Probabilistic Logic Rule Learning from Small Imperfect Open-World Datasets using a Dempster-Shafer Theoretic Approach

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the problem of learning *epistemic uncertainty measures* for probabilistic logic rules from *small imperfect datasets*. While Bayesian approaches have had tremendous success in learning probabilistic parameters for rules from complex relational data, we lack good methods for handling small and incomplete datasets, with imprecise and probabilistic data instances containing mutually-dependent attributes, being obtained from multiple heterogeneous sources in the open world where new attributes are introduced ad hoc. We propose a Dempster-Shafer approach to address these challenges.

1 Introduction

There has been a growing interest in combining *logic-based representations* with *probabilistic reasoning mechanisms* and *machine learning techniques* [1]. Under the rubrics of probabilistic logic learning (PLL) [2] and statistical relational learning (SRL) [3] several approaches combine probabilistic mechanisms (e.g., Bayesian Networks, Markov Networks, Stochastic Grammars), with logical representation schemes (propositional logic, first-order logic), and machine learning techniques that allow for automated learning of probabilistic parameters or relational structure from data. Some popular PLL and SRL paradigms include Bayesian Logic Programs (BLP), PRISM, ICL, LPADs, ProbLog, P-Log, CP-Logic, PITA, Markov Logic Networks (MLN), Probabilistic Relational Models (PRM), Bayesian Logic Networks (BLN) and Relational Dependency Networks (RDN).

These approaches adopt an underlying Bayesian probability framework expressed graphically as Bayesian networks or Markov Networks and can learn probabilistic weights from data [4]. The Bayesian setting is an intuitive one that already has a number of off-the-shelf tools to make inferences, learn parameters and compute estimates relatively efficiently. While Bayesian approaches have enjoyed considerable success and shown potential in handling large datasets having a fairly complex relational structure such as NELL [5], there is little work on applying these approaches to learning from *small imperfect open-world datasets*.

Small imperfect open-world datasets can come from multiple distinct heterogeneous sources of varying levels of capability and reliability, producing streams of incomplete, ignorant instances for known and unknown attributes that may be mutually dependent on each other. This type of data is frequently encountered in autonomous agents and perceptual systems that learn normative behavior or must learn in contextually-charged environments. Bayesian techniques are not well-suited in these situations and can lead to non-intuitive results. Instead, we propose an alternative approach based on *Dempster-Shafer* (DS) *Theory of Belief Functions* [6] together with an algorithm for automatically learning belief-theoretic logic rules from small imperfect datasets in open-world contexts.

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

2 Dempster-Shafer Theory Background

DS-Theory is a measure-theoretic mathematical framework that allows for combining pieces of uncertain evidential information to produce degrees of belief for the various events of interest. It has been extensively used in sensor fusion networks, object tracking, and network security [7–9]. In DS-Theory a set of elementary events of interest is called *Frame of Discernment* (FoD). The FoD is a finite set of mutually exclusive events $\Theta = \{\theta_1, ..., \theta_N\}$. The power set of Θ is denoted by $2^\Theta = \{A : A \subseteq \Theta\}$. Each set $A \subseteq \Theta$ has a certain weight, or *mass* associated with it. A *Basic Belief Assignment* (BBA) is a mapping $m_\Theta(\cdot) : 2^\Theta \to [0, 1]$ such that $\sum_{A\subseteq\Theta} m_\Theta(A) = 1$ and $m_\Theta(\emptyset) = 0$. The BBA measures the support assigned to the propositions $A \subseteq \Theta$ only. The subsets of A with non-zero mass are referred to as *focal elements* and comprise the set $\mathcal{F}_\Theta$. The triple $\mathcal{E} = \{\Theta, \mathcal{F}_\Theta, m_\Theta(\cdot)\}$ is called the *Body of Evidence* (BoE). For ease of reading, we sometimes omit $\mathcal{F}_\Theta$ when referencing the BoE. Given a BoE $\{\Theta, \mathcal{F}_\Theta, m_\Theta(\cdot)\}$, the *belief* for a set of hypotheses A is $Bel(A) = \sum_{B\subseteq A} m_\Theta(B)$. This belief function captures the total support that can be committed to A without also committing it to the complement A^c of A. The *plausibility* of A is $Pl(A) = 1 - Bel(A^c)$. Thus, $Pl(A)$ corresponds to the total belief that does not contradict A. The $uncertainty$ interval of A is $[Bel(A), Pl(A)]$, which contains the true probability $P(A)$. In the limit case with no uncertainty, we get $Pl(A) = Bel(A) = P(A)$.

DS-Theory extends Bayesian theory in several ways. First, it allows for assigning probabilistic measures to sets of these hypotheses (not just individual ones), including the set of all hypothesis. This allows DS-Theory to consider ignorant and ambiguous information. Second, DS-theory does not require assuming any prior distributions, which is useful when priors are difficult to justify, as is the case with many open-world sensing and perception tasks. Third, DS-theoretic uncertainty generally refers to epistemic uncertainty and corresponds to beliefs held by agents about the world. However, probability theoretic uncertainty often refers to aleatory uncertainty as it relates to frequency of occurrence, randomness and chance. Bayesian and DS-theories do share many commonalities and DS-theory is often viewed as being a generalization of Bayesian theory.

The history of DS-Theory is not without controversy and has been criticized by some including Judea Pearl [10], who subsequently recalled this criticism [11]. Nevertheless, major strides have been made to address these concerns including approximation algorithms for reducing the time complexity of computation [12], decision theoretic aspects [13], graphical models [14], approaches to resolve conflicts that arose from Dempster's original rule of combination [15], and the development of DS-theoretic logical operators [16, 17].

One recent development in DS-theory, is an evidence filtering strategy that has upgraded Dempster's original rule of combination of evidence to accommodate the inertia of available evidence and address some challenges with respect to conflicting evidence [18]. In particular consider the BoEs $\mathcal{E}_1 = \{\Theta, \mathcal{F}_1, m_1(\cdot)\}$ and $\mathcal{E}_2 = \{\Theta, \mathcal{F}_2, m_2(\cdot)\}$, and a given $A \in \mathcal{F}_2$. The updated belief (from iteration t to $t + 1$) $Bel_{t+1} : 2^\Theta \to [0, 1]$ and the updated plausibility $Pl_{t+1} : 2^\Theta \to [0, 1]$ of an arbitrary proposition $B \subseteq \Theta$ are [1]:

$$Bel(B)^{\mathcal{E}_1}_{t+1} = \mu_t \cdot Bel(B)^{\mathcal{E}_1}_t + \nu_t \cdot Bel(B|A)^{\mathcal{E}_2}_t$$

$$Pl(B)^{\mathcal{E}_1}_{t+1} = \mu_t \cdot Pl(B)^{\mathcal{E}_1}_t + \nu_t \cdot Pl(B|A)^{\mathcal{E}_2}_t$$

where $\mu_t, \nu_t \geq 0, \mu_t + \nu_t = 1$. The conditional in the above equations are Fagin-Halpern conditionals which can be considered an extension of Bayesian conditional notions [19]. That is for a BoE $\mathcal{E} = \{\Theta, \mathcal{F}, m(\cdot)\}, A \subseteq \Theta$ and an arbitrary $B \subseteq \Theta$, the conditional beliefs and plausibility are given by:

$$Bel(B|A)^{\mathcal{E}} = Bel(A \cap B)^{\mathcal{E}} / \left[Bel(A \cap B)^{\mathcal{E}} + Pl(A \setminus B)^{\mathcal{E}}\right]$$

$$Pl(B|A)^{\mathcal{E}} = Pl(A \cap B)^{\mathcal{E}} / \left[Pl(A \cap B)^{\mathcal{E}} + Bel(A \setminus B)^{\mathcal{E}}\right]$$

We build on this and other developments to provide a unified probabilistic logic learning framework, grounded on a Belief-Theoretic approach.

[1] We specify the BoE superscript for $Bel()$ and $Pl()$ as needed to be precise, especially when we are combining two distinct BoEs.

3 Proposed Belief-Theoretic Approach and its Unique Properties

3.1 Rule System

Consider a propositional alphabet $\mathcal{L}$, in which we have all the standard symbols (variables, predicates, functions) and logical connectives. In this alphabet, we define a belief theoretic rule, as follows:

Definition 1 *(Belief-Theoretic Rule System). A belief-theoretic rule is an expression of the form:*

$$\mathcal{R} := [\alpha, \beta] :: \psi \implies (\neg)\phi$$

where the ψ, ϕ are Uncertain Logic *atoms, i.e., propositional formulas with an associated "uncertainty interval" $[\alpha, \beta]$ defined under Dempster-ShaferTheory [20] as $\alpha = Bel(\mathcal{R}), \beta = Pl(\mathcal{R})$ with $0 \leq \alpha \leq \beta \leq 1$. Thus, under the present formulation both the atoms and the rules have uncertainty intervals.*[2] *The $(\neg)$ indicates that negation is optional in this rule. A Belief-Theoretic Rule System $\mathcal{T}$ is a finite set of belief-theoretic rules $\mathcal{R}$.*[3]

Example 1 *Consider an agent reasoning about actions it can perform in a car and in a house. We can represent this scenario as a Belief-Theoretic Rule System, $\mathcal{T}$, as follows:*

$$\begin{array}{ll} \mathcal{R}_1 := [0.8, 0.95] :: inCar \implies driving & \mathcal{R}_4 := [0, 0.3] :: inHouse \implies running \\ \mathcal{R}_2 := [0.9, 1] :: inCar \implies texting & \mathcal{R}_5 := [0.3, 0.6] :: inHouse \implies smoking \\ \mathcal{R}_3 := [0.9, 1] :: inHouse \implies texting & \end{array}$$

The rules in this example have intuitive semantics, whereby the antecedents correspond to context and the consequents correspond to actions taken. The location of the uncertainty interval between 0 and 1 suggests the degree of truth and falsity for the rule and the width of the interval generally suggests the level of support or evidence for that rule. So, rules $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$ have tight uncertainty intervals close to 1 indicating a confident support for their truth.

The rule system displays the agent's current level of belief or epistemic uncertainty about a certain set of rules that are influenced by various pieces of evidence. However, it is typically the case that the agent is not updating its beliefs about all the rules in a rule-system simultaneously, but instead it is considering only a certain subset that might pertain, for example, to a particular context that the agent is in, and for which evidence arrives together. For instance, when an agent is collecting data by observing cars, they may observe multiple actions being performed together: driving, texting, talking etc. In this setting the agent might need to only consider those actions that are relevant in the context at the time of data collection, i.e., those relevant to the "in car" context.

To formalize these intuitions, consider the rule system $\mathcal{T}$ of Example 1. The subset of rules from the rule system relevant to the context $inCar$ is $\{\mathcal{R}_1, \mathcal{R}_2\} \subset \mathcal{T}$. We model the arriving evidence as well as the agent's growing body of beliefs as DS-theoretic frames of discernment $\Theta_{\mathcal{T}}^{inCar}$ comprising all possible combinations of the rule consequents (and negations) present in the selected subset of rules:

$$\Theta_{\mathcal{T}}^{inCar} = \{(texting, driving), (texting, \neg driving), (\neg texting, driving), (\neg texting, \neg driving)\}$$

Modeling the frame of discernment in this exhaustive way ensures that elementary events in the frame are mutually exclusive of each other. The subset of rules associated with this frame is called a Rule frame $\mathcal{R}_{\mathcal{T}}^{inCar} = \{\mathcal{R}_1, \mathcal{R}_2\}$. Now, if we are interested in measuring the amount of support in favor of, say rule $\mathcal{R}_1$, based on our evidence that is captured in the frame $\Theta_{\mathcal{T}}^{inCar}$, we would measure $Bel(\{(texting, driving), (\neg texting, driving)\})$ as this captures the level of support for just $driving$ irrespective of $texting$, from a body of evidence that captures both.

To generalize, a Rule Frame $\mathcal{R}_{\mathcal{T}}^{\psi}$ is a set of rules in rule system $\mathcal{T}$ that share the same antecedent ψ. Similarly, we can generalize the frame of discernment, $\Theta_{\mathcal{T}}^{\psi}$, by first defining a DS-theoretic elementary event θ as a tuple of all the rule consequents ϕ (or their negations) present in a rule frame $\mathcal{R}_{\mathcal{T}}^{\psi}$. A set of elementary events forms an indexed frame of discernment $\Theta_{\mathcal{T}}^{\psi}$. Defining an elementary event in this way allows us to represent, exhaustively, all possible combinations of the set of consequents $\phi_1, \ldots, \phi_k$ and their negations.

[2] The concepts presented in this paper are not limited to a propositional logic and are extendable to a first-order language. The proposed formulation does not preclude the possibility of multiple consequents and antecedents combined with logical operators.

[3] Rules $\mathcal{R}$ in the rule system $\mathcal{T}$ differ from each other either by antecedent and/or consequent.

3.2 Learning Uncertainty Intervals for Rules from Data

Data Format. For a rule system $\mathcal{T}$ with n rules, consider an indexed FoD $\Theta_{\mathcal{T}}^{\psi}$ and a corresponding rule frame $\mathcal{R}_{\mathcal{T}}^{\psi}$ comprising k rules. Consider a set $S = \{s_1, \ldots, s_m\}$ of m evidence sources. Let the set of evidence sources provide a set of BoEs, defined as $\mathbb{E} = \{\mathcal{E}_1^{\Theta_{\mathcal{T}}^{\psi}}, \ldots, \mathcal{E}_m^{\Theta_{\mathcal{T}}^{\psi}}\}$. Each BoE is a DS-theoretic BoE $\mathcal{E}$ and is associated with an indexed FoD $\Theta_{\mathcal{T}}^{\psi}$. For simplicity, we will assume that all the BoEs in $\mathbb{E}$ correspond to the same indexed frame $\Theta_{\mathcal{T}}^{\psi}$. That is, the sources S provide evidence for the same rule frame $\mathcal{R}_{\mathcal{T}}^{\psi}$. We can then define an observation, the set $O_i = \{o_{i,1}, \ldots, o_{i,k}\}$, made by a source s_i as a form of truth assignment where each $o_{i,j} \in \{0, 1, \epsilon\}$ indicates whether the source observes, for a given antecedent ψ, whether a certain consequent $\phi_j, 1 \leq j \leq k$ is true (1), false (0) or unknown (ϵ). For instance, an observation that a person in a car is texting, but not driving can be represented as $O_i = \{1, 0\}$. We can combine the observation O_i with other information about the source as well as a DS-theoretic mass assignment to form a data instance, defined as follows. A data instance is a tuple $d = (s_i, O_i, m_{\Theta_{\mathcal{T}}^{\psi}}(\cdot)_{s_i})$ comprising a specific source identifier $s_i \in S$, an observation O_i, and a DS-theoretic mass assignment $m_{\Theta_{\mathcal{T}}^{\psi}}(\cdot)_{s_i}$ for source s_i per BoE, $\mathcal{E}_i^{\Theta_{\mathcal{T}}^{\psi}}$ provided by that source. A dataset $\mathcal{D} = \{d_1, \ldots, d_n\}$ is a finite set of n data instances.

Learning Problem. The learning problem can be defined as follows: Given an "unspecified" rule frame ${}^{-}\mathcal{R}_{\mathcal{T}}^{\psi}$ ("$-$" suggests that parameter values are unspecified) with k rules, and a dataset $\mathcal{D}$, compute the parameters of the rule frame $\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k$.

Learning Algorithm. The rule learning algorithm (Algorithm 1) assigns uncertainty parameters to each rule, updating those values as it considers each new data instance. Algorithm 1, displayed below, achieves this form of rule learning. The algorithm iterates though each data instance d in the data set $\mathcal{D}$ (line 5) and, per instance, through each rule $\mathcal{R}$ in the rule frame $\mathcal{R}_{\mathcal{T}}^{\psi}$ (line 6). For each iteration, we first set the hyper-parameters μ and ν (line 7) that specify how much weight the algorithm will place on previous learned knowledge (μ) and on each new data instance (ν). These hyper-parameters are then used to compute a conditional belief and plausibility for a rule given that particular instance of data (lines 8,9). The conditional beliefs and probabilities then yield an updated belief and plausibility for each rule (lines 10, 11). Finally, the algorithm updates the uncertainty interval for each rule with the new belief and plausibility values (lines 13,14). The result is a set of belief-theoretic rules (rules accompanied with uncertainty intervals) (lines 16,17).

4 Comparing the Bayesian Approach with the Proposed Approach

To evaluate the proposed approach, we compare it to a Bayesian approach used in many PLL and SRL formulations (e.g, BLP, ProbLog). Consider a Bayesian clause of the form $\mathcal{R}^B := p :: \psi \implies (\neg)\phi$, where ψ, ϕ are Bayesian atoms and p is a point probability estimate. To learn p from data, we can establish a prior distribution and then update the distribution for each instance. We can assume an uninformative prior over each rule, as a uniform distribution: $\mathbf{p} \sim \text{Beta}(1, 1) = \text{Uniform}(0, 1)$. We now suppose that given a specific value of p_j for rule j, each individual source i will provide an observation $o_{i,j}$. We can compute the conditional, as follows: $P(o_{i,j}|p_j) = p_j^{o_{i,j}}(1 - p_j)^{1-o_{i,j}}$. We can then compute the posterior, for n sources, where each source has a reliability measure or mass m_i' as follows:

$$P(p_j|o_{1,j}, \ldots, o_{n,j}) = \prod_{i=1}^{n} \left[p_j^{o_{i,j}}(1 - p_j)^{1-o_{i,j}}\right]^{m_i'}$$

which simplifies to a Beta distribution:

$$\mathbf{p_j}|o_{1,j}, \ldots, o_{n,j} \sim \text{Beta}\left(\sum_{i=1}^{n} m_j' o_{i,j} + 1, \sum_{i=1}^{n} m_j'(1 - o_{i,j}) + 1\right)$$

In this case, the value of p in a Bayesian rule could be sampled from the distribution $\mathbf{p}$. Since we do have a distribution, we can potentially estimate confidence intervals (or credible intervals) to generate measures more akin to the proposed DS-based approach.

Algorithm 1 getParameters($\mathcal{D}, {}^{-}\mathcal{R}_{\mathcal{T}}^{\psi}$)

Input: $\mathcal{D} = \{d_1, \ldots, d_n\}$: Dataset containing n data instances
Input: ${}^{-}\mathcal{R}_{\mathcal{T}}^{\psi}$: An unspecified rule frame containing k rules $\mathcal{R}$
Initialize a DS Frame $\Theta_{\mathcal{T}}^{\psi} = \{\theta_1, \ldots, \theta_{2^k}\}$
$m(\Theta_{\mathcal{T}}^{\psi}) \leftarrow 1$
$t \leftarrow 0$
for all $d \in \mathcal{D}$ **do**
    Let $\mathcal{E}_d$ be a BoE that corresponds to the data instance d
    Let $\mathcal{E}_\Theta$ be a BoE that corresponds to the indexed frame $\Theta_{\mathcal{T}}^{\psi}$
    for all $\mathcal{R} \in^{-} \mathcal{R}_{\mathcal{T}}^{\psi}$ **do**
        Set learning parameters μ_t and ν_t
        $Bel(\mathcal{R}|d)^{\mathcal{E}_d} = Bel(\mathcal{R} \cap d)^{\mathcal{E}_d} / (Bel(\mathcal{R} \cap d)^{\mathcal{E}_d} + Pl(d \setminus \mathcal{R})^{\mathcal{E}_d})$
        $Pl(\mathcal{R}|d)^{\mathcal{E}_d} = Pl(\mathcal{R} \cap d)^{\mathcal{E}_d} / (Pl(\mathcal{R} \cap d)^{\mathcal{E}_d} + Bel(d \setminus \mathcal{R})^{\mathcal{E}_d})$
        $Bel(\mathcal{R})_{t+1}^{\mathcal{E}_\Theta} = \mu_t \cdot Bel(\mathcal{R})_t^{\mathcal{E}_\Theta} + \nu_t \cdot Bel(\mathcal{R}|d)^{\mathcal{E}_d}$
        $Pl(\mathcal{R})_{t+1}^{\mathcal{E}_\Theta} = \mu_t \cdot Pl(\mathcal{R})_t^{\mathcal{E}_\Theta} + \nu_t \cdot Pl(\mathcal{R}|d)^{\mathcal{E}_d}$
    end for
    Set frame $\Theta_{\mathcal{T}}^{\psi}$ with $Bel(\mathcal{R})_{t+1}$ and $Pl(\mathcal{R})_{t+1}$
    $t \leftarrow t + 1$
end for
for all $\mathcal{R} \in \mathcal{R}_{\mathcal{T}}^{\psi}$ **do**
    $\alpha_{\mathcal{R}} \leftarrow Bel(\mathcal{R})^{\mathcal{E}_\Theta}$
    $\beta_{\mathcal{R}} \leftarrow Pl(\mathcal{R})^{\mathcal{E}_\Theta}$
end for
Output: $\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k$

4.1 Dataset and Experimental Conditions

We consider a small sample dataset $\mathcal{D}$ shown in Table 1 containing ten data instances d from several different sources, corresponding to an unspecified and modified version of the rule frame ${}^{-}\mathcal{R}_{\mathcal{T}}^{\psi}$ from rules $\mathcal{R}_1$ and $\mathcal{R}_2$ in Example 1. For the evaluation, we consider four conditions that are variations on the dataset in Table 1: (I) *Reliable-Complete* – the situation when there are no missing values and the agent considers all the sources to be reliable and certain (ϵ are replaced with a 1 or 0, and the masses all equal 1.0); (II) *Reliable-Incomplete* – while masses are still set to 1.0, several values are missing in the data; (III) *Unreliable-Complete* – masses are set between 0 and 1, but there are no missing values; and (IV) *Unreliable-Incomplete* – masses are set between 0 and 1, and there are several missing values. We **hypothesize** that, even for such a simple dataset, the proposed approach will be able to differentiate between these conditions and provide a richer sense for the uncertainty in the data than the competing Bayesian Approach.

Source ID	driving	texting	mass
1	0	1	0.8
2	0	1	0.95
3	1	0	0.65
4	1	0	0.95
5	1	ϵ/0	0.8
6	ϵ/0	0	0.65
7	0	ϵ/1	0.9
8	ϵ/1	0	0.9
9	0	0	0.5
10	ϵ/0	ϵ/1	0.85

Table 1: Experimental Dataset $\mathcal{D}$. The ϵ refers to missing data entries. Variations of experimental conditions involve replacing ϵ with either a 1 or 0 as indicated, and mass values with 1.0.

We set the learning parameters $\nu_t = 1/|dataset|$, $\mu_t = 1 - \nu_t$ in the algorithm to emphasize that new information will be assigned a weight of 0.1, while the inertia of existing knowledge will be assigned a weight of 0.9. To remove order effects of the DS-based updating process, we ran the algorithm through 100 epochs, randomizing the order prior to each run. This allowed us to both reduce order effects and analyze convergence characteristics when compared to the Bayesian approach. On the Bayesian side, we looked at the maximum a posteriori (MAP) estimate and a 95% credible interval computed from the inverse CDF.

5 Results and Discussion

Ignorance, Reliability and Convergence. In this experiment, we set p of the Bayesian rules to the MAP of $\mathbf{p}$. The general limitation with point estimates is that we cannot distinguish between the actual real-valued truth estimate and uncertainty in the truth value. Thus, for a value of 0.5 we do not know the amount of confidence in this estimate. We plot the MAP estimates (red circles) for both rules across the four conditions (Figure 1). Although there is some slight variation across various cases, the MAP estimates do not shed light on the reliability or completeness of the data.

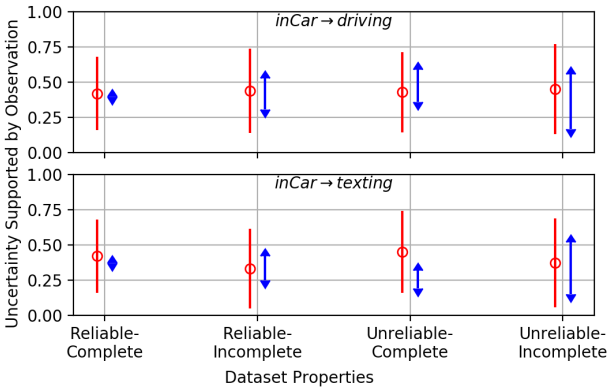

Figure 1: Learned parameters provide richer sense of uncertainty. The proposed approach (blue) exposes more aspects of the epistemic uncertainty associated with small imperfect datasets than a traditional Bayesian approach (red). When the data is reliable and complete, the proposed approach converges to Bayesian. However, when the data displays unreliability or contains missing values, then the proposed approach allows for variable length intervals to express this uncertainty of evidence

We could also extend the definition of a Bayesian Clause to have an interval, corresponding to a credible interval (CI) of $\mathbf{p}$, allowing us to capture some of the richness present in the probability distribution. We computed a 95% CI for the probability distributions for each of the rules. The CI suggests that there is a 95% chance that the calculated confidence interval from some future experiment encompasses the true value of p. It does not, however, suggest that it contains the value of true probability with 95% certainty, whereas the proposed Belief-theoretic uncertainty interval states that the true value exists within the stated interval. Similar to the MAP estimate, the intervals are also not informative, and although there is some widening in the presence of (un)reliability and incompleteness, generally, there is not much variation between the conditions. Moreover, even if the CI is suggestive of uncertainty more broadly, the upper and lower limits of the CI themselves, do not provide any further information about the uncertainty of the rule. The belief-theoretic limits, on the other hand, are well-defined and have specific meaning that pertain to the level of support provided by the evidence. That is, the $Bel()$ (lower limit) specifically represents the measure of evidence supporting a proposition and the $Pl()$ (upper limit) specifically measures evidence that do not contradict the proposition. Thus, conversely, $1 - Bel()$ represents the level of doubt in the evidence for the proposition and $1 - Pl()$ represents the level of disbelief in the evidence. This type of information is not captured in a Bayesian CI.

The proposed approach captures variation in the data (conditions I-IV), while still converging to Bayesian estimates when there is perfect data (condition I). Moreover, in one condition (condition III, $texting$), the MAP estimate lies outside of the belief-theoretic interval suggesting that there is a discrepancy between the different types of uncertainty being captured. We believe that the MAP estimate captures aleatory uncertainty while the belief-theoretic approach captures epistemic uncertainty, and this observed difference results from the selection of a potentially inappropriate prior in the Bayesian approach. Not limited by a prior, the belief-theoretic approach allows for a more dynamic update process and convergence to an estimate of epistemic uncertainty.

One advantageous feature of the proposed approach is that no matter how small the dataset, we can obtain a uncertainty interval based on the evidence received thus far. Although we do not show an instance-by-instance illustration of the algorithm, we can say that the algorithm begins with complete uncertainty $[0, 1]$ and then with each input converges to either a point estimate (as is the case of

condition I of reliable and complete data) or to an interval (as in the cases of conditions II, III, IV). The rate and degree of convergence is also dependent on the selection of learning parameters m, n, which roughly map to a learning rate.

Independence Between Relations. Another desirable feature of the proposed approach is that we can ask a number of other questions of the indexed frame $\Theta^{\psi}_{\mathcal{T}}$ that are not explicitly in the rule system. For example, we might ask what is the uncertainty associated with $(texting, driving)$. In the Bayesian setting, if we assume these two actions are independent, we can multiply point estimates (0.4 and 0.4 for condition I) and generate a non-zero probability of 0.16. In reality, these two actions may not be independent of each other, and therefore it may be improper to make such an assumption. Moreover, there is absolutely no evidence in Table 1 to support this non-zero probability as none of the 10 data instances support both texting as well as driving. In contrast, in the proposed approach, we do not make these assumptions, and instead directly query the same frame that was learned in the experiment thus far. In doing so, we obtain an interval $[0, 0]$. This conforms to our intuitions about the data as well as acceptable traffic norms.

6 Learning in an Open-World

Overall, we are interested in a numerical quantity that represents a degree to which the agent is certain of something, or a degree to which the agent believes it, or a degree to which the evidence supports it [21]. In the open world, measuring this sort of uncertainty requires the ability to process a stream of information from multiple heterogeneous sources, say about a rule like those presented above, and then incorporate and update uncertainty measures on this rule. The challenge is that one information source may be quite different from another source not only in terms of reliability (as discussed earlier) but also its *repertoire of capabilities*. For example, while one source can detect both actions of texting and driving, another might only be able to detect the driving action, while still another source might be able to detect a different action of "talking."

We elaborate this idea by extending the dataset in Table 1 and adding information from three new sources 11, 12 and 13, as shown in Table 2.

Source ID	texting	driving	talking	eating	mass
10	ϵ/0	ϵ/1			0.85
11	0	1	1		0.9
12				1	0.4
13		1	1		0.6
12				ϵ/0	0.4

Table 2: Sample Dataset $\mathcal{D}'$: We extend Table 1 by first re-introducing source 10 and then sequentially incorporating sources 11, 12 and 13 and again source 10, each with different capabilities. Thus, when information from source 11 is received, a new attribute of "talking" must be incorporated into the frame and into the rule frame. When source 12 is processed, a new attribute of "eating" must be incorporated. Note, the agent does not see all these data instances simultaneously, but instead in sequence, which is more common in the open-world.

The first row of Table 2 includes the last entry of Table 1, which shows source 10 as being able to detect $texting$ and $driving$ and has provided observations $\{\epsilon/0, \epsilon/1\}$, accordingly. At this stage, the indexed frame $\Theta^{\psi}{}_{\mathcal{T}}$ is:

$$\Theta^{\psi}_{\mathcal{T}} = \{\theta_1 : (driving, texting), \theta_2 : (driving, \neg texting), \theta_3 : (\neg driving, texting), \theta_4 : (\neg driving, \neg texting)\}$$

Next, the agent receives evidence from source 11, which contains an additional previously unknown attribute $talking$ and provides an observations for all three attributes $texting, driving$, and $talking$. The agent must now expand its indexed frame to incorporate this new previously unseen attribute and generate the following expanded frame ($\Theta^{\psi}{}'_{\mathcal{T}}$):

$$\Theta^{\psi}{}'_{\mathcal{T}} = \{\theta'_1 : (driving, texting, talking), \ldots \theta'_5 : (driving, texting, \neg talking), \ldots\}$$

DS-theory provides some useful notions to determine how $\Theta^{\psi}_{\mathcal{T}}$ relates to $\Theta^{\psi}{}'_{\mathcal{T}}$, which in turn will allows us to dynamically grow and shrink the indexed frame as the agent sees new data. The notion

of "refinement" describes how one frame can be obtained from another by splitting some or all of the elements of the initial frame. The canonical example of this operation is when a frame $\Theta = \{animal, flowers\}$ is split into $\Theta' = \{dog, cat, rose, lily\}$. We can prove that expanding a frame from $\Theta_{\mathcal{T}}^{\psi}$ to $\Theta_{\mathcal{T}}^{\psi'}$ is indeed such a refinement (although we cannot present the full proof here due to space limitations), by leveraging the exhaustive elaboration of consequents ϕ in the elementary events θ in Section 3.1, which in turn allows us to relate, for example $\{(\phi_1)\} \in \Theta$ with the expanded $\{(\phi_1, \phi_2), (\phi_1, \neg\phi_2)\} \in \Theta'$. Showing that the two frames $\Theta_{\mathcal{T}}^{\psi}$ and $\Theta_{\mathcal{T}}^{\psi'}$ are a refinement also lets us establish that they are "compatible" in a DS sense. That means, we can show that the frames agree on the information defined in them, allowing us to prove that for $A \subseteq \Theta, Bel_{\Theta}(A) = Bel_{\Theta'}(\omega(A))$, where ω is a refinement function $\omega : 2^{\Theta} \rightarrow 2^{\Theta'}$ mapping the two frames.

Next, when the agent receives information for a new attribute $eating$ from source 12, the frame is further refined. Because source 12 provides information for *only* $eating$, it doesn't make sense to update the entire refined frame.[4] Fortunately, DS-theory also defines the notion of "coarsening" (inverse of refinement) which allows us to go in the opposite direction from $\Theta_{\mathcal{T}}^{\psi'}$ to $\Theta_{\mathcal{T}}^{\psi}$, an operation that can be performed in $|\Theta| \log |\Theta'|$ time [12]. This ability to coarsen a frame allows us the possibility of coarsening the frame to just one attribute (namely $eating$) and then incorporating the masses assigned by source 12. The agent can then receive information from source 13, which does not induce a refinement because there are no new attributes, ϕ, however, it does induce a coarsening as it does not provide information for all attributes in the currently most-refined frame. Finally, the agent might receive a new data instance from a previously known source, in this case source 12. The task of coarsening the frame for source 12 is made easier this time because the agent already knows the mapping between the frames from the prior computation. Thus, once an agent has encountered a source, it remembers the capabilities of the source, and does not have to recompute these frame mappings.

One of the most exciting aspects of the proposed approach is its ability to account for not just the reliability of the sources, but as we have discussed, their capabilities as well. In doing so, we can expand and grow the set of rules *in real time*, without always having to recompute joint distributions, as we would need to do in a Bayesian approach. We can also update rules more efficiently as updating a rule from a source with limited capacity does not impact existing knowledge about a capacity not captured by the source. The DS-based approach allows us to speed up certain operations, especially when new attributes are added ad-hoc and when sources provide information along a few dimensions; this is different from Bayesian approaches where even small open-world extension would require the recalculation of the whole distribution.

7 General Discussion, Limitations and Conclusion

An advantage of learning belief-theoretic rules is to be able to apply existing DS-theoretic logic formalisms (e.g., *Uncertain Logic* [20, 22]) to perform all manner of *inference* (e.g., modus ponens, AND, OR). This sort of belief-theoretic inference has found applications in many AI and robotic cognitive architectures [23, 24], so learning rule parameters from data would be beneficial.

DS-based operations are typically of exponential time complexity in the size of the frame since we consider all possible subsets of the frame. Although there are efficient implementations of DS-theoretic methods (graphical models and Monte-Carlo) [12], the proposed approach is generally intractable for large datasets. Thus, for large datasets, Bayesian approaches may be preferred.

Epistemic uncertainty is highly relevant to many open-world datasets. "Whereas the Bayesian language asks, in effect, that we think in terms of a chance model for the facts in which we are interested, the belief-function language asks that we think in terms of chance model for the reliability and meaning of our evidence." [25]. In this paper, we proposed a promising new probabilistic logic learning framework that uses a belief-theoretic logical representation combined with a learning methodology that allows for learning interval uncertainty for logical rules. The proposed approach offers several advantages over traditional Bayesian approaches when learning from *small imperfect datasets in the open world*.

[4] Note the distinction between a source that is ignorant (when it is capable of providing an observation on an attribute but is unsure of its value) and the situation when the source is just incapable of providing any value as are the cases represented by red-crosses in Table 2.

References

[1] De Raedt, L. *Logical and relational learning* (Springer Science & Business Media, 2008).

[2] De Raedt, L. & Kersting, K. Probabilistic logic learning. *ACM SIGKDD Explorations Newsletter* **5**, 31–48 (2003).

[3] Getoor, L. *Introduction to statistical relational learning* (MIT press, 2007).

[4] Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference* (Morgan Kaufmann, 2014).

[5] Carlson, A. *et al.* Toward an architecture for never-ending language learning. In *AAAI*, vol. 5, 3 (2010).

[6] Shafer, G. *A Mathematical Theory of Evidence* (Princeton University Press, 1976).

[7] Delmotte, F. & Smets, P. Target identification based on the transferable belief model interpretation of dempster-shafer model. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* **34**, 457–471 (2004).

[8] Seraji, H. & Serrano, N. A multisensor decision fusion system for terrain safety assessment. *IEEE Transactions on Robotics* **25**, 99–108 (2009).

[9] Mahoor, M. H. & Abdel-Mottaleb, M. A multimodal approach for face modeling and recognition. *IEEE Transactions on Information Forensics and Security* **3**, 431–440 (2008).

[10] Pearl, J. Reasoning with belief functions: a critical assessment. *Technical ReportR-136, UCLA* (1989).

[11] Pearl, J. Reasoning with belief functions: An analysis of compatibility. *International Journal of Approximate Reasoning* **4**, 363–389 (1990).

[12] Wilson, N. Algorithms for dempster-shafer theory. In *Handbook of defeasible reasoning and uncertainty management systems*, 421–475 (Springer, 2000).

[13] Smets, P. & Kennes, R. The transferable belief model. *Artificial intelligence* **66**, 191–234 (1994).

[14] Bergsten, U. & Schubert, J. Dempster's rule for evidence ordered in a complete directed acyclic graph. *International Journal of Approximate Reasoning* **9**, 37–73 (1993).

[15] Yager, R. R. On the dempster-shafer framework and new combination rules. *Information Sciences* **41**, 93–137 (1987).

[16] Nunez, R. C. *et al.* DS-based uncertain implication rules for inference and fusion applications. *Information Fusion (FUSION), 2013 16th International Conference on* 1934–1941 (2013).

[17] Tang, Y., Hang, C. W., Parsons, S. & Singh, M. Towards argumentation with symbolic dempster-shafer evidence. *Frontiers in Artificial Intelligence and Applications* **245**, 462–469 (2012).

[18] Dewasurendra, D. A., Bauer, P. H. & Premaratne, K. Evidence filtering. *IEEE Transactions on Signal Processing* **55**, 5796–5805 (2007).

[19] Fagin, R. & Halpern, J. Y. A new approach to updating beliefs. *arXiv preprint arXiv:1304.1119* (2013).

[20] Núnez, R. C., Scheutz, M., Premaratne, K. & Murthi, M. N. Modeling uncertainty in first-order logic: A dempster-shafer theoretic approach. In *8th International Symposium on Imprecise Probability: Theories and Applications* (2013a).

[21] Shafer, G. Non-additive probabilities in the work of bernoulli and lambert. In *Classic Works of the Dempster-Shafer Theory of Belief Functions*, 117–182 (Springer, 2008).

[22] Núnez, R. C., Scheutz, M., Premaratne, K. & Murthi, M. N. Modeling uncertainty in first-order logic: A dempster-shafer theoretic approach. In *8th International Symposium on Imprecise Probability: Theories and Applications* (2013).

[23] Sarathy, V. & Scheutz, M. A logic-based computational framework for inferring cognitive affordances. *IEEE Transactions on Cognitive and Developmental Systems* **8** (2016).

[24] Williams, T., Briggs, G., Oosterveld, B. & Scheutz, M. Going beyond command- based instructions: Extending robotic natural language interaction capabilities. In *Proceedings of AAAI* (2015).

[25] Shafer, G. & Tversky, A. Languages and designs for probability judgment. *Cognitive Science* **9**, 309–339 (1985).