The Thirty-First International Florida
Artificial Intelligence Research Society Conference (FLAIRS-31)

Comparing Approaches to Qualitative Data Mining

Richard Niland, Christian Eichhorn, Gabriele Kern-Isberner
TU Dortmund University,
Dortmund, NRW, Germany
⟨first name⟩.⟨last name⟩@udo.edu

Abstract

The amount of data generated and collected every day is increasing continuously. With it, the task of sorting and analyzing this vast aggregation of information is of growing importance. In particular, it is of interest to discover interrelations between observations—a process known as *data mining*. Usually, *quantitative* data mining is performed, assigning concrete weights or percentages to the extracted pieces of information, which, albeit easier to process by machines, does not coincide well with the way humans think and reason. The latter is achieved more fittingly via *qualitative* data mining, extracting rules such as "*If it is a bird, it usually flies*" without the need for any modifying numeric quantifiers. In this article, we recall two established approaches to qualitative data mining. Subsequently, we introduce a novel distance measure as well as, built thereupon, a preference relation yielding a context-aware ranking of approaches to data mining. This measure is then used to evaluate and compare the aforementioned approaches to one another. Finally, we show that both approaches produce fundamentally different results, and show by way of example which kinds of applications each one is better suited for.

1 Introduction and Overview

Approaches summarized under the term *Data Mining* are processes that extract knowledge in the form of relationships, patterns, or rules from a set of data samples or observations (see, e.g., (Fayyad, Piatetsky-shapiro, and Smyth 1996)). The majority of data mining approaches are *quantitative* in nature as they generate probabilistic rules in forms such as "The conditional probability of B given A is x." *Qualitative* rules in the form of "If A then usually B", on the other hand, are building blocks of qualitative and semi-quantitative commonsense, or *nonmonotonic*, reasoning. Formalized as conditionals with a trivalent interpretation, they are capable of being used in a context with human reasoners. One reason for this is that human reasoning is sometimes prone to fallacies when dealing with probabilities (e.g., (Bar-Hillel 1980)) while "commonsense reasoning", using default rules without quantifiers, can be more accessible (e.g., (Ragni et al. 2017)). Approaches to qualitative data mining thus are helpful in scenarios where human reasoners and data mining processes are to cooperate. Examples include, but are not limited to, conditional knowledge bases set up by an expert which are to be revised with the data mining results, data mining results that are to be assessed by human experts, or mined rules that are to be used as qualitative and human-readable explanations for the observations. There are only few approaches to qualitative data mining. In this article, we recall two established approaches, namely the *Big-stepped Probabilities* method (BSP for short, (Benferhat et al. 2003)) and the (qualitative variant of the) CONDORCKD approach (Kern-Isberner et al. 2009). We show that, even if the general structure of these approaches is similar and, at a first glance, modular, the internal methods are tailored to either approach and cannot be substituted easily. We further illustrate that, since both approaches generate sets of conditionals of different formats, they cannot be compared by just observing their input-output-behavior. Thus, we compare both approaches semantically based on the epistemic states induced by the resulting conditional rules. To this end, we define a context aware distance measure as well as a preference relation over epistemic states encoded as ordinal conditional functions, using the latter to show in which scenario which of the presented approaches is the better choice. The rest of this paper is structured as follows. We start out with a brief recapitulation of the formal prerequisites to the presented approaches in Sect. 2. Next, we recall the BSP method and the QCKD approach in Sect. 3. In Sect. 4, we introduce a means of comparing qualitative data mining approaches, and apply this method to the two approaches presented in the preceding section. Finally, we conclude with both a concise summary of our findings and an outlook on future work on the subject at hand in Sect. 5.

2 Preliminaries and Basic Techniques

In this paper, we use a propositional language $\mathfrak{L}$ via closure of a set of propositional variables $\Sigma = \{V_1, \ldots, V_m\}$ under conjunction ($\wedge$) and negation ($\neg$). We abbreviate conjunction by juxtaposition, writing AB for $A \wedge B$, and negation by overlining, writing $\overline{A}$ for $\neg A$. Interpretations or *possible worlds* ω are represented as complete conjunctions over literals $\dot{v}_i \in \{v_i, \overline{v}_i\}$ of variables in Σ, the set of all possible worlds is denoted by Ω. The evaluation of a formula A in a world ω ($[\![A]\!]_\omega$), satisfaction and entailment ($\models$), the material implication ($\Rightarrow$), and semantic equivalence ($\equiv$) are defined as usual, as are the symbols for tautology ($\top$) and contradiction ($\bot$).

A conditional $(B|A)$ is a three-valued logical entity with an evaluation (Finetti 1974) $[\![(B|A)]\!]_\omega = true$ iff $\omega \models AB$ (verification), $[\![(B|A)]\!]_\omega = false$ iff $\omega \models A\overline{B}$ (falsification), and $[\![(B|A)]\!]_\omega = undefined$ iff $\omega \models \overline{A}$ (non-applicability). Conditionals $(B|A)$ establish connections between a premise A and a conclusion B, formalizing the defeasible rule "If A then usually B." Conditionals $(\dot{v}|A)$ where $\dot{v}$ is a literal and A is a conjunction of literals are called *single-elementary conditionals*. For a conditional $(B|A)$ we call $A \Rightarrow B$ its *material counterpart*. A finite set of conditionals $\Delta = \{(B_1|A_1), \ldots, (B_n|A_n)\} \subseteq (\mathfrak{L} \mid \mathfrak{L})$ is called a *knowledge base*. A set of conditionals Δ *tolerates* a conditional $(B|A)$ iff there is a world that verifies $(B|A)$ without falsifying any conditional $(B_i|A_i)$ in Δ.

Example 2.1 (Employee Benefits) *Consider a company that offers its employees rank-based benefits. While most are cubicle workers, some have an office of their own, a company car, or even access to the private company jet. We encode these benefits using the set of variables $\Sigma_{office} = \{C, J, O\}$. Hence, possible sets of employee benefits (i.e., possible worlds), are $\Omega_{office} = \{c\,j\,o, c\,j\,\overline{o}, c\,\overline{j}\,o, c\,\overline{j}\,\overline{o}, \overline{c}\,j\,o, \overline{c}\,j\,\overline{o}, \overline{c}\,\overline{j}\,o, \overline{c}\,\overline{j}\,\overline{o}\}$. We formulate the defeasible rule "If a person may use the company jet, they also get a company car," formally $\delta_1 = (c\,|\,j)$. Similarly, someone who does not even have an office likely has no company car nor access to the jet, formally $\delta_2 = (\overline{c}\,\overline{j}\,|\,\overline{o})$. A plausible scenario, $c\,\overline{j}\,o$, that is, an employee with an office and a car but no access to the jet, falsifies none of the above rules. On the other hand, $\overline{c}\,j\,\overline{o}$ falsifies both rules; therefore, it is considered odd for someone with access to a jet and an office to not have access to a car. The company could aggregate these rules regarding its employee benefits in the form of a knowledge base $\Delta_{office} = \{\delta_1, \delta_2\}$.*

Let $\Delta = \{(B_1|A_1), \ldots, (B_n|A_n)\} \subseteq (\mathfrak{L} \mid \mathfrak{L})$ be a knowledge base and let $\mathfrak{F}(\Delta) = \{\mathbf{a}_1^+, \mathbf{a}_1^-, \ldots, \mathbf{a}_n^+, \mathbf{a}_n^-\}$ be a set of abstract symbols (Kern-Isberner 2001). The function $\sigma_{\Delta,i} : \Omega \to \mathfrak{F}(\Delta) \cup \{1\}$ assigns to each $\omega \in \Omega$ a symbol out of $\{\mathbf{a}_i^+, \mathbf{a}_i^-, 1\}$ based on the evaluation of the conditional $(B_i|A_i)$ in ω, such that $\sigma_{\Delta,i}(\omega) = \mathbf{a}_i^+$ iff $[\![(B_i|A_i)]\!]_\omega = true$, $\sigma_{\Delta,i}(\omega) = \mathbf{a}_i^-$ iff $[\![(B_i|A_i)]\!]_\omega = false$, and $\sigma_{\Delta,i}(\omega) = 1$ iff $[\![(B_i|A_i)]\!]_\omega = undefined$. In this way, the set $\mathfrak{F}(\Delta)$ is an indicator set for the verification or falsification of each individual conditional in Δ. Using this set as a set of generators, we define the free Abelian group $(\mathfrak{F}(\Delta), \cdot, 1)$ with the invertible, associative, and commutative connective $\cdot$ and neutral element 1. The *conditional structure* (Kern-Isberner 2001) $\sigma_\Delta : \Omega \to (\mathfrak{F}(\Delta), \cdot, 1)$ of a world ω based on a given knowledge base Δ is the combination of all indicator functions $\sigma_{\Delta,i}$, $1 \le i \le n$, on ω inside this group, that is, $\sigma_\Delta(\omega) = \prod_{i=1}^{n} \sigma_{\Delta,i}(\omega)$.

From the elements of Ω we generate the free Abelian group $\widehat{\Omega} = \langle \omega \mid \omega \in \Omega \rangle$, containing all words $\widehat{\omega} = \prod_{1 \le i \le |\Omega|} \omega_i^{r_i}$ where $\omega_i \in \Omega$ and $r_i \in \mathbb{Z}$. We call a word $\widehat{\omega} \in \widehat{\Omega}$ a *generalized world* iff all its exponents are positive, in other words, iff it corresponds to a multiset of worlds. This way, we can represent elements of $\widehat{\Omega}$ as fractions $\frac{\widehat{\omega}_1}{\widehat{\omega}_2}$ of generalized worlds by grouping together the worlds with positive (negative) exponents to form $\widehat{\omega}_1$ ($\widehat{\omega}_2$), and by omitting any world whose exponent is zero. We overload σ_Δ for generalized worlds $\widehat{\omega} = \frac{\widehat{\omega}_1}{\widehat{\omega}_2} \in \widehat{\Omega}$ (allowing only trivial cancellations) and define

$$\sigma_\Delta(\widehat{\omega}) = \sigma_\Delta\left(\frac{\widehat{\omega}_1}{\widehat{\omega}_2}\right) = \Big(\prod_{\omega \in \widehat{\omega}_1} \sigma_\Delta(\omega)\Big) \Big/ \Big(\prod_{\omega \in \widehat{\omega}_2} \sigma_\Delta(\omega)\Big). \quad (1)$$

Example 2.2 *The knowledge base Δ_{office} given in Example 2.1 induces the conditional structures on Ω shown in Table 1. The structural relationship between the employees who have access to the company jet but do or do not enjoy a company car is formalized by the generalized world $\widehat{\omega} = \frac{c\,j\,o \cdot c\,j\,\overline{o}}{\overline{c}\,j\,o \cdot \overline{c}\,j\,\overline{o}}$ and induces the conditional structure $\sigma_{\Delta_{office}}(\widehat{\omega}) = \frac{\sigma_{\Delta_{office}}(\widehat{\omega}_1)}{\sigma_{\Delta_{office}}(\widehat{\omega}_2)} = \frac{\mathbf{a}_1^+\ 1\ \cdot \mathbf{a}_1^+ \mathbf{a}_2^-}{\mathbf{a}_1^-\ 1\ \cdot \mathbf{a}_1^- \mathbf{a}_2^-} = \frac{\mathbf{a}_1^+ \mathbf{a}_1^+}{\mathbf{a}_1^- \mathbf{a}_1^-}$.*

As trivalent logical entity, the evaluation function $[\![\cdot]\!]_\omega$ is not sufficient to give appropriate semantics to conditionals. Hence, conditionals have to be considered within richer structures such as epistemic states in the sense of (Halpern 2005) which can be represented as total transitive orderings on the sets of worlds. In this paper, we use orderings that originate from plausibility and possibility orderings of worlds.

Ordinal conditional functions (OCF, (Spohn 2012)) $\kappa : \Omega \to \mathbb{N}_0^\infty$ assign an implausibility rank $\kappa(\omega)$ to each world $\omega \in \Omega$, that is, the higher $\kappa(\omega)$ is, the less ω is believed by the agent, with the normalization constraint that there have to be maximally plausible worlds, that is, the pre-image of 0 must not be empty, $\kappa^{-1}(0) \neq \varnothing$. The rank of a formula A is the minimal rank of all worlds satisfying the formula, $\kappa(A) = \min\{\kappa(\omega) \mid \omega \models A\}$, and the rank of a conditional is the rank of its verification normalized by that of the premise, formally $\kappa(B|A) = \kappa(AB) - \kappa(A)$. A ranking function κ *accepts* a conditional $(B|A)$, written $\kappa \models (B|A)$, iff the verification of the conditional is more plausible than its falsification, $\kappa(AB) < \kappa(A\overline{B})$, and κ is *admissible with respect to a conditional knowledge base* $\Delta = \{(B_1|A_1), \ldots, (B_n|A_n)\} \subseteq (\mathfrak{L} \mid \mathfrak{L})$, written $\kappa \models \Delta$ iff κ accepts all conditionals in Δ.

A ranking function κ that is admissible with respect to a knowledge base Δ can be generated inductively, given that Δ is consistent; we here recall the established approach of *System Z* (Pearl 1990). For this, the knowledge base Δ is partitioned algorithmically into inclusion-maximal disjoint ordered subsets $(\Delta_0 \uplus \cdots \uplus \Delta_m) = \Delta$ such that each conditional in a subset is tolerated by all conditionals contained

Table 1: Conditional Structure for the Employee Benefits Example (Example 2.2).

ω	*verif.*	*falsif.*	$\sigma_\Delta(\omega)$	ω	*verif.*	*falsif.*	$\sigma_\Delta(\omega)$
$c\,j\,o$	δ_1		$\mathbf{a}_1^+\ 1$	$\overline{c}\,j\,o$		δ_1	$\mathbf{a}_1^-\ 1$
$c\,j\,\overline{o}$	δ_1	δ_2	$\mathbf{a}_1^+ \mathbf{a}_2^-$	$\overline{c}\,j\,\overline{o}$		δ_1, δ_2	$\mathbf{a}_1^- \mathbf{a}_2^-$
$c\,\overline{j}\,o$			$1\ 1$	$\overline{c}\,\overline{j}\,o$			$1\ 1$
$c\,\overline{j}\,\overline{o}$		δ_2	$1\ \mathbf{a}_2^-$	$\overline{c}\,\overline{j}\,\overline{o}$	δ_2		$1\ \mathbf{a}_2^+$

in the union of its subset and all subsets of a higher order, formally, for all $1 \leq i \leq m$ and for all $(B|A) \in \Delta_i$, there is an $\omega \in \Omega$ s.t. $\omega \models A \wedge B \wedge \bigwedge_{j=i}^{m} \big(\bigwedge_{(D|C) \in \Delta_j} C \Rightarrow D \big)$. Given this partitioning, the rank of a world ω is the highest ordinal of subsets in the partition which contains conditionals falsified by ω, increased by 1, i.e.,

$$\kappa_{\Delta}^{Z}(\omega) = \begin{cases} 0 & \text{iff } \omega \text{ does not falsify conditionals in } \Delta \\ \arg \max\limits_{0 \leq i \leq m} \{\omega \models A\overline{B} \mid (B|A) \in \Delta_i\} + 1 & \text{o/w.} \end{cases}$$

Under *possibility theory* (Dubois and Prade 2015) worlds are ordered with functions $\pi : \Omega \to [0, 1]$ representing the possibility that ω is the real world on a scale from *rejected* (or impossible, $\pi(\omega) = 0$) to *totally possible* ($\pi(\omega) = 0$). Here, the normalizing condition is that there have to be worlds that are totally possible, that is, the pre-image of 1 must not be empty, $\pi^{-1}(1) \neq \varnothing$. The possibility of a formula A is the supremum of the possibilities of all worlds satisfying A, $\pi(A) = \sup\{\pi(\omega) \mid \omega \models A\}$. Given a t-norm $\circ$ (usually either the minimum or the product), for conditionals $(B|A)$ it is defined that $\pi(AB) = \pi(B|A) \circ \pi(A)$. We define a possibility distribution to accept a conditional $(B|A)$, written $\pi \models (B|A)$, iff the verification of the conditional is more possible than its falsification, $\pi(AB) > \pi(A\overline{B})$.

The, so to say, classical way of representing epistemic states are *probability measures* $P : \mathfrak{L} \to [0, 1]$ on propositional logic which are defined in the usual way such that $0 \leq P(\omega) \leq 1$, $0 \leq P(A) = \sum_{\omega \models A} P(\omega) < 1$, $P(\bot) = 0$, $P(\top) = 1$, and $P(A) + P(\overline{A}) = 1$. On top of a probability measure we define by $\mathcal{P}$ a function that maps a generalized world to a value in $\mathbb{R}_0^+$ using a probability measure P similarly to (1) such that $\mathcal{P}(\widehat{\omega}) = \mathcal{P}\left(\frac{\widehat{\omega}_1}{\widehat{\omega}_2}\right) = \left(\prod_{\omega \in \widehat{\omega}_1} P(\omega)\right) \cdot \left(\prod_{\omega \in \widehat{\omega}_2} P(\omega)\right)^{-1}$.

3 Qualitative Data Mining

The data mining approaches inspected in this paper can, on an abstract level, be generalized to a degree where they consist of two separate phases: A *clustering* phase in which data points are aggregated by a cluster analysis algorithm so that objects of a common type (with a certain degree of freedom) are grouped together and a *rule mining* phase in which each of these clusters is analyzed for relationships between its contained observations to generate rules from.

The *Big-stepped Probabilities* (BsP) approach (Benferhat et al. 2003) uses *non-linear big-stepped probability distributions* (*nlbsp*) to cluster input data such that the i-th stacked cluster (*stratum*) is more probable than all following strata combined.

Definition 3.1 (Non-linear Big-stepped Probability Distribution (Benferhat et al. 2003)) *Let Ω be the set of possible worlds over $\mathcal{V}$ and let $\mathcal{P} : \Omega \to [0, 1]$ be a probability distribution over Ω. Then, $\mathcal{P}$ is called a* non-linear big-stepped probability distribution *iff there exists a partitioning $\Omega = E_1 \uplus \cdots \uplus E_m$ into strata E_i with rank i, s.t. $|E_i| \geq 1$ for all $i \in \{1, \ldots, m\}$, $|E_m| = 1$ or $\mathcal{P}(E_m) = 0$, and $\mathcal{P}(E_i) > \mathcal{P}(E_j)$ iff $i < j$ for all $i, j \in \{1, \ldots, m\}$.*

Additionally, BsP relies on so-called *common* and *(minimal) discriminating factors:*

Definition 3.2 (Common, Minimal Discriminating Factors (Benferhat et al. 2003)) *Let $\Omega = E_1 \uplus \cdots \uplus E_m$ be a partitioning of the set of possible worlds into strata E_i of rank i s.t. $\mathcal{P}$ is an nlbsp. Then, the set of* common factors *of E_i is the set $comm(E_i) = \{\dot{v} \mid \forall \omega \in E_i : \omega \models \dot{v}\}$. Based on the set of* discriminating factors *of E_i, $disc(E_i) = \{D \subseteq comm(E_i) \mid \bigvee_{\omega \in E_j, j<i} \omega \not\models \bigwedge_{\dot{v} \in D} \dot{v}\}$, the set of* minimal discriminating factors *of E_i is the set $disc^{\downarrow}(E_i) = \{D \in disc(E_i) \mid \nexists D' \in disc(E_i) : D' \subsetneq D\}$.*

The original publication (Benferhat et al. 2003) proposed computing nlbsp's by iteratively merging (sets of) worlds until the resulting partition conforms to Definition 3.1. Given an nlbsp $(E_1, \ldots, E_m)$ for some data set, a rule set Δ^{BsP} can be computed from $comm(E_i)$ and $disc^{\downarrow}(E_i)$ via

$$\Delta^{\text{BsP}} = \left\{ \Big(\bigwedge_{\dot{v} \in comm(E_i) \setminus \{D\}} \dot{v} \mid \bigwedge_{\dot{v}' \in D} \dot{v}' \Big) \right\} \quad (2)$$

where $D \in disc^{\downarrow}(E_i)$, $1 \leq i \leq m$.

Example 3.3 (Employee Benefits) *We extend Example 2.1, assuming that the corporation has 100,000 employees of which most do not receive any benefits. The 2,000 managers are provided with a company car and an office of their own. The 30 members of the board of directors may additionally access the company jet. The three employment tiers form strata of an nlbsp, the common factors being the respective tier's benefits and the discriminating factors listing benefits that set a tier apart from those beneath it. Table 2 lists those strata in more detail. Using Equation* (2)*, we extract the rule $(\overline{c}\,\overline{j}\,\overline{o}|\top)$ from the first stratum, that is, an employee is generally assumed to not receive any particular privileges. From E_2, we extract $(\,c\,\overline{j} \mid o\,)$ and $(\,\overline{j}\,o \mid c\,)$, that is, an employee receiving an office tends to receive a car, and vice versa, but not access to the jet. Finally, from E_3 we extract $(\,c\,o \mid j\,)$, i.e., anyone with access to the jet tends to also get an office and a car. Thus, we end up with the knowledge base $\Delta_{office}^{\text{BsP}} = \{(\overline{c}\,\overline{j}\,\overline{o}|\top), (c\,\overline{j}\,|\,o), (\overline{j}\,o|c), (c\,o|\,j), \}$.*

While the results obtained with BsP in Example 3.3 seem plausible given the data, this is not the case for BsP in general, as shown in Example 3.4.

Example 3.4 (Equivalence) *Consider the distribution in Table 3 for the set $\Sigma_{eq} = \{A, B\}$. A human observer might easily conclude that $a \Leftrightarrow b$ holds in 98% of all cases. The* BsP *approach, however, will attempt to form an nlbsp by combining either of the 49%-worlds with either of the*

Table 2: Strata for Example 3.3.

	Worlds	Freq.	$\mathcal{P}(E_i)$	$comm(E_i)$	$disc^{\downarrow}(E_i)$
E_1	$\overline{c}\,\overline{j}\,\overline{o}$	97970	97.97%	$\{\overline{c}, \overline{j}, \overline{o}\}$	$\{\varnothing\}$
E_2	$c\,\overline{j}\,o$	2000	2.00%	$\{c, \overline{j}, o\}$	$\{\{o\}, \{c\}\}$
E_3	$c\,j\,o$	30	0.03%	$\{c, j, o\}$	$\{\{j\}\}$
E_4	*others*	0	0.00%	$\varnothing$	$\varnothing$

Table 3: Input Distribution for Example 3.4.

ω	Freq.	$\mathcal{P}(\omega)$	ω	Freq.	$\mathcal{P}(\omega)$
$a\,b$	49	49%	$a\,\overline{b}$	1	1%
$\overline{a}\,\overline{b}$	49	49%	$\overline{a}\,b$	1	1%

Table 4: System Z OFC of Examples 3.4 and 3.5.

ω	$\kappa^{\text{BsP}}(\omega)$	$\kappa^{\text{CKD}}(\omega)$	ω	$\kappa^{\text{BsP}}(\omega)$	$\kappa^{\text{CKD}}(\omega)$
$a\,b$	0	0	$\overline{a}\,b$	0	1
$a\,\overline{b}$	0	1	$\overline{a}\,\overline{b}$	0	0

Table 5: System Z OCF of Examples 3.3 and 3.6.

ω	$\kappa^{\text{BsP}}(\omega)$	$\kappa^{\text{CKD}}(\omega)$	ω	$\kappa^{\text{BsP}}(\omega)$	$\kappa^{\text{CKD}}(\omega)$
$c\,j\,o$	2	1	$\overline{c}\,j\,o$	3	2
$c\,j\,\overline{o}$	3	1	$\overline{c}\,j\,\overline{o}$	3	2
$c\,\overline{j}\,o$	1	0	$\overline{c}\,\overline{j}\,o$	2	1
$c\,\overline{j}\,\overline{o}$	2	1	$\overline{c}\,\overline{j}\,\overline{o}$	0	0

1%-worlds. This fails to generate a stratum containing the absolute majority of observations, thus requiring another merge. The two non-merged strata are the only ones with at least one common factor, thus they are merged to form a $50 : 50$-split of observations. These two strata still do not form an nlbsp; merging them results in a single stratum equal to Ω. This, however, is a vacuous nlbsp since $comm(\Omega) = disc^{\downarrow}(\Omega) = \varnothing$. The only rule extractable from Ω is $(\top|\top)$, causing any admissible OFC to assign 0 to any and all worlds in Ω.

The *qualitative* CONDORCKD approach (Kern-Isberner et al. 2009) (QCKD) starts out with an equidistant clustering on the negative logarithms of the relative frequencies of all individual worlds. Unlike BSP, which clusters by common literals first, QCKD attempts to find and group together worlds with "similar enough" probabilities of occurrence.

To extract rules from some input data set, QCKD first finds a set of minimal *null conjunctions* $\mathcal{NC}$, the shortest conjunctions of literals appearing only in unobserved worlds. The world clusters induce a set $\ker_0 \mathcal{P}$ of world products, or *generalized worlds*, such that the multiplied probabilities of these worlds yield a value close enough to 1. The margin of "close enough" is given by the initial clustering of the worlds.

A set of starting rules Δ_0 is generated for each positive literal $v_i \in \Sigma$ such that the (most long) conjunctions of other literals together with v_i do not satisfy any of the null conjunctions.

With some rule set Δ_i, QCKD first determines the conditional structure $\sigma_{\Delta_i}(\Omega)$, then searches for generalized worlds in $\ker_0 \mathcal{P}$ where most but not all conditional effects cancel out. Since world pairs in $\ker_0 \mathcal{P}$ behave (almost) identically, the remaining conditional effects can be assumed to do so, too, permitting to merge two rules (if the generalized world's numerator and denominator both verify or both falsify one of them) or to delete a rule (if the numerator verifies and the denominator falsifies it, or vice versa). This process is repeated iteratively until no more modifications are possible.

Finally, the extracted conditionals are annotated with ranks. For purely qualitative data mining, the concrete values of these ranks are of no importance; however, *negative ranks* are used to negate a rule's conclusion, i.e., a conditional $(\,b\,|\,a\,)$ with a rank $x < 0$ corresponds to $(\,\overline{b}\,|\,a\,)$ with rank $|x|$. This licenses for the discovery of rules with negative conclusions even though the starting rules were limited to positive ones. For more detailed technical information, we refer to (Kern-Isberner et al. 2009).

Example 3.5 (Equivalence Revisited) *Example 3.4 exposed a pitfall for* BSP. *Returning to Example 3.4, we now apply the* QCKD *method, instead. As there are no worlds with a frequency of 0, the set $\mathcal{NC}$ is empty. The starting rules are $\Delta_0^{\text{CKD}} = \{\delta_1 = (\,a \mid b\,), \delta_2 = (\,a \mid \overline{b}\,), \delta_3 = (\,b\,|\,a\,), \delta_4 = (\,b\,|\,\overline{a}\,)\}$. The kernel of $\mathcal{P}$ contains $\widehat{\omega}_1 = \frac{a\,\overline{b}}{\overline{a}\,b}$ and $\widehat{\omega}_2 = \frac{a\,b}{\overline{a}\,\overline{b}}$. We have $\sigma_{\Delta_{eq}^{\text{CKD}}}(\frac{a\,\overline{b}}{\overline{a}\,b}) = \frac{\mathbf{a}_2^+\mathbf{a}_3^-}{\mathbf{a}_1^-\mathbf{a}_4^+}$, allowing for no further reductions, as is the case with $\widehat{\omega}_2$.* QCKD *assigns negative ranks to δ_2 and δ_4, yielding $\Delta_{eq}^{\text{CKD}} = \{\delta_1, \delta_2' = (\overline{a}\,|\,\overline{b}), \delta_3, \delta_4' = (\overline{b}\,|\,\overline{a})\}$, which precisely describes $a \Leftrightarrow b$.*

Example 3.6 (Employee Benefits Revisited) *The handling of Example 3.3 using* QCKD *is slightly more complex. As null conjunctions, we have $\mathcal{NC} = \{\, c\,\overline{o}\,,\ \overline{c}\,o\,,\ \overline{c}\,j\,\overline{o}\,\}$, and the highly different frequencies of the three observed scenarios result in an empty kernel $\ker_0 \mathcal{P}$. We therefore do not iterate over Δ_0 and end up directly with $\Delta_{\mathit{office}}^{\text{CKD}} = \{\delta_1 = (\,c\,|\,j\,), \delta_2 = (\,c\,|\,\overline{j}\,o\,), \delta_3' = (\,\overline{j}\,|\,c\,o\,), \delta_4 = (\,o\,|\,c\,)\}$, with δ_3' originating from the negatively-ranked starting rule $\delta_3 = (\,j\,|\,c\,o)$.*

4 Comparing the Approaches

Our goal is to have a means of comparing BSP with QCKD, ideally one allowing us to prefer one over the other. As the examples from the previous section show, the formats of the mined rule sets are substantially different, as BSP yields conditionals with short premises and longer conclusions while QCKD yields single-elementary conditionals with literals for conclusions and premises of varying length. The material counterparts of the latter are a slight generalization of Horn clauses whereas the prior form far more complex formulas. So, since both differ in the structure of extracted rules, the remaining viable option is to investigate the corresponding epistemic states. Infinitely many possible ranking functions matching the same rule set may exist. Therefore, we focus on the System Z ranking functions (see Table 4 for Examples 3.4 and 3.5, and Table 5 for Examples 3.3 and 3.6), only. This, while again revealing similarities and discrepancies between the approaches, provides us with an ordinal basis of comparison. Table 5 reveals that in this example, QCKD considers managers and basic employees to be equally plausible, even though their frequencies differ by a factor of almost 50, suggesting that QCKD fails to properly reflect the input data distribution. On the other hand, BSP fails to discover coimplicatory relationships between literals (cf. Table 4). Hence, ranking functions alone do not yet offer a viable basis of comparison. To construct a sound preferential framework for approaches to qualitative data mining, we propose the notion of *inherent ranking functions* that act as our ideal or desired epistemic state and allow us to apply a context-aware distance measure.

Definition 4.1 (Inherent Ranking Functions (Niland 2017)) *Let $\mathcal{P}$ be a frequency distribution over the set of worlds Ω. The* inherent ranking functions $\kappa^{\downarrow}, \kappa^{\uparrow} : \Omega \to \mathbb{N}_0^{\infty}$ *are defined as*

- $\kappa^{\downarrow}(\omega) = 0$ *if* $\mathcal{P}(\omega) > 0$ *and* $\mathcal{P}(\omega) = \max_{\omega' \in \Omega} \mathcal{P}(\omega')$*, or* $\mathcal{P}(\omega) = 0$
- $\kappa^{\downarrow}(\omega) = \kappa^{\downarrow}(\omega') + 1$ *iff* $\mathcal{P}(\omega') > \mathcal{P}(\omega) > 0$ *and* $\nexists \omega'' \in \Omega : \mathcal{P}(\omega') > \mathcal{P}(\omega'') > \mathcal{P}(\omega)$
- $\kappa^{\uparrow}(\omega) = \kappa^{\downarrow}(\omega)$ *iff* $\mathcal{P}(\omega) > 0$,
- $\kappa^{\uparrow}(\omega) = 1 + \max_{\mathcal{P}(\omega') > 0} \kappa^{\downarrow}(\omega')$ *if* $\mathcal{P}(\omega) = 0$.

Less formally, ranks are assigned to observed worlds in the inverse order of their frequencies. In this way, $\kappa^{\uparrow}$ treats the raw data as an unbiased, representative sample of adequate size, treating unobserved scenarios as maximally *im*plausible (*closed world assumption*), whereas $\kappa^{\downarrow}$ treats unobserved worlds as maximally plausible, implying that is not clear whether the raw data cover the whole sample space (*open world assumption*).

Assuming either of these inherent ranking functions to be the ideal ranking function κ_0, one computed OCF (and thus one approach) may be favored over another if it differs less from said ideal than the other does. To this end, we define the concepts of context-aware local and global distance, and an induced preferential order over all ranking functions over a shared set of worlds.

Definition 4.2 (Context-aware Local and Global Distance) *Let κ, κ' be ranking functions over a common set of worlds Ω. Then, for any $\omega \in \Omega$, the* local distance from κ to κ' in the context of ω *is* $dist_\omega(\kappa, \kappa') = |\kappa(\omega) - \kappa'(\omega)|$*, and the* global distance from κ to κ' *is* $dist(\kappa, \kappa') = \sum_{\omega \in \Omega} dist_\omega(\kappa, \kappa')$.

Definition 4.3 (Context-aware Preference) *Let $\kappa_0, \kappa_1, \kappa_2$ be ranking functions over a common sets of worlds Ω. Then,* κ_1 is preferred over κ_2 in the context of κ_0*, written $\kappa_1 \preccurlyeq_{\kappa_0} \kappa_2$, iff $dist(\kappa_0, \kappa_1) \leq dist(\kappa_0, \kappa_2)$.*

The global distances of κ^{BsP} and κ^{CKD} to the inherent ranking functions $\kappa^{\downarrow}$ and $\kappa^{\uparrow}$ calculated according to Definition 4.2 for both running examples are given in Table 6, showing that the choice of how to interpret unobserved scenarios may invert the order of preference between the two approaches: For the office example, we prefer the result of BSP (the OCF $\kappa^{\mathrm{BsP}}_{office}$) to the result of QCKD ($\kappa^{\mathrm{CKD}}_{office}$) under closed world assumption (i.e., compared to $\kappa^{\uparrow}$), whilst otherwise (i.e., compared to $\kappa^{\downarrow}$), the preference is reversed. For the equivalence example, this effect does not show up. In other words, whether or not one approach is considered to be better than another strongly depends on both the context

Table 6: Global Distances between Ranking Functions.

from \ to	$\kappa^{\downarrow}_{eq}$	$\kappa^{\uparrow}_{eq}$	from \ to	$\kappa^{\downarrow}_{office}$	$\kappa^{\uparrow}_{office}$
$\kappa^{\mathrm{BsP}}_{eq}$	2	2	$\kappa^{\mathrm{BsP}}_{office}$	13	2
$\kappa^{\mathrm{CKD}}_{eq}$	0	0	$\kappa^{\mathrm{CKD}}_{office}$	9	10

given by the input data and on the intention of the person *using* them; an approach is therefore not judged in general but always with the task at hand in mind.

Since both the BSP and QCKD approaches start by a clustering of worlds followed by some rule extraction method based on said clustering, the idea comes to mind that a recombination of the clustering and extraction methods of both approaches could improve upon their results. However, these stages are heavily tailored to one another within each approach (e.g., the BSP extraction relies on common and discriminating factors, which are completely ignored in the QCKD clustering), yielding worse results in either of the two recombinations, up to the point of empty rule sets.

5 Conclusion

In this paper, we recalled two approaches to qualitative data mining, BSP and CONDORCKD. Both are constructed following the same schema. Still we could show that the approaches yield results in form of conditionals that, by themselves, are incomparable. Thus, to compare both approaches, we concentrated not on the generated conditionals but rather on the epistemic states encoded by these conditionals. To this end, we defined a context-aware distance measure which allowed us to determine the distance of the resulting epistemic states to the inherent ranking functions of the raw data.

On top of this distance, we proposed a context-sensitive preference relation based on both the input data and the choice of how to interpret missing or unobserved scenarios by computing a desired epistemic state from the input and determining which of the approaches extracts a more similar epistemic state. This relation revealed that neither of the approaches is generally superior to the other, but that ranking them is possible under consideration of the concrete application of the approach: When unobserved scenarios are considered to be highly unlikely ($\kappa^{\uparrow}$, closed world assumption), the better choice appears to be to use the big-stepped method, and when unobserved scenarios are not to be penalized in terms of likelihood ($\kappa^{\downarrow}$, open world assumption), the CONDORCKD results more aptly describe the potentially underlying knowledge. This means that, for concrete applications of approaches to qualitative data mining where we can assume that each possible outcome has been observed (but not necessarily with the correct frequencies), that is, that it is highly unlikely that any future event might yield an outcome not represented in the raw data, our results recommend using BSP, whilst if we do not know whether the raw data covers the whole sample space, our results recommend using CONDORCKD. Additional examples and analyses in (Niland 2017) support this observation; future work includes comparing the approaches using examples from the standard set of benchmarks for approaches to data mining. We omitted these in this article due to their increased variable and sample sizes and instead used compact examples to illustrate the methodological differences rather than the statistical details.

Further work includes testing how this relation behaves for different pairs of approaches, different examples, or some other "ideal" epistemic state, along with a formal rather than empiric approach to these questions, as well as more sophisticated distance measures. Recent work (Beierle, Eichhorn, and Kern-Isberner 2017) points up ways of normalizing knowledge bases. This step could be used as post processing of the compared methods. Moreover, it would be interesting to test the presented approaches against "classical" probabilistic methods under this aspect.

Acknowledgements

We thank the anonymous reviewers for their valuable hints that helped us improving the paper. This work was supported by DFG research unit FOR 1513 on "Hybrid Reasoning for Intelligent Systems" and DFG-Grant KI1413/5-1 to Gabriele Kern-Isberner as part of the priority program "New Frameworks of Rationality". R. Niland is supported by FOR 1513, C. Eichhorn is supported by Grant KI1413/5-1.

References

Bar-Hillel, M. 1980. The base-rate fallacy in probability judgments. *Acta Psychologica* 44(3):211–233.

Beierle, C.; Eichhorn, C.; and Kern-Isberner, G. 2017. A Transformation System for Unique Minimal Normal Forms of Conditional Knowledge Bases. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 14th European Conference*, 236–245. Cham: Springer.

Benferhat, S.; Dubois, D.; Lagrue, S.; and Prade, H. 2003. A big-stepped probabilities approach for discovering default rules. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 11(supp01):1–14.

Dubois, D., and Prade, H. 2015. Possibility theory and its applications: Where do we stand? In Kacprzyk, J., and Pedrycz, W., eds., *Springer Handbook of Computational Intelligence*. Berlin: Springer. 31–60.

Fayyad, U.; Piatetsky-shapiro, G.; and Smyth, P. 1996. From Data Mining to Knowledge Discovery in Databases. *AI Magazine* 17:37–54.

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

Halpern, J. Y. 2005. *Reasoning About Uncertainty*. Cambridge, MA, USA: MIT Press.

Kern-Isberner, G.; Thimm, M.; Finthammer, M.; and Fisseler, J. 2009. Mining default rules from statistical data. In *Proceedings of the Twenty-Second International Florida Artificial Intelligence Research Society Conference*.

Kern-Isberner, G. 2001. *Conditionals in Nonmonotonic Reasoning and Belief Revision – Considering Conditionals as Agents*. Number 2087 in Lecture Notes in Computer Science. Berlin, DE: Springer Science+Business Media.

Niland, R. 2017. Qualitative Data Mining with Ranking Functions. Bachelor thesis, TU Dortmund University.

Pearl, J. 1990. System Z: A natural ordering of defaults with tractable applications to nonmonotonic reasoning. In Parikh, R., ed., *Proceedings of the 3rd conference on Theoretical aspects of reasoning about knowledge*, 121–135. San Francisco: Morgan Kaufmann.

Ragni, M.; Eichhorn, C.; Bock, T.; Kern-Isberner, G.; and Tse, A. P. P. 2017. Formal nonmonotonic theories and properties of human defeasible reasoning. *Minds and Machines* Volume 27, Number 1:79–117.

Spohn, W. 2012. *The Laws of Belief: Ranking Theory and Its Philosophical Applications*. Oxford: Oxford University Press.