

Let's demonstrate how a complex real-life problem can be formulated in terms of probabilistic DL reasoning.

Consider the domain of cancer, more precisely, women's breast cancer, yet more specifically, the issue of breast cancer risk assessment [1]. Very roughly, the central problem is to combine all the risk factors that apply to a particular woman and come up with a credible number reflecting her chance of developing breast cancer, either in her lifetime or in the short term (normally in the next 10 years). There are a few models that do that, basically, just by computing an empirically inferred function of input parameters (risk factors).

Pronto offers a different way of approaching the problem. It supports a wide use of all the background knowledge captured in a classical ontology—for example, of the sort maintained in the NCI Thesaurus [2] —but also allows us to augment the classical KB with probabilistic statements, such that the risk of developing cancer can be computed as an ontological inference. That makes modeling a lot more explicit and illustrative—especially with support of Pellet's explanations—than using a “black box” function.

So, consider a classical part of an ontology [3] for modeling the breast cancer domain—by the way, we're not claiming it's correct or useful from medical point of view; you may also consider Matt Williams' version of clinical ontology [4] if you're concerned with correctness of terms and stuff like that.

The ontology defines risk factors that are relevant to breast cancer, i.e., subclasses of RiskFactor. Then it also defines different categories of women, first, those that have certain risk factors (subclasses of WomanWithRiskFactors); and, second, those distinct in terms of the risk of developing cancer (subclasses of WomanUnderBRCRISK). The basic task is to compute the probability that a certain woman is an instance of some WomanUnderBRCRISK subclass given that she is an instance of some WomanWithRiskFactors subclass. In addition, it will be useful to infer the generic probabilistic subsumption between classes under WomanUnderBRCRISK and under WomanWithRiskFactors.

The first thing to do in order to enable such probabilistic reasoning is to express the uncertain background knowledge about the domain. This is done by listing the conditional constraints [5] in the form of OWL 1.1 axiom annotations. The constraints can either be in a separate file that imports the classical OWL ontology or be embedded into the classical part. For this example, the constraints express how individual risk factors influence the risk of developing cancer (numbers taken from “Risk Factors and Prevention” [6]). The job of Pronto is to combine factors that apply to a particular woman and compute the probability that she is an instance of some WomanUnderBRCRISK subclass.

Let's now quickly go through the individuals to illustrate the reasoning:

- Julie is a woman in her thirties. The only risk factor that applies to her is AgeUnder50, so the Pronto concludes that Julie:(WomanWithBRCInShortTime|owl:Thing)[0.0;0.027] (her chance of developing cancer in next 10 years is no higher than 2.7%)
- Mary is known to have BRCA1 gene mutation which is known to be a hugely important risk factor. Using the generic constraint (WomanUnderGreatBRCRISK|WomanWithBRCAMutation)[1;1], Pronto puts her in the category of women with the highest relative risk of cancer (this example, also shows, that conditional constraints, with some caveats, can model certain subsumption relationships)
- For Ann we know two risk factors – her mother had BRC and she is an Ashkenazi Jew, so she has an increased chance of having inherited gene mutation. Using the combination of risk factors without overriding, Pronto concludes that she has a 31.25% chance of being in the category of 3x increased risk and over 2.5% of being in the highest risk category.

- Helen is the most interesting case. For her we again know 2 risk factors – her age is over 50 and her mother had cancer. Using overriding we can specify how these two factors strengthen or weaken each other to produce the actual risk. This can be done by defining a generic constraint
(WomanUnderGreatBRCRisk|SeniorWomanWithMotherBRCAffected)[0.9;1] that overrides constraints for both factors individually. Thus, Pronto entails that she is in the highest risk category with more than 90% probability.

Finally, I want to mention explanations, which are an important part of Pellet's reasoning services. I begin by pointing out that DL reasoning can be difficult to understand and probabilistic reasoning can also be difficult to understand. Not surprisingly, the hybrid reasoning that Pronto is capable of can be very difficult to understand. This is both a limitation and an opportunity to extend Pellet's explanation services to Pronto.

In other words, we want to extract the minimal set of conditional constraints that are sufficient to produce the given probabilistic entailment. For the breast cancer example above that would imply filtering out all the irrelevant risk factors and leaving only those which were taken into account during reasoning. We've got some initial work done in extending explanations in Pronto, but there's much more work to do, including extending debug and repair features of Pellet to Pronto.

Resources:

[1] NCI Breast Cancer Risk Calculator: <http://www.cancer.gov/bcrisktool/>

[2] National Cancer Institute Thesaurus: <http://www.nci.nih.gov/cancerinfo/terminologyresources>

[3] Available at: [examples/brc/ontologies/cancer_ra.owl](#)

[4] Available at: <http://acl.icnet.uk/%7Emw/MDMDomain0.1.owl>

[5] Available at: [examples/brc/ontologies/cancer_cc.owl](#)

[6] Available at: <http://cms.komen.org/komen/AboutBreastCancer/RiskFactorsPrevention/index.htm>