Algoritmos Eficientes para Aprendizado de Lógicas Descritivas Probabilísticas

Trabalho de Mestrado

Raphael Melo Thiago (Aluno)[1], Kate Revoredo (Orientadora)[1], Aline Paes[2] (Co-orientadora)

Programa de Pós Graduação em Informática (PPGI)

[1]Departamento de Informática Aplicada – Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
[2]Universidade Federal Fluminense (UFF)

{raphael.thiago, katerevoredo}@uniriotec.br, alinepaes@ic.uff.br

Ano de Ingresso no Programa de Mestrado: 2012

Época esperada de conclusão: Abril de 2014

Etapas já concluídas: disciplinas cursadas, revisão inicial da literatura e definição geral da proposta.

Resumo. *As lógicas descritivas (DLs) tem recebido atenção dado o seu papel fundamental dentro da Web Semântica. A adição de conceitos probabilísticos às DLs torna os modelos mais expressivos e representativos. Dada a escala e magnitude dos dados existentes fazem-se necessários mecanismos eficientes de aprendizado automático de modelos na DL escolhida. Nesta dissertação novos mecanismos escaláveis de aprendizado de DLs probabilísticas serão propostos e comparados com os mecanismos não escaláveis existentes.*

Palavras-chave. *Aprendizado de lógicas descritivas, Lógica Descritiva probabilística.*

1. Introdução

A geração de modelos que representem o conhecimento sobre o "mundo" sempre foi uma necessidade para o raciocínio computacional [Brachman e Levesque 2004]. Essa necessidade é ainda maior no campo da *web* semântica, onde o seu sucesso está intimamente relacionado com o corpo de conhecimento já modelado. Assim, um grande campo de estudo da *web* semântica tem sido a geração automática de modelos, já que a definição manual é uma atividade não trivial e custosa [Maedche e Staab 2001].

As lógicas descritivas [Baader e Nutt 2010] são linguagens de representação de conhecimento utilizadas dentro da *web* semântica. Elas são fragmentos decidíveis de lógica de primeira ordem utilizados para representar conhecimento determinístico. Em alguns casos, dada a natureza de um cenário, a representação através de um modelo determinístico não é suficiente para representá-lo fielmente, assim a adição de conceitos probabilísticos se torna atraente. Por exemplo, em uma rede social, onde os nós são

Gostaria de agradecer à Capes por seu apoio com a bolsa de mestrado.

indivíduos e as arestas indicam algum relacionamento entre eles, a predição de links [Liben-Nowell e Kleinberg 2003] é um problema que tem benefícios ao considerar a semântica do domínio dos indivíduos para optar pela inclusão ou não de um novo link. Além disso, em [Ochoa et.al., 2013] foi mostrado o beneficio de se considerar a incerteza inerente ao problema, através da consideração de uma lógica descritiva probabilística para a modelagem do domínio.

Como dito anteriormente, o principal obstáculo para as definições de modelos formais são os custos associados. Este problema é ainda mais acentuado ao serem adicionados componentes probabilísticos. Dada essa problemática e o grande número de modelos necessários à utilidade da *web* semântica, mecanismos automáticos de geração de modelos se tornam indispensáveis.

O objetivo dessa dissertação é a definição de um *framework* de aprendizado que considere algoritmos de aprendizado escaláveis para a geração de modelos em lógica descritiva probabilística. Em [Ochoa-Luna et al. 2011], um mecanismo automático de aprendizado para lógicas descritivas probabilísticas foi apresentado, no entanto, este mecanismo não é escalável. Ele será então utilizado como base para a avaliação de algoritmos alternativos escaláveis.

2. Lógicas Descritivas Probabilísticas

As lógicas descritivas (DL) formam uma família de formalismos lógicos que servem para representar conhecimento [Baader e Nutt 2010] por meio da definição de conceitos relevantes (sua terminologia - TBox), e então usa esses conceitos para especificar propriedades de objetos e indivíduos que ocorrem em seu domínio (a descrição do mundo - ABox). A arquitetura das lógicas descritivas pode ser dividida na base de conhecimento e nos mecanismos de inferência.

Nas lógicas descritivas a criação de conceitos complexos é realizada utilizando os construtores definidos para cada linguagem particular. Os construtores da DL *ALC* [Baader e Nutt 2010] são conjunção (C ⊔ D), disjunção (C ⊓ D), negação ($\neg C$), restrição existencial ($\exists r.C$) e restrição de valor ($\forall r.C$). As definições da coluna esquerda na Figura 1 estão definidas em DL *ALC*.

Humano ≡ Animal ⊓ Racional	$P(\text{Animal}) = 0.9$
Besta ≡ Animal ⊓ ¬Racional	$P(\text{Racional}) = 0.6$
Pai ≡ Humano ⊓ ∃temFilho.Humano	$P(\text{temFilho}) = 0.3$
CanguruPai ≡ Canguru ⊓ ∃temFilho.Canguru.	$P(\text{Canguru}\|\text{Besta}) = 0.4$
	$P(\text{Canguru}\|\neg\text{Besta}) = 0.0$

Figura 1. Exemplo de terminologia descrita utilizando a lógica descritiva probabilística cr*ALC* [Cozman e Polastro 2009]

As lógicas descritivas probabilísticas (PDL) são divididas de acordo com o tipo de inclusões probabilísticas permitidos: *i.* (Semântica baseada em domínio) $P_D(Professor) = \alpha$, um objeto aleatório é um professor com a probabilidade α; *ii.* (Semântica baseada em interpretações) $P\big(Professor(João)\big) = \alpha$, α é a probabilidade do conjunto de interpretações onde $João$é um $Professor$.

A cr*ALC* [Cozman e Polastro 2009] é uma PDL baseada na DL *ALC* que adota a semântica baseada em interpretações. As terminologias em cr*ALC* nada mais são que terminologias em *ALC* com a adição de inclusões probabilísticas (Figura 1). As inclusões probabilísticas podem ser de três tipos (veja a Figura 2); onde C é um nome de conceito, D é um conceito e r é um nome de papel.

$$\begin{aligned} P(C) &\in [\underline{\alpha}, \overline{\alpha}], \\ P(C|D) &\in [\underline{\alpha}, \overline{\alpha}], \\ P(r) &\in [\underline{\beta}, \overline{\beta}]. \end{aligned}$$

Figura 2. Inclusões Probabilísticas

2.1. Inferências

Além das inferências normais presente nas DLs as PDLs possuem um tipo de inferência específico. Essa inferência consiste em encontrar um limite superior/inferior para a probabilidade de uma sentença, formalmente: $P(A_0(a_0)|\mathcal{A})$ dada uma terminologia e um ABox $\mathcal{A} = \{A_1(a_1), A_2(a_2), \ldots, A_j(a_j)\}$ (detalhes em [Cozman e Polastro 2009]).

2.2. Aprendizado em Lógica Descritiva Probabilísitca cr*ALC*

O aprendizado de DLs é um campo que vem sendo estudado e possui resultados reconhecidos pela academia. Por outro lado, pesquisas de algoritmos de aprendizado de PDLs ainda é incipiente, necessitando de mais trabalhos em diversas frentes, a escalabilidade dos algoritmos de aprendizado é uma delas.

O problema de aprendizado de lógicas descritivas pode ser sumarizado da seguinte forma:

Dado:

Uma base de conhecimento K

Um conceito alvo $Target$, onde $Target \notin K$

Um conjunto $E, E = E_p \cup E_n$

Encontrar:

Definição determinística: $C(Target \equiv C)$ tal que $K \cup C \vdash E_p$ e $K \cup C \nvdash E_n$

sse, uma definição determinística não for encontrada:

P(C|Condição), onde Condição é obtida a partir da definição de C encontrada.

Figura 3. Problema de aprendizado formalizado [Ochoa-Luna et al. 2011]

Com essa definição um algoritmo geral de aprendizado de lógicas descritivas probabilísticas pode ser definido com a adição dos seguintes elementos [Ochoa-Luna et al. 2011]:

Operador de refinamento: define a árvore de busca de conceitos determinísticos.

Função de busca: define como percorrer a árvore de busca.

Função de avaliação: responsável por retornar a qualidade de um conceito em relação ao conjunto E; geralmente utiliza a cobertura dos exemplos pela definição do conceito.

Em [Ochoa-Luna et al. 2011] um algoritmo de aprendizado para PDLs é proposto (Algoritmo 1), nele a árvore de busca é expandida até que o critério de parada seja obedecido. O nó com o maior valor para a função de avaliação `valor` é considerado o melhor candidato para a definição de $Target$, se o seu valor for maior que um limite inferior ele é retornado como a definição determinística para o conceito $Target$. Se não, uma inclusão probabilística deve ser encontrada para explicar os exemplos. Este algoritmo tem como vantagem a sua abordagem conjunta de aprendizado. No entanto existem algumas evidências que apontam que este mecanismo não se comporta bem em bases de dados grandes [Polastro et al. 2012], principalmente em relação às inferências realizadas com os modelos gerados.

Algoritmo 1. Algoritmo geral de aprendizado de conceitos em Lógicas descritivas probabilísticas [Ochoa-Luna et al. 2011]

```
Sejam:  K=<T, A> - uma base de conhecimento, Target - conceito alvo e
E - conjunto de treinamento
Entrada (K, Target, E)
ArvoreDeBusca = {C=T, h=0}
Repetir
        Escolha um no N={C,h} com a melhor avaliação em ArvoreDeBusca
        Expandir o nó para o tamanho h+1:
                Adicione todos os nós D ∈ (operador de refinamento(C)) com
                tamanho = h+1
                aprenda os valores dos nós em D
                N = {C, h+1}
até critério de parada
N'=melhor nó em ArvoreDeBusca
se valor(N')>limite então
        retornar conceito determinístico C'∈ N' (Target≡C')
senão
        retornar inclusão probabilística(ArvoreDeBusca, Target)
fim se
```

3. Pontos de Modificação

Para encontrar um possível ponto de melhoria para os mecanismos existentes foi realizado um estudo da estrutura geral de aprendizado e utilização da PDL cr*ALC*, a Figura 4 mostra o resultado desta análise. Cada caixa presente na Figura 4 representa um ponto de interesse no *framework* da PDL cr*ALC*, isso o aprendizado da terminologia (1) e os mecanismos de inferência (2).

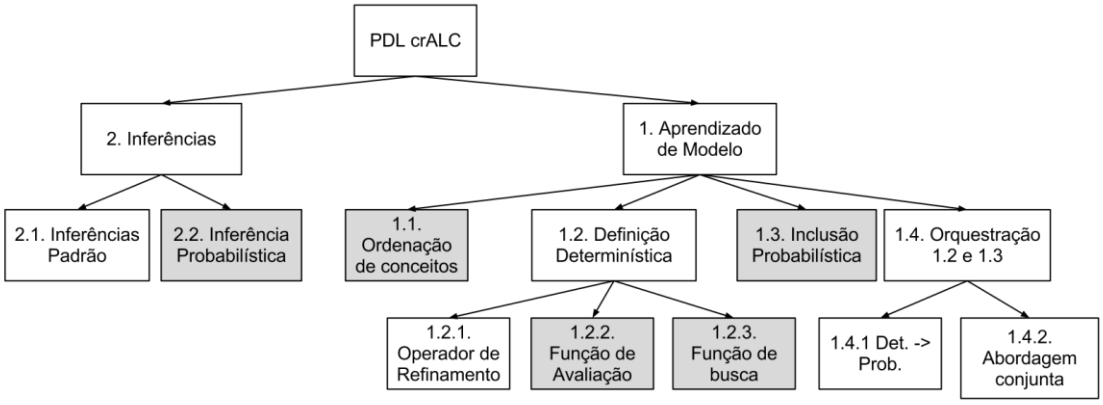

Figura 4 Arquitetura solução PDL cr*ALC*

As caixas cinza na Figura 4 representam os pontos de modificação mais promissores. São eles:

1. Algoritmo de inferência (2.2): modificar para realizar inferências *lifted* [Kisynski e Poole 2009];
2. Ordenação de conceitos para aprendizado (1.1): levando em consideração que a tarefa de aprendizado é realizada repetidamente para vários conceitos de um mesmo domínio, a mudança na ordenação de cada problema de aprendizado implica em diferentes modelos aprendidos ao fim.
3. Algoritmo de aprendizado (1.2.2 ,1.2.3, 1.3).

4. Trabalhos Relacionados

Existem três algoritmos de aprendizado da PDLs cr*ALC* na literatura [Ochoa-Luna e Cozman 2009] [Revoredo et al 2010] [Ochoa-Luna et al 2011], cada um apresenta uma abordagem ligeiramente diferente, culminando no *framework* proposto em [Ochoa-Luna et al 2011] que utiliza uma abordagem de aprendizado conjunta, esta abordagem será utilizada como base para as modificações realizadas neste trabalho.

Todos os algoritmos de aprendizado são inspirados por conceitos de programação em lógica indutiva (ILP) [Lavrac e Dzeroski 1994] e compartilham algumas de suas suposições.

Em [Lehmann e Pascal 2010] é apresentado o *framework* DL-Learner, ele pode ser considerado o padrão *de facto* para o aprendizado de DLs. O DL-Learner é importante, pois ele pode ser utilizado como o operador de refinamento no algoritmo de aprendizado.

5. Contribuições da pesquisa

Definição de pontos de melhoria no *framework* geral da cr*ALC*.

Criação de um algoritmo que aprenda ontologias descritas em cr*ALC* e que possa ser utilizado em cenários que as bases de dados sejam grandes.

Análise do comportamento do algoritmo proposto em [Ochoa et al. 2011] em relação ao aumento de tamanho das bases geradoras.

6. Estado Atual do Trabalho

O trabalho encontra-se no estágio inicial do desenvolvimento e pesquisa sobre as melhorias na escalabilidade, tendo sido realizada a revisão inicial da literatura e uma análise dos pontos de melhoria mais interessantes.

Referências

Baader, F. e Nutt, W. (2010) “Basic description logics”. In: The Description Logic Handbook, pages 47-100, Editado por Franz Baader, Deborah L. McGuinness, Daniele Nardi e Peter F. Patel-Schneider, Cambridge University Press.

Brachman, R. J. e Levesque, H. J. (2004) "Knowledge Representation and Reasoning", Elsevier.

Maedche, A. e Staab, S. (2001) "Ontology learning for the semantic web", IEEE Intell Syst Vol. 16, Issue 2, p. 72-79.

Liben-Nowell, D. e Kleinberg, J. (2003) "The link prediction problem for social networks", Proceedings of the twelfth international conference of Information and Knowledge Management, p. 556-559, ACM.

Ochoa-Luna J. E., Revoredo, K e Cozman, F. G. (2011) "An Experimental Evaluation of a Scalable Probabilistic Description Logic Approach for Semantic Link Prediction", Journal of Brazilian Computer Science, 2013, a ser publicado.

Ochoa-Luna, J. E., Revoredo, K. e Cozman, F. G. (2011) "Learning Probabilistic Description Logics: A Framework and Algorithms", Advances in Artificial Inteligence, Lecture Notes in Computer Science Vol. 7094, p. 28-39.

Cozman, F. G. e Polastro, R. B. (2009) "Complexity analysis and variational inference for interpretation-based probabilistic description logics", Conference on Uncertainty in Artificial Intelligence, p. 117-125.

Lehmann, J. e Hitzler, P (2010) "Concept learning in description logics using refinement operators", Machine Learning, Vol. 78, Issue 1-2, p. 203-250.

Polastro, R. B., Cozman, F. G., Takiyama, F. I. e Revoredo, K. C. (2012) “Computing Inference for Credal ALC Terminologies”, Proceedings of the 8th International Workshop on Uncertainty Reasoning for the Semantic Web, p. 94-97.

Kisynski, J. e Poole, D. (2009) “Lifted Aggregation in directed first-order probabilistic models”, Int. Joint Conf. on Artificial Intelligence, p. 1921-1929.

Ochoa-Luna, J., Cozman, F. (2009) “An algorithm for learning with probabilistic description logics”, Proceedings of the 5th International Workshop on Uncertainty Reasoning for the Semantic Web, p. 63–74.

Revoredo, K., Ochoa-Luna, J. E. e Cozman, F. G. (2010) “Learning terminologies in probabilistic description logics”, Advances in Artificial Intelligence SBIA 2010, p. 41–50.

Lavrac, N.; Dzeroski, S. (1994) “Inductive Logic Programming: Techniques and Applications”, New York: Ellis Horwood.