning. In *Proceedings ECSQARU'01*, 604–615. Springer LNAI 2143.

Mierswa, I.; Wurst, M.; Klinkenberg, R.; Scholz, M.; and Euler, T. 2006. YALE: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*. ACM Press.

Sombé, L. 1990. *Reasoning Under Incomplete Information in Artificial Intelligence*. Wiley & Sons.

Spohn, W. 1988. Ordinal conditional functions: a dynamic theory of epistemic states. In Harper, W., and Skyrms, B., eds., *Causation in Decision, Belief Change, and Statistics*, volume 2. Kluwer Academic Publishers. 105–134.