第37卷 第24期
Vol.37 No.24

计 算 机 工 程
Computer Engineering

2011年12月
December 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)24—0158—03 文献标识码：A 中图分类号：TP391

Markov逻辑网在迁移学习中的应用

熊忠阳，舒方俊，张玉芳，孔 润

(重庆大学计算机学院，重庆 400044)

摘 要：为充分利用过期训练数据和数据结构相关性进行新领域的学习，提出一种基于Markov逻辑网的迁移学习方法。该方法对源域与目标域的谓词进行自动映射后，通过自我诊断、结构更新和新公式挖掘3个步骤对映射结构进行优化，使之更适用于目标域数据。实验结果证明，与传统的机器学习方法相比，该方法使概率推理所获结果的准确率更高，所需的学习时间与训练数据更少。

关键词：迁移学习；Markov逻辑网；自动映射；机器学习；一阶逻辑

Application of Markov Logic Network in Transfer Learning

XIONG Zhong-yang, SHU Fang-jun, ZHANG Yu-fang, KONG Run

(College of Computer Science, Chongqing University, Chongqing 400044, China)

【Abstract】In order to take advantages of expired training data and correlation data structures to achieve the aim of learning new areas, this paper proposes a transfer learning approach based on Markov Logic Network(MLN). It autonomously maps the predicates on the source and target, and optimizes the mapping structure by self-diagnosis, structure update and new formula mining. Experimental results show that compared with traditional method, the probabilistic reasoning to the target domain MLN structure to gain higher accuracy of the results with less learning time and training data.

【Key words】transfer learning; Markov Logic Network(MLN); autonomous mapping; machine learning; first-order logic

DOI: 10.3969/j.issn.1000-3428.2011.24.053

1 概述

在传统的机器学习框架下，学习的任务就是在给定充分训练数据的基础上学习一个模型，然后利用这个学习到的模型对测试数据进行预测与评估。然而，当前的数据挖掘研究中存在一个关键的问题：在一些新出现的领域中，大量训练数据很难得到。传统的机器学习需要对每个领域标注大量训练数据，这将会耗费大量的人力与物力；而没有大量的标注数据会使很多与学习相关的研究与应用无法开展。其次，传统的机器学习假设训练数据与测试数据服从相同的数据分布。然而在许多情况下，这种同分布假设并不满足。不满足同分布假设的情况往往发生在训练数据过期，而标注新数据非常昂贵费时。由此得到了大量在不同分布下的过期训练数据，完全丢弃这些数据会造成很大的浪费。在这种情况下，迁移学习就变得非常重要了。

迁移学习是将从一个学习环境中学到的知识用于帮助新环境中学习任务的一种机器学习方法。因此，迁移学习不会像传统机器学习那样作同分布假设。它可以从现有的数据中迁移知识，用来帮助将来的学习[1]。本文使用基于Markov逻辑网(Markov Logic Network, MLN)的学习方法来实现迁移学习。

2 一阶逻辑和MLN

一阶逻辑主要由常量、变量、谓词和公式组成[2]。常量(可以有类型)用来描述域中的对象；变量作为占位符可以被实例化；谓词用来表示域中存在的关系，如WorkedFor；公式用来表示域中多个对象之间的关系。公式由常量、变量、谓词和逻辑连接词(如∨和∧)组成。没有变量的公式称为闭公式。一个可能世界就是为域中每个可能的闭谓词指定真值。公式的子句形式是将公式用析取式来表现，如：!P∨Q与P=>Q是逻辑等价的。

一阶逻辑知识库是一阶逻辑公式和句子所构成的集合，亦可看作是施加于可能世界上的一个约束的集合。在一阶逻辑中，一个世界如果违反了一个公式，则该世界不存在(即一个可能世界必须满足所有公式，否则该世界不存在)。Markov逻辑网的基本思想就是放松这个约束：一个世界违反了一个公式，则其在知识库中发生的概率降低，但并非不发生。一个世界违反的公式越少，其发生的可能性就越大。在Markov逻辑网中，每个公式都用权值来反映其约束的强弱：权值越大，满足该公式世界的发生概率与不满足该公式世界的发生概率之间的差就越大。Markov逻辑网的例子可以表示如下：

1.0 Director(A)∨Actor(A)
0.1 MovieMember(M,A)∨!Director(A)
1.2 !WorkedFor(A,B)∨!Director(A)
0.3 Director(A)∨MovieMember(M,A)∨!WorkedFor(B,A)

Markov逻辑是公式附加权值的Markov网[3]，表达式前的系数便是公式的权值。在Markov逻辑网中，每个节点对应一个闭谓词；一个特征值对应一个公式。Markov逻辑网推理的关键任务是在给定Markov逻辑网和常量的情况下，以其他公式为条件计算出待求公式的概率。

Markov逻辑网的权值可从一个或多个关系数据库中学习得到。本文采用判别式训练方法[4]。判别式方法是把一个

基金项目：中央高校研究生科技创新基金资助项目(CDJXS11180013)

作者简介：熊忠阳(1962－)，男，教授、博士生导师，主研方向：数据挖掘，网格计算，并行处理，互联网技术；舒方俊，硕士研究生；张玉芳，教授；孔 润，硕士研究生

收稿日期：2011-06-03 **E-mail：**shufj@qq.com

世界中的闭原子分为证据原子集合 X 和查询原子集合 Y，在给定 X 的情况下，Y 的条件概率为：

$$P(y \mid x)=\frac{1}{Z_X}\exp(\sum_{i\in F_y} w_i n_i(x,y))$$

其中，F_y 是所有 Markov 逻辑网子句的集合(至少有一个闭子句涉及到查询原子)；$n_i(x, y)$是涉及到查询原子的第 i 个子句的真闭子句个数。

需要说明的是，Markov 逻辑网只需简单地添加对应的公式就可以精确地表示人工智能中许多常用的模型。更为重要的是，Markov 逻辑网所构建的是一个非 IID 的模型，还可以作为很多统计关系学习任务的统一框架[5]。

3　基于 MLN 的迁移学习

基于 Markov 逻辑网的迁移学习主要分为 2 个步骤：首先将源域中的 MLN 结构与目标域进行映射，然后对映射得到的结构进行优化[6]。

3.1　结构映射

结构映射的目的是找到源域的 MLN 结构到目标域的最优映射。通常产生映射的方式有 2 种：全局映射与局部映射。其中，全局映射是指建立每个源谓词到目标谓词的映射，然后用它来迁移整个源域的 MLN 结构；局部映射是指将 MLN 结构中的每个公式独立出来，单独寻找每个公式中出现的所有谓词的最优映射。由于全局映射的空间复杂度是按源域中谓词个数呈指数增长的，对于谓词较多的 MLN 结构来说，想要寻找最优的全局映射比较困难；而通常情况下，一个公式中谓词的个数是远小于 MLN 结构中谓词总量的，这就使局部映射变得更为可行。所以，实验中采用局部映射的方法。

算法 寻找源域某一公式的一个合理映射

```
Procedure LEGAL_MAPPING(srcClause,tarPreds)
predsMapping ← ∅
typeConstraints ← ∅
Repeat
    Pick an unmapped source predicate srcPred
    For each unmapped target predicate tarPred do
        If isConmpatible(srcPred,tarPred) then
            Add this mapping to predsMapping
            Update typeConstraints
            Exit this for loop
        End if
    End for
Until All predicates in srcClause are mapped
Return predsMapping
End procedure
```

本文采用穷举法寻找一个公式最优的谓词映射，首先通过上述算法找到所有合理的谓词映射，然后从中找到最优的映射。如果源域中一个公式中的每一个谓词都能与目标域中的一个相兼容的谓词或者空谓词作映射，那么这个映射是合法的。2 个谓词兼容必须满足如下条件：它们有相同的参数个数，而且它们的参数类型与当前类型约束一致。例如：谓词 Publication(title, person)要与谓词 Gender(person, gend)兼容，则类型约束中就需要加入 title→person 和 person→gend 这 2 条约束。该公式中其他谓词映射的兼容性也必须满足这 2 条约束。

合理的谓词映射建立后，利用由目标域公式建立的 MLN 结构的加权伪对数似然性(Weighted Pseudo Log-likelihood, WPLL)[7]值对该映射进行评价。WPLL 值最大的谓词映射便是最优的。递归调用上述算法可以找到所有的合理映射，由此可以找到一个公式的最优谓词映射。

3.2　结构优化

结构优化是指将得到的映射结构进行优化，使之更适合目标域数据，主要有 3 个步骤：

(1)自我诊断

自我诊断主要用于查找结构中存在的不准确部分，即检查源域 MLN 结构中是否有公式需要加长或者缩短。每一个公式都可以写成推导形式，因此，有条件和结论 2 个部分，结论只有在条件满足的情况下成立。对于结论是错误的公式，就需要增加条件，使其条件尽量无法满足，从而达到结论很难成立的目的。而对于有些结论正确但由于条件冗长而无法成立的公式，就要相应地减少条件，使其条件容易得到满足，从而达到结论易于成立的目的。

(2)结构更新

根据自我诊断得到的结果，对需要处理的公式进行增加或者缩短。

(3)新公式挖掘

前面 2 个步骤都是对于源域迁移学习得到的公式进行更新优化，而无法去挖掘目标域中特有关系的公式。为了解决这个问题，使用关系寻径(Relational Pathfinding, RPF)[6]的方式，来寻找属于目标域自己的公式。一个简单数据库可以表示为：

Director(jack)　　MovieMember(movie1, jack)
Actor(jill)　　MovieMember(movie1, jill)
WorkedFor(jill, jack)　　MovieMember(movie2, jill)

关系寻径是将数据库看成是一个图 G，如图 1 所示，数据库中的常量是图中的节点，各个常量之间的关系作为边。

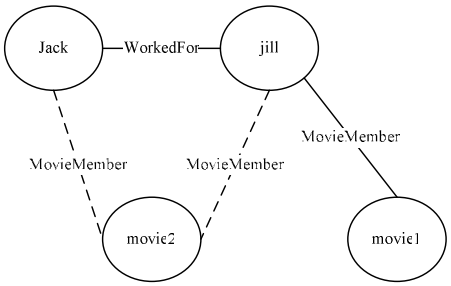

图 1　数据库的关系图

对于图中形成环路的部分，关系寻径方法会将环路中 2 个节点直接相连的边作为结论，而将其他边作为这个结论的条件，如图 1 中虚线部分与 WorkedFor 形成了环路，则通过抽象可以得到如下公式：

Actor(P)∧Director(P')∧MovieMember(M,P)∧MovieMember(M, P')=>WorkedFor(P,P')

4　实验与结果分析

4.1　数据集

本文采用 UW-CSE 数据集[8](源域数据集)与 IMDB 数据集[9](目标域数据集)进行迁移学习的实验。UW-CSE 数据集给出了华盛顿大学计算机科学与工程系的人员和这些人员之间关系等的描述；且将整个数据集按研究方向分为 5 个子集：ai，graphics，language，systems，theory。IMDB 数据集是由国际电影数据库给出的，被划分为 5 个子集，每个子集包含 4 部电影以及电影的导演和一线演员。2 个数据集的谓词结构如表 1 所示。其中，UW-CSE 数据集包含 1 323 个常量、9 种类型和 15 个谓词；IMDB 数据集包含 316 个常量、4 种类型和 10 个谓词。

表 1 数据集的谓词描述

UW-CSE	IMDB
TaughtBy(course,person,semester)	Director(person)
CourseLevel(course,level)	Actor(person)
Position(person,pos)	Movie(title,person)
AdvisedBy(person,person)	Gender(person,gend)
ProjectMember(project,person)	WorkedUnder(person,person)
Phase(person,phas)	Genre(person,genr)
TempAdvisedBy(person,person)	SamePerson(person,person)
YearsInProgram(person,year)	SameMovie(title,title)
TA(course,person,semester)	SameGenre(genr,genr)
Student(person)	SameGender(gend,gend)
Professor(person)	
SamePerson(person,person)	
SameCourse(course,course)	
SameProject(project,project)	
Publication(title,person)	

4.2 实验方法

为验证基于 Markov 逻辑网迁移学习的可行性，本文对 3 种方法进行了实验比较，包括：直接通过目标域数据集进行权值学习与概率推理(S-M)的传统机器学习方法，通过源域 MLN 结构映射得到目标域结构然后进行概率推理(M-M)，通过源域映射与结构优化得到的目标域结构然后进行概率推理(R-M)[6]。

实验预处理在 Microsoft Visual Studio 2003 下进行。Markov 逻辑网相关的迁移学习、权值学习和概率推理在 Alchemy[10]下进行。Alchemy 是一个基于 Markov 逻辑表示的软件包，提供了一系列统计关系学习和概率逻辑推理方面的算法。实验具体步骤如下：

(1)源域数据集的 Markov 逻辑表示：即上一节提到的源域的模型，根据 3.1 节中的谓词描述，构建源域的.mln 文件。

(2)知识库构建：根据谓词定义，生成包含各种不同谓词的知识库。

(3)迁移学习：使用源域中的 MLN 结构映射得到目标域的结构，通过迁移学习，得到目标域谓词映射的 MLN 结构与优化后的 MLN 结构。

(4)谓词推理：根据上一步学习得到的 MLN 结构和测试知识库，采用 MC-SAT 算法进行推理。

4.3 结果分析

本文使用查全率查准率曲线下面积(Area Under Curve, AUC)与条件对数似然性(Conditional Log-likelihood, CLL)[3] 2 个评价指标来评价实验结果的好坏。AUC 能很好地反映算法结果的准确性，CLL 能直接衡量通过优化近似方法估计出的概率分布的质量。图 2 给出 3 种方法得到的查全率-查准率曲线。

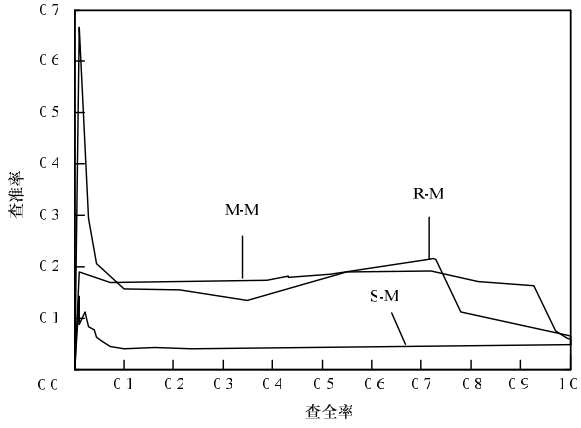

图 2 3 种方法的查准率-查全率比较

通过 Matlab 计算可得三者的 AUC 值，由于本文使用的查询谓词是 WorkedUnder，CLL 值也是对于 WorkedUnder 在数据中正确出现的概率取对数进行计算的，计算结果如表 2 所示，可以看出，R-M 与 M-M 的实验结果明显优于 S-M，而 R-M 是效果最好的。

表 2 3 种实验方法的 AUC 与 CLL 比较

实验方法	AUC	CLL
R-M	0.147	−1.917 3
M-M	0.136	−1.995 1
S-M	0.022	−3.816 7

综上所述，采用 Markov 逻辑网进行迁移学习的方式使通过目标域 MLN 结构进行概率推理得到的结果的准确率有很大的提高，由此证明基于 Markov 逻辑网的迁移学习方法可以获得比传统机器学习方法更好的效果。而且其需要的时间及数据集也比传统的机器学习方法少，这是因为通过相关域中已有的结构得到自己的知识结构比从海量的数据中分析得到要容易得多。

5 结束语

本文采用 Markov 逻辑网进行迁移学习，为新领域的学习提供了一个新思路。Markov 逻辑网还可以作为构建其他统计关系学习模型的框架，应用到文本分类、实体解析、信息抽取等诸多领域。下一步的工作是将 Markov 逻辑网应用到其他领域，如推荐系统。

参考文献

[1] 戴文渊. 基于实例和特征的迁移学习算法研究[D]. 上海: 上海交通大学, 2009.

[2] Russell S, Norvig P. Artificial Intelligence: A Modern Approach Upper Saddle River[M]. 2nd ed. [S. l.]: Prentice Hall, 2003.

[3] Richardson M, Domingos P. Markov Logic Networks[D]. Seattle, Washington, USA: University of Washington, 2004.

[4] Singla P, Domingos P. Discriminative Training of Markov Logic Networks[C]//Proceedings of the 20th National Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2005.

[5] Domingos P, Richardson M. Markov Logic: A Unifying Framework for Statistical Relational Learning[C]//Proceedings of International Conference on Machine Learning. Banff, Canada: IMLS Press, 2004.

[6] Mihalkova L, Huynh T, Mooney R J. Mapping and Revising Markov Logic Networks for Transfer Learning[C]//Proceedings of the 22nd Conference on Artificial Intelligence. Vancouver, Canada: [s. n.], 2007.

[7] Mihalkova L S. Improving Learning of Markov Logic Networks Using Transfer and Bottom-up Induction[D]. Texas, USA: Department of Computer Sciences University of Texas at Austin, University of Texas at Austin, 2009.

[8] UW-CSE Dataset[DB/OL]. [2010-09-22]. http://www.cs.washington.edu/ai/mln/database.html.

[9] IMDB Dataset[DB/OL]. [2010-09-19]. http://www.imdb.com.

[10] Kok S, Singla P, Richardson M, et al. The Alchemy System for Statistical Relational AI[Z]. Department of Computer Science and Engineering, University of Washington, 2005.

编辑 张 帆