软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(8):1699−1713 [doi: 10.3724/SP.J.1001.2011.04053]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

马尔可夫逻辑网络研究*

徐从富, 郝春亮+, 苏保君, 楼俊杰

(浙江大学 人工智能研究所,浙江 杭州 310027)

Research on Markov Logic Networks

XU Cong-Fu, HAO Chun-Liang+, SU Bao-Jun, LOU Jun-Jie

(Institute of Artificial Intelligence, Zhejiang University, Hangzhou 310027, China)

+ Corresponding author: E-mail: chunliang_hao@163.com

Xu CF, Hao CL, Su BJ, Lou JJ. Research on Markov logic networks. *Journal of Software*, 2011,22(8): 1699−1713. http://www.jos.org.cn/1000-9825/4053.htm

Abstract: Markov logic networks (MLNs) is a statistical relational learning (SRL) model, which combines Markov network and first order logic. It has been applied widely in nature language processing, complex networks, information extraction, etc. This paper addresses the theoretical model of Markov logic networks (MLNs), weight and structure learning of MLNs comprehensively, and finally presents future works of MLNs.

Key words: Markov logic network; statistical relational learning; probabilistic graphical model; inference; weight learning; structure learning

摘 要: 马尔可夫逻辑网络是将马尔可夫网络与一阶逻辑相结合的一种统计关系学习模型,在自然语言处理、复杂网络、信息抽取等领域都有重要的应用前景.较为全面、深入地总结了马尔可夫逻辑网络的理论模型、推理、权重和结构学习,最后指出了马尔可夫逻辑网络未来的主要研究方向.

关键词: Markov 逻辑网;统计关系学习;概率图模型;推理;权重学习;结构学习

中图法分类号: TP181 **文献标识码**: A

如何有效地处理复杂性和不确定性等问题,一直是人工智能界的难点之一.为此,国内外学者进行了长期的探索研究.近年来,先后提出了统计关系学习(statistical relational learning,简称 SRL)[1−3]和概率图模型(probabilistic graphical model,简称 PGM)[4,5]等重要方法,并引起了极大关注.简单来说,统计关系学习通过集成关系/逻辑表示、概率推理、不确定性处理、机器学习和数据挖掘等方法,以获取关系数据中的似然模型.目前,关系/逻辑表示仍以一阶谓词逻辑为主,用以紧凑表达领域知识和有效处理逻辑问题,主要包括逻辑程序(logic programming)[6−8]和描述逻辑(description logic)[9].而概率图模型则是一种通用化的不确定性知识表示和处理方法,主要涵盖了贝叶斯网络(Bayesian networks,简称 BNs)[10−12]、隐马尔可夫模型(hidden Markov model,简称 HMM)[13,14]、马尔可夫决策过程(Markov decision process,简称 MDP)[15−17]、神经网络(neural network,简称 NN)[18,19]等.一个很自然的想法是,将统计关系学习(尤其是关系/逻辑表示)和概率图模型进行综合集成.前几年,国内外在这方面的研究工作主要侧重于将概率统计和一阶谓词逻辑的某个特定子集(如 Horn 从句[20]、数据库

* 基金项目: 国家自然科学基金(60970081); 国家重点基础研究发展计划(973)(2010CB327903)
收稿时间: 2010-06-22; 修改时间: 2010-10-29; 定稿时间: 2011-04-28

© 中国科学院软件研究所 http://www.jos.org.cn

查询语言等)结合起来,但通常都相当复杂.

2004 年,美国华盛顿大学的 Domingos 和 Richardson[1]首次提出了马尔可夫逻辑网络(Markov logic networks,简称 MLNs,以下简称 Markov 逻辑网),并论证了 Markov 逻辑网作为统计关系学习的统一框架的可能性[1].其主要理由如下:一方面,从概率统计的角度来看,Markov 逻辑网不仅为简洁明了地描述庞大的 Markov 网(Markov networks,简称 MNs)提供了一种紧凑有效的手段,而且它还能够灵活地在 Markov 网中融入模块化知识域的能力;另一方面,从一阶谓词逻辑的角度来看,Markov 逻辑网给一阶谓词逻辑加入了出色的不确定性处理能力,并且能够容忍知识域中存在的不完整性和矛盾性等问题.因此,统计关系学习中的许多重要工作(如聚合分类、关系预测、社会网络及目标识别等)都可统一到 Markov 逻辑网的体系中来[21].鉴于此,当前国际人工智能界普遍公认 Markov 逻辑网是一种较完美地结合一阶谓词逻辑和概率图模型的复杂性和不确定性问题表示和处理方法[22,23],具有十分重要的研究价值和广阔的应用前景,已成为人工智能、机器学习、数据挖掘等领域的研究热点.

近年来,Markov 逻辑网的研究进展非常迅猛,涌现出一大批研究成果.2005 年,Kok 等人提出了一种自顶而下的从关系数据库中学习 Markov 逻辑网结构的算法[24].Singla 等人为解决伪似然参数学习中非邻接变量之间推理结果不理想的问题[1],提出了一种判别训练(discriminative training)的参数学习方法[25].2006 年,Singla 等人提出了一种基于 Markov 逻辑网的实体解析方法[26].2007 年,Singla 等人将 Markov 逻辑网推广至无限域[27],Lowd 等人提出了几种高效的权重学习算法[28].Mihalkova 等人提出了一种自底而上的 Markov 逻辑网结构学习算法,极大地提高了结构学习的搜索效率和准确率[29].2008 年,Wang 等人为解决 Markov 逻辑网只能处理离散变量的问题,提出了混合 Markov 逻辑网(hybrid Markov logic network,简称 HMLN)[30].Kok 等人采用基于 Markov 逻辑网的非监督方法从大量文本中抽取出语义信息网[31].2009 年,Kok 等人提出了基于 Hypergraph Lifting(即视一个关系数据库为一张超图,而闭原子和关系分别视作超图的节点和边)的 Markov 逻辑网结构学习算法[32].Nath 等人提出了基于 Markov 逻辑网的关系决策理论框架,且证明了经典的规划理论和 Markov 决策过程均为其特殊形式[33].Davis 等人采用二阶 Markov 逻辑网进行深度迁移学习(deep transfer learning),并成功地应用于分子生物学、社会网络等领域[34].Poon 等人采用基于 Markov 逻辑网的非监督方法进行语义分析,从一系列生物医学文本中抽取出一个知识库,并基于该知识库回答相关的医学问题[35].2010 年,Kok 等人针对经典 Markov 逻辑网结构学习算法难以适应较长从句的缺点,提出了基于结构主题(structural motifs)的学习方法[36].

本文在简要介绍 Markov 网和一阶逻辑的基础上,重点论述 Markov 逻辑网的推理、参数学习及结构学习的最新研究进展,总结了 Markov 逻辑网的典型应用,最后探讨了 Markov 逻辑网的未来研究方向.

1 Markov 逻辑网

1.1 Markov网和一阶逻辑

Markov 网[37]也称 Markov 随机场(Markov random field,简称 MRF)[4],是一个变量集合 $X=(X_1,X_2,\ldots,X_n)\in\chi$的联合分布模型.它由一个无向图 G 和定义于 G 上的一组势函数ϕ_k组成.其中,无向图的每个节点都代表一个随机变量,而 G 中的每一个“团(clique)”都对应着一个势函数(为非负实函数),表示团的一个状态.Markov 网所代表的变量集的联合分布表示为

$$P(X=x)=\frac{1}{Z}\prod_k \phi_k(x_{\{k\}}) \tag{1}$$

其中,$x_{\{k\}}$表示 Markov 网中第 k 个团的状态,即对应于第 k 个团中所有变量的取值状态.Z 是归一化因子,且 $Z=\sum_{x\in\chi}\prod_k \phi_k(x_{\{k\}})$.通常,公式(1)表示为对数线性模型,以便于对 Markov 网中蕴含的实体特征进行具体化,并使得推理和学习等过程更加方便.若把 Markov 网中每个团的势函数表示为指数函数,指数项为对应团的加权特征量,则可得:

$$P(X=x)=\frac{1}{Z}\exp\left\{\sum_j \omega_j f_j(x)\right\} \tag{2}$$

© 中国科学院软件研究所 http://www.jos.org.cn

其中,ω_j 表示权重,$f_j(x)$表示特征函数.理论上,这里的特征函数可以是任意的实值函数,然而为方便讨论,本文涉及的特征函数均为二元函数.从用势函数表达的公式(1)来看,可直观地认为每个特征量对应于一个团的某一状态,即团中变量集的一个取值,且该特征量的权重等于 $\log\varphi_k(x_{\{k\}})$.在 Markov 网中,最常用的近似推理算法是马尔可夫链蒙特卡洛(Markov chain Monte Carlo,简称 MCMC)方法[9],其代表是吉布斯采样(Gibbs sampling).简单地说,吉布斯采样是在给定一个变量的 Markov 覆盖(Markov blanket,也称 Markov 毯)的条件下,根据概率 $P(x|MB(x))$进行采样,其中,$MB(x)$表示变量 x 的 Markov 覆盖.在 Markov 网中,$MB(x)$仅表示变量 x 的所有邻居节点.当需要计算边缘概率时,比如 $P(X=x)$,仅需在所有抽到的样本中计算满足 $X=x$ 的样本数,然后除以样本总数.当计算条件概率时,比如 $P(X=x|Y=y)$,需要先设置变量 Y 的值为 y,然后进行采样.另一个较常用的推理方法是信念传播算法(也称和积算法)[4].

一阶谓词知识库[37]由若干一阶谓词规则组成.一阶谓词规则由 4 种类型的符号组成,即常量、变量、函数和谓词.其中,常量指定义域里一个简单的对象;变量可以指定义域里若干对象;函数(如 *SonOf*())表示一组对象到一个对象的映射;谓词指定义域中若干对象之间的关系(如 *Friends*())或者对象的属性(如 *Red*()).变量和常量可以有类型,一个类型的变量仅能从定义类型的对象集中取值.一个项可以是任意地表示一个对象的表达式,如 *SonOf*(*x*).原子是作用于一组项的谓词(如 *SonOf*(*x*,*BrotherOf*(*y*))).一个常项是指没有变量的项.一个闭原子(ground atom)或闭谓词(ground predicate)是指所有参数均为常项的原子或谓词.一般地,规则是从原子开始,用连接词(如蕴含关系、等价关系等)和量词(如全称量词和存在量词)递归地建立起来.在 Markov 逻辑网中,通常把规则表示成从句形式.一个可能的世界(a possible world)是指给所有可能出现的闭原子都赋予了真值.关于一阶谓词逻辑的详细知识可参见文献[38].

1.2 Markov逻辑网的定义和示例

一阶逻辑知识库可看作是在一个可能世界的集合上建立一系列硬性规则,即如果一个世界违反了其中的某一条规则,那么这个世界的存在概率即为 0.Markov 逻辑网的基本思想是让那些硬性规则有所松弛,即当一个世界违反了其中的一条规则时,这个世界存在的可能性将降低,但并非不可能.一个世界违反的规则越少,这个世界存在的可能性就越大.为此,给每个规则都加上了一个特定的权重,它反映了对满足该规则的可能世界的约束力.一个规则的权重越大,对于满足和不满足该规则的两个世界而言,它们之间的差异就越大. Markov 逻辑网的定义如下:

定义 1[37]. Markov 逻辑网 L 是一组二元项(F_i,ω_i),其中,F_i 表示一阶逻辑规则,ω_i 是一个实数.这组二元项(F_i,ω_i)与一组有限常量集 $C=\{c_1,c_2,\ldots,c_n\}$一起定义了一个 Markov 网 $M_{L,C}$:

(1) L 中的任意闭原子(ground atom)都对应了 $M_{L,C}$ 中的一个二值节点.若此闭原子为真,则对应的二值节点取值为 1;若为假,则取值为 0.

(2) L 中的任意闭规则(ground formula)都对应着一个特征值,若此闭规则为真,则对应的特征值为 1;若为假,则特征值为 0.并且,这个特征值的权重为二元项中该规则 F_i 对应的权重ω_i.

由定义 1 可知,Markov 网 $M_{L,C}$ 的节点是由 Markov 逻辑网 L 中每个闭原子生成的,边是由闭原子之间的关系生成的.因此,Markov 逻辑网可看作是一个用以生成 Markov 网的模板.给定相同的 Markov 逻辑网和不同的有限常量集合 C,可以产生不同的 Markov 网,并且这些 Markov 网在规模上的差别可能会非常巨大.但是,从同一个 Markov 逻辑网中产生的不同 Markov 网在结构和参数上都有若干共同点.比如,有相同的团的数目,同一个规则的所有可能的常量取值有相同的权重,等等.根据这种方式产生的每一个 Markov 网可称为闭 Markov 网(ground Markov network).从上述 Markov 逻辑网定义和公式(1)、公式(2)可知,一个闭 Markov 网中所蕴含的可能世界 x 的概率分布为

$$P(X=x)=\frac{1}{Z}\exp\left\{\sum_i \omega_i n_i(x)\right\}=\frac{1}{Z}\prod_i \phi_i(x_{\{i\}})^{n_i(x)} \tag{3}$$

其中,$n_i(x)$表示关于规则 F_i 的取值为真的对应闭规则的个数;$x_{\{i\}}$表示出现在规则 F_i 中的原子集合的状态,且 $\phi(x_{\{i\}})=e^{w_i}$.公式(3)中的第 1 个等式给出了 Markov 逻辑网的对数线性模型,第 2 个等式采用了等价的势函数乘

© 中国科学院软件研究所 http://www.jos.org.cn

积形式.直觉上,权重极大的规则代表着此规则的约束力非常强,由公式(3)可知,若某个世界违反了此规则,则该世界存在的概率将趋向于 0.事实上,规则 F 中涉及的变量通常会出现在其他规则之中,当改变规则 F 的值时,并不能保证其他规则的值保持不变.所以,在这种常见的情况中,规则 F 的权重与其概率之间并不存在一一对应关系.如果把规则的权重视为最大熵分布,或者把规则的权重视为经验概率,那么,所有规则的概率共同决定了规则 F 的权重.在一个学习成熟的 Markov 逻辑网中,规则的权重可以由包含规则的经验概率共同计算得到.由上可知,若一个世界违反了规则 F,而规则 F 的权重并不十分巨大,那么这个世界是有可能存在的.因此, Markov 逻辑网可以有效地处理知识域的矛盾性,容忍其内在缺陷.需要特别指出的是,知识域的矛盾性并不一定是其内在缺陷,因为知识域往往只是表示真实世界的某一方面,而真实世界却总是矛盾的综合体.

从 Markov 逻辑网的定义出发,很容易得到一个图结构:当两个节点所表示的闭原子出现在同一个闭规则之中时,这两个节点之间存在一条边.显然,所有出现在同一个闭规则里的闭原子组成了一个(可能不是最大的)团.引入变量的类型可以将变量的取值范围限定在常量集合的子集中,从而极大地减小由 Markov 逻辑网产生的闭 Markov 网的规模.然而,即使采用该方式,所产生的闭 Markov 网也依旧可能非常庞大.因而,在 Markov 逻辑网中推理不可能生成整个闭 Markov 网.

一个简单的 Markov 逻辑网实例见表 1.该 Markov 逻辑网包含两个二元项(F_1,1.5),(F_2,2.2).值得注意的是,如果一个规则包含多个从句,那么这些从句将平分该规则的权重[37].

Table 1 A simple example of Markov logic network[37]

表 1 一个简单的 Markov 逻辑网实例[37]

Proposition	First-Order logic	Clausal form	Weight
Smoking causes cancer.	F_1:$\forall x, Sm(x)\Rightarrow Ca(x)$	$\neg Sm(x)\vee Ca(x)$	1.5
If two people are friends, either both smoke or neither does.	F_2:$\forall x\forall y, Fr(x,y)\Rightarrow (Sm(x)\Leftrightarrow Sm(y))$	$\neg Fr(x,y)\vee Sm(x)\vee\neg Sm(y)$	1.1
		$\neg Fr(x,y)\vee\neg Sm(x)\vee Sm(y)$	1.1

在表 1 中,x 和 y 分别表示个体变项,$Sm(x)$,$Ca(x)$和 $Fr(x,y)$是谓词,分别表示 x 吸烟与否、x 患癌症与否、x 与 y 为朋友与否.规则 F_1 表示吸烟致癌,规则 F_2 表示若 x 与 y 为朋友,则他们的吸烟习惯相同.由表 1 可知,规则 F_1 和 F_2 的权重分别为 1.5 和 2.2,规则的权重指出了满足该规则的世界和不满足该规则的世界之间的对数差异.例如,在其他条件相同时,n 个抽烟的人不患癌症的概率比所有抽烟的人都患癌症的概率小 $e^{1.5n}$.当给定个体常项集合 $C=\{A,B\}$时,则生成如图 1 所示的闭 Markov 网[37].

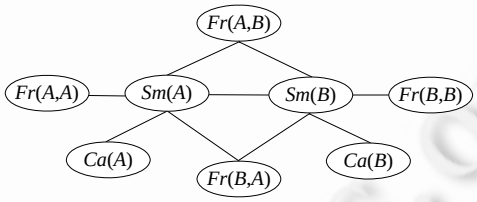

Fig. 1 An example of ground Markov network[37]

图 1 闭 Markov 网示例[37]

图 1 所示的闭 Markov 网中有 6 个团,其相应的权重见表 2.

Table 2 Cliques and weights of ground Markov network in Fig.1

表 2 图 1 所示闭 Markov 网中的团及其权重

First-Order logic	Variable assignment	Clique	Weight
F_1:$\forall x, Sm(x)\Rightarrow Ca(x)$	$x=A$	$\{Sm(A),Ca(A)\}$	1.5
	$x=B$	$\{Sm(B),Ca(B)\}$	1.5
F_2:$\forall x\forall y, Fr(x,y)\Rightarrow(Sm(x)\Leftrightarrow Sm(y))$	$x=A, y=A$	$\{Fr(A),Sm(A)\}$	1.1
	$x=A, y=B$	$\{Fr(A,B),Sm(A),Sm(B)\}$	1.1
	$x=B, y=A$	$\{Fr(B,A),Sm(A),Sm(B)\}$	1.1
	$x=B, y=B$	$\{Fr(B,B),Sm(B)\}$	1.1

© 中国科学院软件研究所 http://www.jos.org.cn

例如,对应于闭规则 $Sm(A)\Rightarrow Ca(A)$的团是$\{Sm(A),Ca(A)\}$,其相应的权重是 1.5.需要特别指出的是,对应于同一规则的团具有相同的权重.

2 Markov 逻辑网的推理

在概率图模型中,推理的基本问题包括计算边缘概率、条件概率以及对于最大可能存在状态的推测(本文称为最大可能性问题).对于 Markov 逻辑网,其推理主要在生成的闭 Markov 网上进行.因此,Markov 逻辑网的推理也属于概率图模型的范畴.然而,由于 Markov 逻辑网引入了知识化模块,所以它又有其自身的特点.

2.1 最大可能性问题

最大可能性问题是概率图模型推理中的重要内容,其基本过程可表述为:给定证据变量集 x,求变量集 y 最可能处于的状态,即求

$$\max_y P(y|x) \tag{4}$$

在 Markov 逻辑网中,结合公式(3),则问题转化为求

$$\max_y \sum_i w_i n_i(x,y) \tag{5}$$

其中,w_i 表示从句 i 的权重,$n_i(x,y)$表示从句 i 的闭从句的真值数量.

因此,计算最大可能性问题可转换为典型的最大化带权的可满足性问题,即寻找一组变量的赋值,使得所有被满足的从句的权重之和最大.通常,可采用任何带权的可满足性问题解决器来处理这类问题,且其效率并不比任何类型的标准逻辑推理方法低.事实上,如果 Markov 逻辑网体系中某些硬性的规则(该规则的权重趋向无穷大)能够适当放松,则推理效率可能会更快.文献[37]成功地运用 MaxWalkSAT 算法解决了这类问题.测试结果表明,MaxWalkSAT 算法可以在几分钟内处理包含上千个变量的可满足性问题. MaxWalkSAT 算法的大致流程为:首先对所有变量随机赋值,然后在所有未满足的从句中随机选取一个从句,改变从句中一个变量 x 的值.对于 x 的选取,有两种情况:根据一定的概率随机选取;或选取这样一个变量,当该变量的值改变时,所有已满足的从句的权重之和将达到最大.变量x的选取体现了“随机”与“贪心”相结合的思想,从而避免了在局部寻优搜索中被卡住而导致算法不能正常收敛.

但是,MaxWalkSAT 算法会碰到一个很严重的问题,即它需要把所有从句和所有原子都转变为闭从句和闭原子.这种处理方式对内存的需求量将会随着从句数量的增加呈指数级上升,而在实际应用中,从句的数量非常大.文献[39]采用一种“需要时再激活”的“惰性”思想,提出了一种 LazySAT 算法,从而很好地解决了该问题. LazySAT 算法利用了一阶知识域中的稀疏性,即大多数闭原子的取值都为假,从而隐性地决定了大多数闭从句的取值都为真.LazySAT 算法对内存的需求量不像传统的 MaxWalkSAT 算法那样由所有的闭原子和闭从句的数量决定,而只由搜索进行当中那些潜在的未满足的闭原子和闭从句决定.

LazySAT 算法[39]以加权的一阶逻辑知识库(weighted first-order KB)为输入,且使用两个队列 *active_atoms* 和 *active_clauses* 来分别保存激活的原子和从句.这两个队列使得 LazySAT 算法比 MaxWalkSAT 算法更有效地利用内存,大大降低了算法的空间复杂度,同时也在一定程度上降低了算法的时间复杂度.然而,LazySAT 算法需要的初始化原子值序列实际上也是一个标准的 MaxWalkSAT 算法的初始化原子值序列,且实验结果表明:当二者的初始化原子值序列相同时,LazySAT 算法的精度并无显著提高,甚至可能与 MaxWalkSAT 算法的结果完全相同.因此,LazySAT 算法除了提高推理速度之外,并未有效地提高推理精度.关于进一步加快推理速度的方法参见文献[40,41],LazySAT 算法的其他扩展参见文献[42].

2.2 边缘概率和条件概率

在概率图模型中,另一种重要的推理形式是计算边缘概率和条件概率,通常采用 MCMC 算法、BP 算法等.对于 Markov 逻辑网,求边缘概率类型的推理问题可表述为:给定 Markov 逻辑网和常量集,求一个规则取值为真的概率.若在求边缘概率中给定了若干其他规则的取值情况,则为条件概率.

由 Markov 逻辑网的定义及条件概率的数学表达式可知,当给定 Markov 逻辑网 L、常量集 C,以及给定规

© 中国科学院软件研究所 http://www.jos.org.cn

则 F_2 取值为真时,规则 F_1 取值为真的概率为

$$P(F_1 \mid F_2, L, C) = \frac{P(F_1 \wedge F_2 \mid M_{L,C})}{P(F_2 \mid M_{L,C})} = \frac{\sum_{X \in \theta_{F_1} \cap \theta_{F_2}} P(X = x \mid M_{L,C})}{\sum_{X \in \theta_{F_2}} P(X = x \mid M_{L,C})} \tag{6}$$

其中,θ_F 表示使规则 F 取值为真的任意可能存在的世界,$P(F_2|M_{L,C})$是所有使 F_2 取值为真的世界存在的概率之和.因此,如果单从公式(6)计算 $P(F_2|M_{L,C})$,那么随着式中闭原子的增加,计算量将呈指数级上升.实际问题所涉及的闭原子往往成千上万,故需寻求解决该问题的其他有效途径.另一个近似但更有效的折中方法是运用 MCMC 算法[43].通常,Markov 逻辑网涉及到的一阶谓词逻辑规则都可转化为从句形式,从而可进一步优化策略[37].

在从句形式下,可以很方便地应用基于知识的模型结构,从而在所生成的闭 Markov 网中只涉及到计算所需的最小部分.在这个最小的闭 Markov 网中,应用 MCMC 方法可以在很大程度上提高推理速度.文献[37]给出了生成该最小闭 Markov 网的算法.该算法改进了传统的闭 Markov 网的生成过程,添加了如下步骤:当查询规则中的原子节点不在证据规则中时,循环加入该原子节点的 Markov 边界包含的原子节点.生成的最小闭 Markov 网将和剩余部分的闭 Markov 网在给定证据规则的条件下形成条件独立状态,从而有效地缩小推理范围.虽然文献[37]指出在最坏情况下该算法无法缩小原来的闭 Markov 网,但在绝大多数实际应用中,该算法都能极大地降低空间和时间复杂度.

事实上,在利用 MCMC,BP 等算法求解概率图模型时,都可能(甚至必然)在一定程度上破坏知识域中数据集的某些确定性关系[44],即 $\omega_i \to \infty$.因为在类似吉布斯采样算法中,采样进行中的各个状态之间必须满足如下关系:

$$\forall x \forall y, P(x)Q(x \to y) = P(y)Q(y \to x) \tag{7}$$

其中,Q 是 Markov 链各个状态之间转换的概率.当数据域中存在某些确定或近似确定关系时,给定某个变量的 Markov 边界,对该变量进行采样,该变量将很难被改变,致使收敛速度非常缓慢,从而极大地增加了计算边缘概率和条件概率的时间复杂度.对于常用的吉布斯采样,一种加快收敛速度的方法是融入模拟退火的思想[45].然而当权重非常大时,在模拟退火过程中互换若干 Markov 链的某些变量状态的操作变得不太现实.当权重趋向无穷大时,就会破坏 Markov 链的遍历特性.文献[46]提出了用于计算条件概率和边缘概率的 MC-SAT 算法.它是融合了切片采样[47]的 MCMC 算法.切片采样在 MCMC 算法族中被广泛采用,其主要思路是:在原有的 MCMC 算法中引入一个辅助参数,以捕捉变量之间的某些确定性关系,并用可满足性解决器和模拟退火步骤来对切片进行采样.文献[48]证明了 MC-SAT 算法严格遵从 MCMC 算法所需要的遍历原则和细节平衡原则.实验结果表明,该算法所得到的概率非常准确.Domingos 研究团队在 Markov 逻辑网实验和开发平台 Alchemy 中,结合了 MC-SAT 算法和上文提到的“惰性”思想,避免了不必要的原子和从句的展开.实验结果表明,“惰性”版本的 MC-SAT 算法是一种更高效的算法.

然而,MC-SAT 算法在有效地处理知识域中确定性关系的同时,却忽略了知识域中的另一种特性,即知识域的稀疏性.MC-SAT 算法在其执行过程中大规模地采用随机选取、概率分流、随机变量赋值等无视知识域稀疏性的方法,实际上添加了很多无效的迭代步骤.如果能够有效地利用知识域的稀疏性,即在大规模采用随机选取、概率分流等方法之前适当采取一系列导向性的判断方式,或在 MC-SAT 算法之前先对知识域进行过滤,可有效地筛选出实用的原子取值分布,从而进一步加快推理算法的执行速度.

3 Markov 逻辑网的学习

3.1 参数学习

参数学习是指在 Markov 逻辑网结构确定的前提下,进一步学习和优化模型的参数.对于 Markov 逻辑网而言,其参数就是公式(3)中的参数 ω_i,这是 Markov 逻辑网体系中的重要研究内容之一.通常情况下,一个模型的参数学习可采用最大似然方法.在 Markov 逻辑网中,一个规则 F_i 的权重 ω_i 的对数似然函数梯度可表示为

$$\frac{\partial}{\partial \omega_i} \log P_\omega(X = x) = n_i(x) - \sum_{x'} P_\omega(X = x') n_i(x') \tag{8}$$

© 中国科学院软件研究所 http://www.jos.org.cn

其中,$n_i(x)$为世界 x 中规则 F_i 的真值个数,$\sum_{x'} P_\omega(X=x')n_i(x')$ 是在所有可能的世界 x'中求和,而 $P_\omega(X=x')$的计算则建立在当前的规则权重向量集$\omega=\{\omega_1,\omega_2,\ldots,\omega_i,\ldots\}$之上.换句话说,规则$F_i$的权重$\omega_i$的对数似然函数梯度,等于当前世界 x 中规则 F_i 的真值个数与在所有可能的世界中规则 F_i 的真值个数的数学期望之差.然而,计算一个规则在世界中的闭规则的个数是不明智的[25,49,50],且直接计算规则 F_i 的真值个数的数学期望更是不可行.因此,直接用最大似然方法实际上并不可行.目前,替代最大似然估计的参数学习方法主要有两种,即伪最大似然估计和判别训练[25].

3.1.1 伪最大似然估计

文献[1]中的伪最大似然估计的最大似然函数可描述为

$$P_\omega^*(X=x)=\prod_{l=1}^{n} P_\omega(X_l=x_l \mid MB_x(X_l)) \tag{9}$$

其中,x_l 表示第 l 个闭原子的真值,$MB_x(X_l)$表示 X_l 的 Markov 边界.计算伪最大似然估计及其梯度无需推理,因此学习速度将显著加快.文献[48]将伪最大似然估计和 L-BFGS 优化器结合起来,即使在一个有上百万闭原子的数据库中,伪最大似然估计的学习效率也相当不错.然而,采用伪最大似然估计会导致非邻接变量之间的推理结果不甚理想.

3.1.2 判别训练

训练 Markov 逻辑网所用的数据集实际上是一个关系数据库,即在封闭世界假设的前提下,训练所用的数据集是一组真值闭原子的集合.所谓的封闭世界是指在数据库中未出现的闭原子的值都定为假.在许多情形下,人们对数据库中的数据会有一种先验知识,即知道哪些数据经常会成为证据数据,哪些数据经常会成为查询数据,且主要目标就是为了正确地预测查询数据.因此,很自然地将数据库分割成两个集合,即证据谓词 X 和查询谓词 Y.通过对条件概率

$$P_\omega(y \mid x)=\frac{1}{Z_x}\exp\left(\sum_{i\in F_Y}\omega_i n_i(x,y)\right) \tag{10}$$

求最大似然来学习权重ω_i.其中,F_Y 是一个从句集合,该集合中的每个从句至少包含 1 个闭查询谓词;$n_i(x,y)$是数据库中第 i 个从句的闭从句的真值个数.对公式(10)的条件对数似然函数(conditional log-likelihood,简称 CLL)求偏导,可得:

$$\frac{\partial}{\partial\omega_i}\log P_\omega(y \mid x)=n_i(x,y)-E_\omega[n_i(x,y)] \tag{11}$$

由公式(11)可知,当数据库中第 i 个从句的闭从句的实际真值个数大于它本身在所有世界中真值个数的期望时,ω_i 的值将在本次学习迭代中提高,反之则会降低.事实上,精确地求 $E_\omega[n_i(x,y)]$的时间复杂度非常巨大.

训练 Markov 逻辑网权重的高效算法主要包括 VP(voted perceptron)算法、CD(contrastive divergence)算法等[28].下面分别简要论述 VP 算法、CD 算法和其他学习算法.

(1) Voted Perceptron(VP)算法

VP 算法的基本思想是:用 MAP(maximum a posteriori probability)状态下的计数值来近似公式(11)中的 $E_\omega[n_i(x,y)]$.它通过捕捉 $y_\omega(x)$的 MAP 状态 $y_\omega^*(x)$ 来近似计算 $n_i(x,y_\omega^*)$.当概率 $P_\omega(y|x)$的分布集中在 $y_\omega^*(x)$ 附近时,这个近似估计的效果非常理想.VP 算法将所有权重初始化为 0,然后进行 T 次梯度下降,最终返回每次迭代的权重之和的平均值,即 $\omega=\sum_{t=1}^{T}\omega_{i,t}/T$.VP 算法最初是在 Viterbi 算法捕捉 MAP 状态之后用来处理隐马尔可夫模型的.Singla 和 Domingos 拓展了 VP 算法,用 MaxWalkSAT 算法取代 Viterbi 算法用以捕捉 MAP 状态,然后引入到 Markov 逻辑网学习体系中,用来训练规则的权重[25].然而值得注意的是,MaxWalkSAT 算法并不能保证寻找到全局的 MAP 状态,这是权重学习中潜在的误差源头.因此,为了提高估算精度,通常在 MaxWalkSAT 算法返回的状态上再进行吉布斯采样,然后对产生的所有样本进行重新评估(实际上大多取均值).

然而,在 Markov 逻辑网体系中应用 VP 算法有个缺陷,即在 Markov 逻辑网体系中由于不同闭原子值代入同一原子变量,使得 VP 算法的迭代次数非常多.由于在 Markov 逻辑网体系中 MAP 状态实际上不是唯一的,而 MaxWalkSAT 算法并不能找到所有的 MAP 状态,所以 VP 算法在运行时也可能遭遇困难.

© 中国科学院软件研究所 http://www.jos.org.cn

(2) Contrastive Divergence(CD)算法

CD 算法[28]和 VP 算法比较类似,其区别在于,CD 算法并没有捕捉原子的 MAP 状态,而是用 MCMC 算法产生的一部分样本取代 MAP 状态.由于 MCMC 算法最终可收敛到真实的期望,因此 CD 算法得到的结果更加准确和稳定.然而,由于 MCMC 算法的收敛速度可能会非常慢,因此更实际的做法是用 MC-SAT 算法取代 CD 算法中的 MCMC 算法步骤.这样做的另一个好处是,用 MC-SAT 算法产生的样本会比用 MCMC 算法产生的样本更为独立,并且携带更多有用信息,收敛也更快.

VP 算法和 CD 算法都是简单的梯度下降参数学习方法,这两种算法极易受到病态问题的影响.在 Hessian 式中,当最大绝对特征值与最小绝对特征值的比例小于 1 时就会出现病态问题.在病态问题中,采用梯度下降的参数学习方式,参数的收敛速度将会非常缓慢,因为没有一个合适的学习速率可以适应病态问题.在 Markov 逻辑网中,Hessian 式即为从句的协方差矩阵.由于有些从句所对应的闭从句的真值个数远大于其他从句的闭从句的真值个数,而 Hessian 式中最大绝对特征值会由于该从句的闭从句的真值个数迅速膨胀,故病态问题可能变得更加严重.病态问题的主要处理方式是,在梯度下降的参数学习方式中不断调整学习速率,从而达到缓和病态问题的目的.

(3) 其他学习算法

DN(diagonal Newton)算法[28]是对 Newton 算法的一种改进,Newton 算法使用逆 Hessian 矩阵进行梯度下降来实现二次函数的优化问题,而 DN 算法使用对角 Hessian 矩阵代替逆 Hessian 矩阵来实现梯度下降,这样可用于 Markov 逻辑网的 CLL 的非二次函数的优化问题.Markov 逻辑网的负 CLL 的 Hessian 矩阵(即协方差矩阵)元素形式为

$$\frac{\partial^2(-\log P_w(y\mid x))}{\partial w_i \partial w_j} = E_w[n_i n_j] - E_w[n_i]E_w[n_j] \tag{12}$$

采用类似梯度的处理方式,公式(12)可通过 MC-SAT 的样本估计出来.按照梯度下降法权重学习的迭代公式为

$$\omega_i = \omega_i - \alpha \frac{E_\omega[n_i] - n_i}{E_\omega[n_i^2] - (E_\omega[n_i])^2} \tag{13}$$

其中,步长α的取值可固定.但通常为了保证精度,给定搜索方向 d 和 Hessian 矩阵 H,α的取值计算公式为

$$\alpha = \frac{-d^\tau g}{d^\tau H d + \tau d^\tau d} \tag{14}$$

其中,τ值在每次迭代后均需调整,其作用是将学习步长限制在二次近似较为准确的区间内进行结构学习,从而可以灵活地调节模型置信区间的大小.然而,在 Markov 逻辑网体系中从句之间的关联十分强烈的情况下,DN 算法却不够有效.

SCG(scaled conjugate gradient)[28]算法是对共轭梯度算法的改进,它使用 Hessian 矩阵而非直线搜索来选择梯度下降中的步长.在 Markov 逻辑网的参数学习中,SCG 算法采用和 DN 算法一样的步长计算方法,如公式(14)所示.

由于判别训练算法的前提是假设数据集可分割为证据谓词集合和查询谓词集合,故该类算法也有其内在缺陷.例如,在基于 Markov 逻辑网的实体解析(entity resolution)中,判别训练在参数学习时往往不能正确处理矛盾规则.事实上,若参数 k 涉及的规则集中包含矛盾规则对,判别训练在学习参数 k 时仅仅随机抓取了矛盾规则对中的某一个规则,却忽略了与之对应的另一个规则,因而参数 k 就不能正确反映知识域的实际情况.一般地,在随后的实体解析过程中,必须经过调整才能有效地解决该问题.

3.2 结构学习

结构学习是指从数据(库)中学出某个模型尽可能合理的结构,就 Markov 逻辑网的结构学习而言,就是从数据(库)中学出最优或次优的网络结构.结构学习一直是人工智能、机器学习等领域公认的研究难点,一种出色的

© 中国科学院软件研究所 http://www.jos.org.cn

结构学习算法对于模型构建的意义十分巨大[51].简而言之,Markov 逻辑网的结构就是需要附加权重的那些规则的集合.一般地,可用任意的归纳逻辑程序设计(inductive logic programming,简称 ILP)技术来学习或修改结构,然而由于 Markov 逻辑网表示的是一个概率分布,而传统的 ILP 采用精确性和覆盖范围作为评价函数往往得不到很好的效果,所以在 Markov 逻辑网的结构学习中,常用对数似然估计或条件对数似然估计作为评价函数,这样可以得到更好的结果.当然,不足之处是这些评价标准的计算代价非常大.在数据集上的实验结果表明,与传统的结构学习算法相比,Markov 逻辑网的结构搜索算法能找到更好的 Markov 逻辑网规则.下面主要介绍评价标准和两种最优化伪似然估计的结构学习算法.

3.2.1 评价标准

如上所述,结构学习主要包括两个难题:一是如何搜索潜在的结构,大部分算法的差异性也就在此;二是如何为搜索到的结构建立起一个评价标准,即如何筛选出最优结构.目前公认较好的评价标准是 Domingos 研究团队设计的带权重的伪对数似然估计(weighted pseudo-log-likelihoood,简称 WPLL):

$$\log P_w^{\bullet}(X=x)=\sum_{r\in R}c_r\sum_{k=1}^{g_r}\log P_w(X_{r,k}=x_{r,k}\mid MB_x(X_{r,k})) \tag{15}$$

其中,R 是一阶原子的集合,g_r 是一阶原子 r 的常量取值的个数,$x_{r,k}$ 是一阶原子 r 的第 k 个常量取值的真值.每个原子的权重 C_r 取决于不同的问题,在默认情况下,可简单地设 $C_r=1/g_r$,这意味着所有一阶谓词的权重均相同.若在建模过程中发现有些谓词不重要,则将其所对应的权重设为 0.

3.2.2 自顶而下的结构学习

自顶而下的结构学习(top-down structure learning,简称 TDSL)[24]每次只学习或修改 Markov 逻辑网中的一个从句.可以用一个空的网络或已有的知识库作为算法初始化的结构,且最好在一开始就把所有的单个原子全部加入到需要学习的 Markov 逻辑网里.为寻找最优结构,必须不断地修改模型,TDSL 有两种搜索策略:一种使用 Beam 搜索寻找最优的从句,然后加入 Markov 逻辑网;另一种是在搜索长度为 l+1 的从句之前,把所有长度为 l 的较优从句加入 Markov 逻辑网.在数据集上的实验结果表明,TDSL 结构学习方法的能力大大超越了以往的 ILP 系统及一些纯概率方法(如朴素贝叶斯算法等).

3.2.3 自底而上的结构学习

事实上,上述 TDSL 结构学习方法也有其不足,因为 Markov 逻辑网的搜索空间随着问题规模的增大而呈指数增长,而且 TDSL 方法的两种搜索策略并非完美,再加上生成测试从句的过程是独立的,故求解过程很容易收敛到局部最优解.此外,尽管 TDSL 方法无须全面搜索,但其效率仍然太低.

自底而上的结构学习(bottom-up structure learning,简称 BUSL)[29]克服了上述缺点,它通过对训练数据的考察而直接建立或修改已有的结构模型.实验结果表明,BUSL 可以比 TDSL 更快地得到更精确的模型.BUSL 通过训练数据自身的特点,建立模板网络来控制候选从句的生成.模板网络由模板节点组成,而模板节点则是通过寻找拥有一组共同常量且同时为真的闭原子,然后将其中的常量抽象替换成变量而生成的.随后不断地生成模板节点,同时使用带权重的伪对数似然估计作为评价标准将好的结构加入模型.

4 算法比较和分析

4.1 与基于Bayesian网的统计关系学习算法比较

基于 Bayesian 网的 SRL 算法是在传统 Bayesian 网的基础上进行扩展的 SRL 方法,其众多模型均采用图型结构和逻辑子句结构表示数据及数据间的关系.前者最典型、最有代表性的方法是概率关系模型(probabilistic relational models,简称 PRM)[3],而后者则是贝叶斯逻辑程序模型(Bayesian logic programs,简称 BLP)[52].与 PRM 和 BLP 相比,Markov 逻辑网有 3 大优势:(1) Markov 逻辑网能够在对象实体和关系不断变化中自动调整其网络结构;(2) Markov 逻辑网能够动态地表示关系优先级的变化,即通过自动更新 Markov 网来调整规则的权重;(3) Markov 逻辑网能够正确识别同一实体的多个符号表示,即对象的动态识别.Markov 逻辑网的这 3 个优点是 PRM 算法和 BLP 算法均不具备的.

© 中国科学院软件研究所 http://www.jos.org.cn

4.2 与基于随机文法的统计关系学习算法比较

基于随机文法的统计关系学习方法是将随机文法提升到一阶逻辑的一种方法,主要包括随机逻辑程序(stochastic logic program,简称 SLP)[53]和统计建模程序设计(programming in statistical modeling,简称 PRISM)[54].这类方法通过对逻辑成分附加概率来处理关系和不确定性.该类方法以逻辑程序来描述模型结构,因而表示能力较强.随机文法和 Markov 逻辑网有一定的相似性,即都是通过对逻辑成分附加概率来处理关系和不确定性.然而,基于随机文法的统计关系学习算法对数据的容错性极差.其主要原因是,随机文法的逻辑成分结构通常较为庞大,与 Markov 逻辑网相比,其推理和学习算法较为呆板,因而碰到质量较差的数据时,基于随机文法的统计关系学习算法表现差强人意,而 Markov 逻辑网受到的影响则较小.

4.3 与基于HMM的统计关系学习算法比较

众所周知,HMM 是在 Markov 模型的基础上进行了扩展,是一种典型的双重随机过程,可以有效地解决动态序列化问题.然而 HMM 有一个很大的缺陷,就是 HMM 在处理数据时有一个前提:对所有的数据块进行连续观测时,数据块之间是独立的(而 Markov 逻辑网却没有此问题).事实上,所有的 HMM 都必须基于这个假设.此外,由 Markov 性(即 t 时刻的状态完全取决于 t–1 时刻的状态)可知,HMM 算法在某些实际应用中(如语音识别等)[33]仍存在不足.而 Markov 逻辑网在其推理和学习算法中有意淡化了这种假设(如通过引入模拟退火算法等),因而能够在这些应用方面上更加游刃有余.

5 Markov 逻辑网的应用

5.1 应用概况

当前,Markov 逻辑网备受国内外人工智能、机器学习等领域研究学者的广泛关注[55].多年来,Domingos 研究团队不断地完善 Markov 逻辑网的理论体系[50,56],并提供了一个学习和发展该理论体系的平台 Alchemy.Markov 逻辑网作为统计关系学习的统一框架,其应用前景十分广阔.Domingos 研究团队在实体解析[26]、信息抽取[57]、社会网络[58]、语义网络[31,35,59]、自然语言处理[60]、分子生物学[34]等方面的实践充分证明了 Markov 逻辑网的实用价值.

国际上,其他研究团队也将 Markov 逻辑网应用到了自然语言处理、地理信息系统和计算机视觉等方面.在自然语言处理领域,Fiona 等人[61]利用 Markov 逻辑网进行文本知识抽取和语义理解,Yu等人[62,63]利用 Markov 逻辑网进行中文命名实体识别(Chinese named entity recognition,简称 Chinese NER),并取得了较好的效果,Cheng 等人[64]将 Markov 逻辑网运用于主题发现,Aron 等人[65]利用 Markov 逻辑网来处理信息抽取中的指代消解问题.在地理信息系统领域,Lin[66]利用 Markov 逻辑网来识别基于位置的活动.在计算机视觉领域,Markov 逻辑网也显出其应用优势[67].

在国内,吉林大学刘大有教授的科研团队较早跟踪 Markov 逻辑网的研究,其发表的论文为国内有志进入该领域的研究者积累了十分宝贵的经验[2,68].于鹏等人[69]针对 Markov 逻辑网学习中采用确定性优化方法时经常出现所求的解不够简洁且易陷入局部极值问题,定义了谓词模板、子句模板以缩小搜索空间,提出一种采用子句模板编码的遗传算法来学习 Markov 逻辑网的结构,并用粒子群算法(PSO)学习 Markov 逻辑网的权参数.孙舒杨等人[70]针对伪似然估计方法存在的主要不足,即虽然降低了计算复杂度但没有考虑参数的先验分布因素,提出了一种基于后验概率的参数估计方法.该方法采用正态先验分布,用伪似然概率替代似然概率,通过最大化伪后验概率来学习 Markov 逻辑网的模型参数.张玉芳等人[71]结合了判别式训练的学习算法以及 MC- SAT、吉布斯采样和模拟退火等推理算法,将 Markov 逻辑网应用于中文文本分类.韩磊等人[72]在进行人体行为分析时,特别是在对两人交互行为进行语义建模的过程中,人工建立了一个表示领域知识的一阶逻辑知识库,并训练 Markov 逻辑网用于两人交互行为的推理.楼俊杰等人[73]在 Domingos 等人提出的基于 Markov 逻辑网的实体解析算法基础上,进一步引入一个可变权重的规则,从而有效地解决了该算法无法处理实体二义性的问题.

© 中国科学院软件研究所 http://www.jos.org.cn

5.2 应用举例

为了更具体地说明 Markov 逻辑网的应用方法,下面以垃圾邮件过滤为例,简要介绍 Markov 逻辑网的应用过程.我们在 Domingos 等人研制的 Markov 逻辑网应用平台——Alchemy 系统上,使用 Python 编程语言,对垃圾邮件过滤数据集 Spamassassin 进行垃圾邮件过滤实验.主要步骤如下:

步骤 1. 构建垃圾邮件过滤的 Markov 逻辑表示,即构建“.mln”文件,主要内容有:

(a) HasToken(token,mail);

(b) Spam(mail);

(c) HasToken(+token,mail)→Spam(mail).

步骤 2. 构建邮件训练数据集,即构建训练用的“.db”文件.对于训练数据集中的每封邮件 mailx,如果属于垃圾邮件,则往该文件中添加文本 Spam(mailx);否则,添加文本!Spam(mailx),同时把邮件中含有的 token 以文本 HasToken(token,mailx)的形式加入该文件中.同时,构建邮件测试数据集,即构建测试用的“.db”文件.对于测试数据集中的每封邮件 maily,把邮件中含有的 token 以文本 HasToken(token, maily)的形式加入该文件中.

步骤 3. 权重学习.使用 Alchemy 系统中的 learnwts 命令对训练数据集进行权重学习,获得含有公式权重的“.mln”文件.

步骤 4. 邮件分类测试.根据上一步学习到的权重,使用 Alchemy 系统中的 infer 命令对测试数据集进行分类,获得含有各个邮件为垃圾邮件(Spam)的概率文件.

步骤 5. 设定阈值,将概率超过该阈值的邮件判定为垃圾邮件(Spam),否则为正常邮件(Ham).

6 述 评

不可否认,人工智能经过 50 多年的曲折发展,如今已取得了举世瞩目的巨大成功.然而必须清醒地看到,当前人工智能、机器学习、模式识别、数据挖掘等领域中依然存在着很多鸿沟和困境.国际权威 AI 专家、加州大学 Berkeley 分校计算机系主任 Stuart Russell 教授在 1995 年获得 IJCAI-1995 颁发的“计算机与思维奖(Computers and Thought Award)”时,在其所做的学术报告“Rationality and Intelligence”[74]中特别指出:“AI 是一个由其研究的问题而非方法所定义的领域”.他认为,在未来相当长时间内,AI 等领域中存在的最大问题是如何填补基于抽象、非底层表示(ungrounded representations)的高层推理(high-level reasoning)与建立底层表示(grounded representations)的传感数据解释(interpreting raw sensor data)之间的巨大鸿沟.类似地,国际机器学习权威、Stanford 大学计算机系教授 Daphne Koller 在 2001 年获得 IJCAI-2001 颁发的“计算机与思维奖”时所做的学术报告中强调指出,传统 AI 中被广泛采用的分析、分解方法正面临着巨大挑战,人们在解决复杂问题时往往很自然地采用分而治之的方法将其分解为每个小片,希望等到每个小片都取得长足进展后再试图进行综合集成,以得到所期望的最终结果.但非常遗憾的是,人工智能、机器学习、模式识别、数据挖掘等相关领域的研究现状却表明事与愿违,因为事实上,每个子问题都各自分家且相互远离,而且是离得越来越远,结果造成 AI 的各个研究分支支离破碎、分崩离析,最后,显然很难将它们有机地综合集成起来.

为了突破这个困境,国际上很多权威 AI 学者提出了不少设想,并进行了有益的尝试.其中,Russell 教授在文献[38]中就基于 Agents 及多 Agents 系统(MAS)的观点,将 AI 领域的所有研究内容统括在一个完整的 MAS 体系中.Koller 教授围绕着如何填补上述 AI 鸿沟问题,提出建立连接的 3 座概念桥梁[75],即表示、推理和学习,她强调重点以概率图模型来达成该目的.另一种代表性的解决方法是美国德克萨斯大学奥斯汀分校的 Peter Stone 教授在获得 IJCAI-2007“计算机与思维奖”时所做的学术报告[76]中指出,应该建立一种完全自主的 Agents,这些 Agents 具有高度的鲁棒性和灵活性,它们可以感知环境,进行高层认知和决策,在环境中进行自主执行,即具有学习、交互、组合及合作等能力.他认为,这种研究方法可以分为两条基本路线:一是基本算法研究,主要包括机器学习、多 Agents 系统;二是应用研究,主要包括实现面向特定的复杂环境的完全自主 Agents,以及从面向特定应用的自主 Agents 实现中总结发现普遍规律.

© 中国科学院软件研究所 http://www.jos.org.cn

不难发现,本文重点论述的美国华盛顿大学 Domingos 教授研究团队所提出的 Markov 逻辑网也是一种独具特色的弥补 AI 鸿沟的重要途径,其主要贡献是:(1) 将传统的一阶谓词逻辑与当前主流的统计学习方法有机地结合起来.这样做的最大优点是,既能让传统的一阶谓词逻辑、产生式规则、专家系统及其开发工具等重焕生机,又能充分利用当前基于概率统计的机器学习等方法的最新成果,而且将它们较完美地集成起来以形成优势互补,从而有力地促进了人工智能、机器学习、模式识别、数据挖掘等领域中各种不同方法的融合与发展;(2) 有效地填补了 AI 等领域中存在的高层与底层之间的巨大鸿沟.一般来说,一阶谓词逻辑更适用于高层知识的表示与推理,而基于概率统计的机器学习方法则擅长于对底层数据进行统计学习.然而长期以来,这两种方法一直都未能有机地融合起来,它们都在朝着不同的方向发展,而且有一种渐行渐远的趋势.因此,Markov 逻辑网对促进这两个重要研究方向的有机融合提供了十分宝贵的经验和借鉴.

Markov 逻辑网结合了一阶逻辑和概率图模型,发展出一系列的推理和学习算法,并且已成功应用于许多领域.然而,Markov 逻辑网的理论还不成熟,还有很大的完善空间.将来可以从以下几个方面来进行深入研究:(1) 增强算法的学习能力,使其可以从缺值数据中学习;(2) 提高真值闭从句的计算速度,解决结构学习算法效率的瓶颈问题;(3) 从一阶逻辑和 Markov 网这两个方面完善 Markov 逻辑网的理论;(4) 增强 Markov 逻辑网模型的实用性,从而更好地解决实际应用问题.

致谢 谨向对本文提出宝贵建议的匿名审稿专家、编辑以及参与本文讨论和修改的王帅、吴宇明、陈雅芳、弭真真、刘强、王铖微、王海涛等研究生表示衷心的感谢!

References:

[1] Domingos P, Richardson M. Markov logic: A unifying framework for statistical relational learning. In: Proc. of the ICML-2004 Workshop on Statistical Relational Learning and its Connections to Other Fields. Banff, 2004. 49−54.

[2] Liu DY, Yu P, Gao Y, Qi H, Sun SY. Research progress in statistical relational learning. Journal of Computer Research and Development, 2008,45(12):2110−2119 (in Chinese with English abstract).

[3] Koller D. Probabilistic relational models. In: Proc. of the 9th Int'l Workshop on Inductive Logic Programming (ILP'99). LANI 1634, Berlin: Springer-Verlag, 1999. 3−13.

[4] Jordan MI. Graphical models. Statistical Science (Special Issue on Bayesian Statistics), 2004,19(1):140−155.

[5] Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques. Cambridge: The MIT Press, 2009.

[6] Baral C, Gelfond M. Logic programming and knowledge representation. Journal of Logic Programming, 1994,19:73−148. [doi: 10.1016/0743-1066(94)90025-6]

[7] Kowalski RA. The early years of logic programming. Communications of the ACM, 1988,31(1):38−43. [doi: 10.1145/35043.35046]

[8] Lloyd JW. Foundations of Logic Programming. New York: Springer-Verlag, 1984.

[9] Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF. The Description Logic Handbook: Theory, Implementation, Applications. Cambridge: Cambridge University Press, 2003.

[10] Ben-Gal I. Bayesian networks. In: Ruggeri F, Kenett R, Faltin F, eds. Proc. of the Encyclopedia of Statistics in Quality and Reliability. New York: John Wiley & Sons, 2007.

[11] Jensen FV, Nielsen TD. Bayesian Networks and Decision Graphs. 2nd ed., New York: Springer-Verlag, 2007.

[12] Zhang LW, Guo HP. Introduction to Bayesian Networks. Beijing: Science Press, 2006 (in Chinese).

[13] Cappé O, Moulines E, Rydén T. Inference in Hidden Markov Models. New York: Springer-Verlag, 2005.

[14] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE, 1989,77(2):257−286. [doi: 10.1109/5.18626]

[15] Bellman RE. A Markovian decision process. Journal of Mathematics and Mechanics, 1957,6(5):679−684.

[16] Bellman RE. Dynamic Programming. Princeton: Princeton University Press, 1957.

[17] Puterman ML. Markov Decision Processes: Discrete Stochastic Dynamic Programming. New York: John Wiley & Sons, 1994.

[18] Arbib MA. The Handbook of Brain Theory and Neural Networks. Boston: MIT Press, 2003.

© 中国科学院软件研究所 http://www.jos.org.cn

[19] Bertsekas DP, Tsitsiklis JN. Neuro-Dynamic Programming. Belmont: Athena Scientific, 1996.

[20] Horn A. On sentences which are true of direct unions of algebras. Journal of Symbolic Logic, 1951,16(1):14−21. [doi: 10.2307/2268661]

[21] Domingos P, Kok S, Poon H, Richardson M, Singla P. Unifying logical and statistical AI. In: Proc. of the 21st National Conf. on Artificial Intelligence (AAAI 2006). Boston, 2006. 2−7.

[22] Domingos P. What's missing in AI: The interface layer. In: Cohen P, ed. Proc. of the Artificial Intelligence: The 1st Hundred Years. Menlo Park: AAAI Press, 2006. http://www.cs.washington.edu/homes/pedrod/papers/a

[23] Kok S, Domingos P. Statistical predicate invention. In: Proc. of the 24th Int'l Conf. on Machine Learning (ICML 2007). Corvallis, 2007. 433−440. [doi: 10.1145/1273496.1273551]

[24] Kok S, Domingos P. Learning the structure of Markov logic networks. In: Proc. of the 22nd Int'l Conf. on Machine Learning (ICML 2005). Bonn, 2005. 441−448. [doi: 10.1145/1102351.1102407]

[25] Singla P, Domingos P. Discriminative training of Markov logic networks. In: Proc. of the 20th National Conf. on Artificial Intelligence (AAAI 2005). Pittsburgh, 2005. 868−873.

[26] Singla P, Domingos P. Entity resolution with Markov logic. In: Proc. of the 6th IEEE Int'l Conf. on Data Mining (ICDM 2006). Hong Kong, 2006. 572−582. [doi: 10.1109/ICDM.2006.65]

[27] Singla P, Domingos P. Markov logic in infinite domains. In: Proc. of the 23rd Conf. on Uncertainty in Artificial Intelligence (UAI 2007). Vancouver, 2007. 368−375.

[28] Lowd D, Domingos P. Efficient weight learning for Markov logic networks. In: Proc. of the 11th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2007). Warsaw, 2007. 200−211. [doi: 10.1007/978-3-540-74976-9_21]

[29] Mihalkova L, Mooney RJ. Bottom-Up learning of Markov logic network structure. In: Proc. of the 24th Int'l Conf. on Machine Learning (ICML 2007). Corvalis, 2007. 625−632. [doi: 10.1145/1273496.1273575]

[30] Wang J, Domingos P. Hybrid Markov logic networks. In: Proc. of the 23rd National Conf. on Artificial Intelligence (AAAI 2008). Chicago, 2008. 1106−1111.

[31] Kok S, Domingos P. Extracting semantic networks from text via relational clustering. In: Proc. of the 19th European Conf. on Machine Learning (ECML 2008). Antwerp, 2008. 624−639. [doi: 10.1007/978-3-540-87479-9_59]

[32] Kok S, Domingos P. Learning Markov logic network structure via hypergraph lifting. In: Proc. of the 26th Int'l Conf. on Machine Learning (ICML 2009). Montreal, 2009. 505−512. [doi: 10.1145/1553374.1553440]

[33] Nath A, Domingos P. A language for relational decision theory. In: Proc. of the 6th Int'l Workshop on Statistical Relational Learning (SRL 2009). Leuven, 2009. http://www.cs.kuleuven.be/~dtai/ilp-mlg-srl/papers/SRL09-3.pdf

[34] Davis J, Domingos P. Deep transfer via second-order Markov logic. In: Proc. of the 20th Int'l Conf. on Machine Learning (ICML 2009). Montreal, 2009. 217−224. [doi: 10.1145/1553374.1553402]

[35] Poon H, Domingos P. Unsupervised semantic parsing. In: Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing (EMNLP 2009). Singapore, 2009. 1−10.

[36] Kok S, Domingos P. Learning Markov logic networks using structural motifs. In: Proc. of the 27th Int'l Conf. on Machine Learning (ICML 2010). Haifa, 2010. 551−558.

[37] Richardson M, Domingos P. Markov logic networks. Machine Learning, 2006,62(1-2):107−136.

[38] Russell S, Norvig P. Artificial Intelligence: A Modern Approach. 2nd ed., Upper Saddle River: Pearson Education, Inc., 2003.

[39] Singla P, Domingos P. Memory-Efficient inference in relational domains. In: Proc. of the 21st National Conf. on Artificial Intelligence (AAAI 2006). Boston, 2006. 488−493.

[40] Gogate V, Webb WA, Domingos P. Learning efficient Markov networks. In: Advances in Neural Information Processing Systems (NIPS 2010). Vancouver, 2010. 748−756.

[41] Lowd D, Domingos P.Approximate inference by compilation to arithmetic circuits. In: Advances in Neural Information Processing Systems (NIPS 2010). Vancouver, 2010. 1477−1485.

[42] Poon H, Domingos P, Sumner M. A general method for reducing the complexity of relational inference and its application to MCMC. In: Proc. of the 23rd National Conf. on Artificial Intelligence (AAAI 2008). Chicago, 2008. 1075−1080.

[43] Gilks WR, Richardson S, Spiegelhalter DJ. Markov Chain Monte Carlo in Practice. London: Chapman and Hall, 1996.

© 中国科学院软件研究所 http://www.jos.org.cn

[44] Yedidia JS, Freeman B, Weiss Y. Generalized belief propagation. In: Leen T, Dietterich T, Tresp V, eds. Advances in Neural Information Processing Systems (NIPS 2001). Cambridge: MIT Press, 2001. 689−695.

[45] Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science, 1983,220(4598):671−680. [doi: 10.1126/science.220.4598.671]

[46] Poon H, Domingos P. Sound and efficient inference with probabilistic and deterministic dependencies. In: Proc. of the 21st National Conf. on Artificial Intelligence (AAAI 2006). Boston, 2006. 458−463.

[47] Neal RM. Slice sampling. The Annals of Statistics, 2003,31(3):705−767. [doi: 10.1214/aos/1056562461]

[48] Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Mathematical Programming, 1989,45(1-3): 503−528. [doi: 10.1007/BF01589116]

[49] Milch B, Zettlemoyer LS, Kersting K, Haimes M, Kaelbling LP. Lifted probabilistic inference with counting formulas. In: Proc. of the 23rd National Conf. on Artificial Intelligence (AAAI 2008). Chicago, 2008. 1062−1068.

[50] Singla P, Domingos P. Lifted first-order belief propagation. In: Proc. of the 23rd National Conf. on Artificial Intelligence (AAAI 2008). Chicago, 2008. 1094−1099.

[51] Dietterich TG, Domingos P, Getoor L, Muggleton S, Tadepalli P. Structured machine learning: The next ten years. Machine Learning, 2008,73(1):3−23.

[52] Kersting K, De Raedt LD. Adaptive Bayesian logic programs. In: Proc. of the 11th Conf. on Inductive Logic Programming. LNCS 2157, Berlin: Springer-Verlag, 2001. 104−117. [doi: 10.1007/3-540-44797-0_9]

[53] Muggleton S. Stochastic logic programs. In: Proc. of the 5th Int'l Workshop on Inductive Logic Programming. Amsterdam: IOS Press, 1996. 254−264.

[54] Sato T, Kameya Y. PRISM: A language for symbolic-statistical modeling. In: Proc. of the 15th Int'l Joint Conf. on Artificial Intelligence. San Francisco: Mrogan Kaufmann Publishers, 1997. 1330−1339.

[55] Domingos P, Lowd D. Markov Logic: An Interface Layer for Artificial Intelligence. San Rafael: Morgan and Claypool, 2009.

[56] Lowd D, Domingos P. Recursive random fields. In: Proc. of the 20th Int'l Joint Conf. on Artificial Intelligence (IJCAI 2007). Hyderabad, 2007. 950−955.

[57] Poon H, Domingos P. Joint inference in information extraction. In: Proc. of the 22nd National Conf. on Artificial Intelligence (AAAI 2007). Vancouver, 2007. 913−918.

[58] Singla P, Kautz H, Luo JB, Gallagher A. Discovery of social relationships in consumer photo collections using Markov logic. In: Proc. of the CVPR Workshop on Semantic Learning and Applications in Multimedia. Anchorage, 2008. 1−7. [doi: 10.1109/CVPRW.2008.4563047]

[59] Domingos P, Lowd D, Kok S, Poon H, Richardson M, Singla P. Just add weights: Markov logic for the semantic Web. In: Costa PCG, d'Amato C, Fanizzi N, Laskey KB, Laskey KJ, Lukasiewicz T, Nickles M, Pool M, eds. Proc. of the Uncertain Reasoning for the Semantic Web I. New York: Springer-Verlag, 2008. [doi: 10.1007/978-3-540-89765-1_1]

[60] Poon H, Domingos P. Joint unsupervised coreference resolution with Markov logic. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP 2008). Honolulu, 2008. 650−659.

[61] McNeill F, Halpin H, Klein E, Bundy A. Merging stories with shallow semantics. In: Proc. of the Workshop on Knowledge and Reasoning for Language Processing (KRAQ 2006). Trento, 2006. 37−42.

[62] Yu XF. Chinese named entity recognition with cascaded hybrid model. In: Proc. of Human Language Technologies 2007: The Conf. of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2007). Rochester, 2007. 197−200.

[63] Yu XF, Lam W, Chan SK. A framework based on graphical models with logic for Chinese named entity recognition. In: Proc. of the 3rd Int'l Joint Conf. on Natural Language Processing. Hyderabad, 2008. 335−342.

[64] Cheng V, Li CH. Topic detection via participation using Markov logic network. In: Proc. of the Signal-Image Technologies and Internet-Based System (SITIS 2007). Shanghai, 2007. 85−91. [doi: 10.1109/SITIS.2007.55]

[65] Culotta A, Wick M, Hall R, McCallum A. First-Order probabilistic models for coreference resolution. In: Proc. of the Human Language Technologies 2007: The Conf. of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2007). Rochester, 2007. 81−88.

[66] Lin L. Location-Based activity recognition [Ph.D. Thesis]. Washington: University of Washington, 2006.

© 中国科学院软件研究所 http://www.jos.org.cn

[67] Tran SD, Davis LS. Event modeling and recognition using Markov logic networks. In: Proc. of the 10th European Conf. on Computer Vision (ECCV 2008). Marseille, 2008. 610−623. [doi: http://dx.doi.org/10.1007/978-3-540-88688-4_45]

[68] Sun SY, Liu DY, Sun CM, Huang GL. Survey of Markov logic networks. Application Research of Computers, 2007,24(2):1−3 (in Chinese with English abstract).

[69] Yu P, Liu DY, Ouyang DT. Research on learning Markov logic networks based on GA and PSO. Acta Electronica Sinica, 2006, 34(12A):2551−2555 (in Chinese with English abstract).

[70] Sun SY, Liu DY, Sun CM. Pseudo-Posterior parameters learning of Markov logic networks. Journal of Jilin University (Science edition), 2006,44(6):946−950 (in Chinese with English abstract).

[71] Zhang YF, Huang T, Ai DM, Xiong ZY. Markov logic network and its application in text classification. Journal of Computer Applications, 2009,29(10):2729−2732 (in Chinese with English abstract). [doi: 10.3724/SP.J.1087.2009.02729]

[72] Han L, Li JF, Jia YD. Human interaction recognition using spatial-temporal works. Chinese Journal of Computers, 2010,33(4): 776−784 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2010.00776]

[73] Lou JJ, Xu CF, Hao CL. Improvement of entity resolution based on Markov logic networks. Computer Science, 2010,37(8): 243−247 (in Chinese with English abstract).

[74] Russell S. Rationality and intelligence. In: Proc. of the 14th Int'l Joint Conf. on Artificial Intelligence (IJCAI'95). Montreal, 1995. 950−957.

[75] Koller D. Representation reasoning learning. In: IJCAI2001 Computers and Thought Award Talk. 2001. [http://robotics.stanford.edu/~koller/CnT-web_files/v3_document.htm]

[76] Stone P. Learning and multiagent reasoning for autonomous agents. In: Proc. of the 20th Int'l Joint Conf. on Artificial Intelligence (IJCAI 2007). Hyderabad, 2007. 13−30.

附中文参考文献:

[2] 刘大有,于鹏,高滢,齐红,孙舒杨.统计关系学习研究进展.计算机研究与发展,2008,45(12):2110−2119.

[12] 张连文,郭海鹏.贝叶斯网引论.北京:科学出版社,2006.

[68] 孙舒扬,刘大有,孙成敏,黄冠利.统计关系学习模型 Markov 逻辑网综述.计算机应用研究,2007,24(2):1−3.

[69] 于鹏,刘大有,欧阳丹彤.基于遗传与粒子群算法的 Markov 逻辑网学习研究.电子学报,2006,34(12A):2551−2555.

[70] 孙舒杨,刘大有,孙成敏.基于后验概率的 Markov 逻辑网参数学习方法.吉林大学学报(理学版),2006,44(6):946−950.

[71] 张玉芳,黄涛,艾东梅,熊忠阳.Markov 逻辑网及其在文本分类中的应用.计算机应用,2009,29(10):2729−2732. [doi: 10.3724/SP.J.1087.2009.02729]

[72] 韩磊,李君峰,贾云得.基于时空单词的两人交互行为识别方法.计算机学报,2010,33(4):776−784. [doi: 10.3724/SP.J.1016.2010.00776]

[73] 楼俊杰,徐从富,郝春亮.基于马尔可夫逻辑网络的实体解析改进算法.计算机科学,2010,37(8):243−247.

徐从富(1969－),男,浙江开化人,博士,副教授,主要研究领域为人工智能,机器学习,数据挖掘,信息融合.

苏保君(1986－),男,硕士,主要研究领域为人工智能,机器学习,数据挖掘.

郝春亮(1986－),男,硕士,主要研究领域为人工智能,机器学习,数据挖掘,软件工程.

楼俊杰(1981－),男,硕士,主要研究领域为人工智能,机器学习.

© 中国科学院软件研究所 http://www.jos.org.cn