文章编号：2095-2163(2020)01-0265-06 中图分类号：TP181 文献标志码：A

基于规则过滤和谓词覆盖的 MLN 迁移算法研究

吁 松，何 慧，王 星
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要： 迁移学习的目标是将源领域的知识迁移到目标领域，从而实现在数据稀缺的目标领域上获得良好的效果。在处理关系型数据时，研究将迁移学习与马尔科夫逻辑网络相结合，得到一种基于一阶逻辑公式映射的迁移学习算法。本文的迁移算法针对的是目标域数据极少的情况。为了提升迁移的效果，研究基于迁移规则和依据权重覆盖谓词的策略对映射的公式进行适当的筛选，迁移对目标域价值最大的公式，最终提升整体迁移效果。研究用 3 个从现实世界收集的关系型数据来验证本文算法，并与现有的算法进行对比，结果显示本文的算法具有出色的表现。

关键词： 迁移学习；马尔可夫逻辑网络；谓词映射；规则筛选；谓词覆盖

MLN transfer learning by rule filter and predicate cover

YU Song, HE Hui, WANG Xing
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 The goal of transfer learning is to transfer knowledge from the source domain to the target domain so as to achieve good results in the target domain where the label data is rare. When dealing with relational data, the paper combines transfer learning with Markov logic networks to obtain a transfer learning algorithm based on first-order logic formula mapping. The proposed transfer learning algorithm is aimed at the situation where the target domain data is minimal. In order to improve the effectiveness of the transfer, the paper appropriately filters the formulas of the mapping by the rule filter and predicate cover strategy, and transfers the formulas with the greatest value to the target domain, finally improves the overall transfer effect. The proposed algorithm is validated with three relational data collected from the real world and compared with existing algorithms. The results show that the proposed algorithm has excellent performance.

【Key words】 transfer learning; Markov Logic Network; predicate mapping; rule filter; predicate cover

0 引 言

随着机器学习领域研究的不断深入，迁移学习作为机器学习领域的一个重要方向而引起研究者的高度关注与重视。相对于传统的机器学习针对一个领域进行训练、测试和运用，迁移学习针对的是 2 个不同但有关联的领域，在一个领域上训练，在另一个领域上测试和运用，前者被称为源领域，后者被称为目标域。这种训练数据与测试数据处于不同领域的要求，正是因为人们在运用传统的机器学习解决问题时，一些获取有标签数据代价昂贵或者难以收集的领域受到有标签数据不足的困扰[1]。而这类领域正是迁移学习可以发挥作用的地方。

迁移学习从其迁移的内容来看，可以分为迁移参数、迁移实例、迁移特征和迁移关系[1]。迁移关系的算法核心是寻找源领域和目标领域之间共享的关系，Li 等人[2]就利用 bootstrapping 的方法迭代构建领域之间的关系。针对马尔科夫逻辑网络[3]的迁移算法，大部分就是一种基于关系的迁移。研究者往往是利用马尔科夫逻辑网络模型的逻辑公式作为源领域和目标领域关系的桥梁。其中，一类算法是将 MLN 模型中的一阶逻辑公式转换成高阶形式，然后进行模型的迁移。Davis 等人[4]通过引入谓词变量将一阶逻辑公式转换成二阶公式，并将这些二阶公式合并成团，再对每个二阶团进行评估，将分数最高的 k 个二阶团迁移到目标域。Haaren 等人[5]同样是将一阶公式转换成二阶公式，但在二阶转换成一阶的过程引入了偏置，实现了效果更好的迁移。另一类算法则是通过谓词映射的方式直接生成目标域的一阶逻辑公式，然后采用不同的策略来调整、变换公式，最终筛选公式进入 MLN 模型[6]，或者更简单一点，运用筛选策略不加调整地迁移公式到 MLN 模型中[7]。

本文设计的迁移算法是一类对目标域数据量要求不高的算法，在谓词映射算法[6]的基础上，提出了基于规则的公式迁移策略和依据权重进行谓词覆

基金项目： 国家自然科学基金(61472108)；国家重点研发计划(2017YB0801801, 2017YFB0803300)。

作者简介： 吁 松(1993-)，男，硕士研究生，主要研究方向：迁移学习；何 慧(1974-)，女，博士，教授，博士生导师，主要研究方向：迁移学习、移动网络安全；王 星(1981-)，男，博士，主要研究方向：网络与信息安全、网络舆情监控、知识迁移。

收稿日期： 2018-06-18

盖的迁移策略,实现了马尔科夫逻辑网络模型[8-10]的迁移。

1 MLN 迁移算法

本文提出的马尔可夫逻辑网络迁移算法是一种迁移马尔可夫逻辑网络模型中一阶逻辑公式的方法。目标领域只需要提供一个单实体为中心的实例就可以实现迁移。迁移算法总体设计思想是将源领域的一阶逻辑公式通过谓词映射转换成目标域的一阶逻辑公式,然后利用数据验证、规则迁移和基于权重的谓词覆盖等手段迁移公式,最后生成目标域的 MLN 模型。本文的最大贡献就是提出了基于规则的公式迁移策略以及基于权重的谓词覆盖策略。

1.1 数据验证

首先,MLN 迁移学习算法需要通过谓词映射生成目标域公式。本文采用局部谓词映射方法,即类型的一致性约束只针对单个公式,不同的公式的类型约束可以是不同的。在映射过程中,研究的算法还额外要求谓词一致性,即在针对单个公式进行迁移时,源领域的谓词与目标域的谓词是一一对应的。谓词一致性约束同样是局部的,即不同公式中的源领域谓词可以对应目标域中的不同谓词。增加这样 2 种局部的一致性约束一方面可以节省算法的运行时间;另一方面相对于全局性的一致性约束而言,有利于生成多样化的迁移公式,进而提高迁移效果。

在得到目标领域的公式之后,研究利用仅有的目标域数据对这些映射公式进行验证。在本文中,可以用数据来验证的公式即称为被数据验证的公式,简称为已验证的公式。同时,进一步对被数据验证过的公式进行细分,将上一步得到的已验证的公式分为 2 种。一种是经过数据验证成立的公式,即可行公式,另一种是数据验证后不成立的公式,即不可行公式。研究借鉴 Lilyana 等人[7]的思想,提取所有可行公式中包含的映射得到可行谓词映射集合,提取所有不可行公式包含的谓词映射组成不可行谓词映射集合,利用其来筛选公式。而且,考虑到不可行公式中的谓词映射并不都是不可行的映射,有的只是由于公式中的某个谓词的映射不好而导致公式验证失败。因此,研究拟使用不可行谓词映射集合与可行谓词映射集合做差,得到的真正的不可行谓词映射集合。此后用这种真正的不可行映射来筛选未被数据验证的公式,也称之为未验证公式,得到候选公式。

在前面提到过,本文算法对目标域数据量的设定是少量数据,所以不会得到太多已验证的公式,而未验证的公式数量则较多。这些未被数据验证过的目标公式中既含有对目标域推理有价值的公式,也含有对目标域推理无价值的公式,因此就需要通过其它手段—基于规则的筛选和根据权重进行谓词覆盖—获得更多对目标域推理有用的公式。

1.2 基于规则的迁移策略

通过对已有的非迁移马尔可夫逻辑网络模型中逻辑公式的观察,研究发现大部分公式都具有如下特征:公式的前置条件参数之间的相互关联,最后推导出与之有关的结论参数的关系。对比迁移得到的未验证的公式,有很多公式违背这个特征,存在一些相互关联的变量推导出与之无关的变量之间的关系,因此推导出一些不合理的结果。

表 1 是非迁移的方式得到 uwcse 领域的 MLN 模型中的 2 条公式。从第一条中可以看到,前置条件是 a_1 在 a_4 学期教 a_3,a_2 在 a_4 学期教 a_3,由此就可以得到一个可能成立的结论 —a_1 和 a_2 是同一人。同理,第二条的前置条件中给出了 a_1 出版 a_2,a_3 出版 a_2,a_1 不是学生,a_3 是学生,于是可以推出 a_1 是教授的结论。 当然上述结论并不都是一定正确,只是有较大概率成立而已,但由于 MLN 的公式有权重来进行描述,因此可以允许这种推导出非确定性结论的公式。表 2 是从 imdb 域向 uwcse 域迁移得到的 MLN 模型的公式中选取了 2 条。第一条,前置条件描述了 a_2 和 a_3 的关系,但结论给出的却是 a_1 和 a_2 的关系,相对而言并不合理。 同样,表 2 中的第 2 条,前置条件分别描述的 a_1 与 a_2 的关系和 a_1 与 a_3 的关系,结论却给出了 a_4 和 a_2 的关系,也不是合理的推导。 基于上述的观察和特征的归纳,本次研究提出一个迁移候选公式的规则,筛选出的公式则称为符合规则的公式,简称为规则公式。

表 1 UWCSE MLN 模型公式

Tab. 1 MLN formulas in UWCSE

序号	模型公式
1	$\text{taughtBy}(a_3,a_1,a_4) \wedge \text{taughtBy}(a_3,a_2,a_4) \Rightarrow \text{samePerson}(a_1,a_2)$
2	$\text{publication}(a_2,a_1) \wedge \text{publication}(a_2,a_3) \wedge \neg\,\text{student}(a_1) \wedge \text{student}(a_3) \Rightarrow \text{professor}(a_1)$

表 2 IMDB 迁移到 UWCSE 域得到的迁移公式

Tab. 2 Transferred formulas from IMDB to UWCSE

序号	迁移公式
1	$\text{position}(a_2,a_3) \Rightarrow \text{tempAdvisedBy}(a_1,a_2)$
2	$\neg\,\text{publication}(a_1,a_2) \wedge \text{publication}(a_1,a_3) \Rightarrow \text{advisedBy}(a_4,a_2)$

为了进一步解释基于规则的公式迁移原理,研究使用图 1 和图 2 分别解释了一个符合规则的公式

验证过程和一个不符合规则的公式验证过程。图 1~2 中的每个圆代表一个集合，2 个圆相交的部分是 2 个集合共同的元素。红色虚线圆圈表示的是作为结论的谓词的实参元素集合，箭头右边的实线圆圈代表可能被推导出关系的元素集合。图 1、图 2 中的 2 个公式的前提条件都是 2 个原子公式构成，这 2 个原子公式的实参都含有 a_1，因此，a_1 成为 2 个集合之间关系的桥梁，2 个本没有关系的集合如今可能存在某种关系，于是并成一个集合，即箭头右边的集合。该集合中任何元素之间都有可能存在某种可推理的关系，或者可以认定某人的身份，因此如果红色圆圈代表的集合是该集合的子集，即图 1 所示的情况，则相应的关系可能存在，该公式被认为是符合规则的公式，反之，如果公式中作为结论的谓词的实参不完全在前置条件实参构成的集合中，即图 2 所示的状态，则这个结论由条件关系推导出的可能性就低，该公式被认为是不符合规则的公式。

$$\text{taughtBy}(a_2,a_3,a_4) \wedge \text{taughtBy}(a_1,a_3,a_4) \Rightarrow \text{same Course}(a_1,a_2)$$

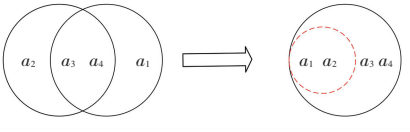

图 1　符合规则的公式示意图

Fig. 1　A formula that conforms to the rule

$$\neg\,\text{publication}(a_1,a_2) \wedge \text{publication}(a_1,a_3) \Rightarrow \text{adviseBy}(a_4,a_2)$$

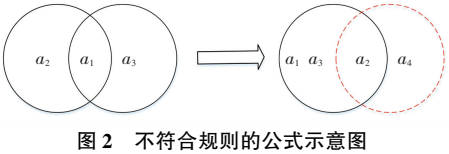

图 2　不符合规则的公式示意图

Fig. 2　A formula that does´t conform to the rule

需要强调的是，符合规则的公式推导出的结论也有可能是错误的，反过来，不符合规则的公式推导出的结论也是有存在可能的。但这不会造成太多不良影响，因为马尔可夫逻辑网络是一种软化逻辑公式硬约束的模型，故而不需要模型中的公式是绝对正确的。

1.3　基于谓词覆盖的迁移策略

这里，首先要明确谓词被覆盖和谓词没被覆盖的定义。所谓谓词覆盖是指目标域中的谓词存在于 MLN 模型的某个公式中，而所谓谓词没被覆盖是指目标域中存在谓词不在研究的 MLN 模型的公式中。谓词覆盖就是指去覆盖那些没有被覆盖的谓词，而依据权重意味着需要优先使用权重高的公式去覆盖谓词。研究中要尽量去覆盖所有目标域的谓词，是因为通过初步试验发现，如果某个谓词没有公式去覆盖的话，那么针对该谓词的推理就不能取得良好的效果。因为该谓词没有公式覆盖，那么推理程序就没有推理的依据，故而得不到正确的结果。因此，本文研发的算法考虑在经过前 2 步公式迁移后，在未能覆盖所有谓词的情况下，挑选剩余的公式中可以覆盖这些谓词的公式作为 MLN 模型公式的补充。研究将使用 *alpha* 参数来指示每个谓词需要被多少个公式覆盖，对于不同的源领域和目标领域可以设置不同的值以达到最好的效果。设计流程步骤是：先根据权重大小排序，然后统计目前 MLN 模型公式还未能覆盖的谓词或者说覆盖的公式数量还没能达到 *alpha* 参数要求的谓词，最后将公式依据权重大小补充进迁移的 MLN 模型中，使其尽量满足 *alpha* 参数规定的数目。

2　实验评估

在这一部分，研究将对算法进行实验评估。这里对比了 2 个 MLN 迁移学习算法，分别是 TAMAR 算法[6] 和 SR2LR 算法[7]。这 2 个算法都有谓词映射的步骤，除谓词映射之外的迁移手段是不同算法之间的主要区别，因此非常适合用于与本文提出算法的对比。

为了分析算法的表现，采用了 2 种典型的用于分析马尔可夫逻辑网络的度量方法——*AUC - PR* 和 *CLL*。研究可知，PR 曲线是精确度-召回率曲线，*AUC - PR* 是指 PR 曲线下的面积。如果用一般的正确率这种方式来衡量则容易被大量不存在的关系的正确率所影响，导致评估结果与真实使用情况有差距。条件对数似然（conditional log-likehood，*CLL*）则主要用于评估马尔科夫逻辑网络推理的质量，是对 *AUC - PR* 的一种补充。*CLL* 值越大，则推理质量越高；反之，值越低，推理质量越差。仍需看到，文中的评估方法比较简单，由此反映得出的推理质量并非精确可靠，如果模型能够生成足够有区别的阈值的话，*CLL* 高低并不重要。

实验中使用了 3 个公开的关系型数据集，分别是 IMDB、UWCSE 和 WebKB。这 3 个数据集都是从现实世界中收集而来的，在时下研究的实验中得到了广泛的使用。其中，UWCSE 数据收集自华盛顿大学的计算机科学与工程系，记录了课程、教授、学生等身份信息，并记录了个体之间的关系，例如，advisedBy、taughtBy 等等。IMDB 数据集是 Lily Mihalkova 采集自 IMDB 数据库的电影领域的相关信息，具体包含了导演、演员、电影等信息以及不同个体之间的关系。WebKB 数据集则记录了 4 所大

学计算机系的 Web 网页和超链接信息。

实验中,测试了 3 个数据集构成的共 6 个迁移场景:IMDB→UWCSE、IMDB→WebKB、UWCSE→IMDB、UWCSE→WebKB、WebKB→IMDB、WebKB→UWCSE。其中,箭头前方是源领域,箭头后方是目标域。在本文的后面章节,会着重展示这 4 个算法在 6 种迁移场景中的表现,而后在这 6 种迁移场景中测试本文提出的迁移策略的效果,最后将基于实验讨论分析规则迁移、谓词覆盖和 *alpha* 参数在本文提出算法中的作用。

3 实验验证

3.1 实验结果对比

实验中,首先测试了 4 种算法在 6 种迁移场景下的表现,4 个算法的 *AUC - PR* 值和 *CLL* 值分别见表 3、表 4。表格的第一列是目标域,第二列是源领域。RFPC 是本文提出的迁移学习算法,其 *alpha* 的参数取值为 4。

从表 3 和表 4 可以看出,在本文的实验中 TAMAR 算法和 SR2LR 算法表现较为相近,且 TAMAR 算法还略好于 SR2LR 算法。2 个算法在迁移的目标域为 WebKB 时,表现几乎一致,*AUC - PR* 都是 0.49,但是在 *CLL* 这个指标上,SR2LR 又略好于 TAMAR。测试的迁移场景是 UWCSE 向 IMDB 迁移时,SR2LR 表现较好,*AUC - PR* 值比 TAMAR 高 0.05 左右。而当迁移场景是 WebKB 向 UWCSE 迁移时,TAMAR 表现较好,*AUC - PR* 比 SR2LR 高 0.1 左右。在其它迁移场景下,2 个算法的 *AUC - PR* 的相差不大。

接下来,将 2 个已有的迁移学习算法与本文提出的 RFPC 进行比较。从表 3 可以看出,在 IMDB 向 WebKb 迁移时,本文的算法比 SR2LR 效果要略有逊色,但在剩余的全部数据上,本文的算法在 *AUC - PR* 这个指标上是超过 SR2LR 算法的,因此在整体上来看,本文的迁移算法得到的马尔可夫逻辑网络模型能够做出更好的推理。从表 4 可以看到,本文的迁移算法在 *CLL* 指标上普遍比 SR2LR 算法差,除了 UWCSE 向 IMDB 迁移时,本文的算法的 *CLL* 指标比 SR2LR 高之外,这意味着本文算法的推理结果概率普遍低于 SR2LR 算法。这可能是因为本文的算法未能进行权值的调整。但是如前述分析可知,如果推理概率的阈值选择恰当,就不会影响本文算法的预测效果,因此这也不会意味着本文的算法比 SR2LR 算法更差。在与 TAMAR 算法比较时,本文的算法在迁移目标域为 UWCSE 时,*AUC - PR* 的值比其略有不及,但在其它迁移场景中,本文的算法均是优于 TAMAR 的,并且在 UWCSE 向 WebKB 和 WebKB 向 IMDB 迁移时,本文的算法在 *AUC - PR* 指标上将远远高于 TAMAR。因此,从整体上来看,本文的算法比 SR2LR 和 TAMAR 都是要好的。

表 3 不同算法在不同数据集上的平均 *AUC-PR* 的值

Tab. 3 The average *AUC-PR* of different algorithms

目标领域	源领域	TAMAR	SR2LR	RFPC
IMDB	UWCSE	0.310 7	0.361 5	0.364 9
IMDB	WebKB	0.305 5	0.292 6	0.454 4
UWCSE	IMDB	0.332 8	0.314 2	0.322 4
UWCSE	WebKB	0.325 6	0.224 3	0.283 1
WebKB	IMDB	0.490 0	0.490 0	0.499 0
WebKB	UWCSE	0.490 0	0.490 0	0.985 3

表 4 不同算法在不同数据集上的平均 *CLL* 的值

Tab. 4 The average *CLL* of different algorithms

目标领域	源领域	TAMAR	SR2LR	RFPC
IMDB	UWCSE	−1.092	−1.506	−1.290
IMDB	WebKB	−1.127	−0.292	−0.651
UWCSE	IMDB	−1.294	−1.901	−1.051
UWCSE	WebKB	−1.304	−0.283	−1.621
WebKB	IMDB	−0.301	−0.300	−0.199
WebKB	UWCSE	−0.301	−0.283	−2.118

3.2 实验结果分析

基于前述仿真测试研究过程,这里拟将探讨剖析 RFPC 算法中不同部分发挥的作用。同样,研究在 6 种迁移场景下对比这些算法的效果,运行后详情见表 5、表 6。其中,Only-Data 代表只使用被目标域数据验证过的公式生成 MLN 模型,Data+Rule 代表了用数据验证过的公式和规则迁移公式构成的 MLN 模型。最后一列代表了完整的 RFPC 算法(*alpha* 参数的取值为 4),该算法迁移得到的 MLN 模型中包含了数据验证的公式,规则迁移的公式和依据权重进行谓词补充的公式。

从表 5 中可以看到,除了向 UWCSE 领域的迁移之外,Data+Rule 的方案都比只有 Data 的方案要好,而且大部分情况下均是如此。例如,Data+Rule 在 WebKB 领域向 IMDB 领域迁移时,*AUC - PR* 值比 Only-Data 高出了 0.157 4。更为明显的是当源领域为 UWCSE、目标域为 WebKB 的情况,Data+Rule 的 *AUC - PR* 值接近 Only-Data 的 *AUC - PR* 值的 2 倍。即便是在向 UWCSE 迁移时,Data+Rule 只比 Only-Data 在 *AUC - PR* 值上低了一点点。从 IMDB 向 WebKB 迁移的时候,Only-Data 和 Data+Rule 则

显出劣势，*AUC－PR* 为 0 是因为数据验证和规则迁移两种策略都不能得到合适的公式，此时 MLN 不会在这种情况下进行推理，故而对应的 *AUC － PR* 值为 0。此种情况下，根据权重进行谓词覆盖的作用就得以体现，在补充了一些权重较高的公式之后，RFPC 算法依旧能够取得较好的效果。在前 2 种公式迁移策略能够有效发挥作用时，根据权重进行谓词覆盖的效果将不再直观明显，但如果 *alpha* 设置合理，那么对推理效果也能有一些提升，关于该点将在下文予以阐释分析。

表 5　不同算法在不同场景下的平均 *AUC-PR* 的值

Tab. 5　The average *AUC-PR* of different algorithms

目标领域	源领域	Only-Data	Data+Rule	RFPC
IMDB	UWCSE	0.326 8	0.367 4	0.364 9
IMDB	WebKB	0.293 0	0.450 4	0.454 4
UWCSE	IMDB	0.323 8	0.323 3	0.322 4
UWCSE	WebKB	0.279 2	0.267 7	0.283 1
WebKB	IMDB	0	0	0.499 0
WebKB	UWCSE	0.490 0	0.985 1	0.985 3

表 6　不同算法在不同场景下的平均 *CLL* 的值

Tab. 6　The average *CLL* of different algorithms

目标领域	源领域	Only-Data	Data+Rule	RFPC
IMDB	UWCSE	−0.927	−1.182	−1.290
IMDB	WebKB	−0.279	−0.588	−0.651
UWCSE	IMDB	−1.182	−1.182	−1.051
UWCSE	WebKB	−0.273	−1.557	−1.621
WebKB	IMDB	0	0	−0.199
WebKB	UWCSE	−0.301	−2.119	−2.118

由于研究的算法中有一个 *alpha* 参数，指示了研究在覆盖谓词时需要选择多少公式，所以研究有必要对其进行深入分析，探寻考察不同的 *alpha* 参数会对算法效果产生的影响。同样测试了在 6 种迁移场景下，*alpha* 参数对 RB 算法的影响，*alpha* 参数设置的范围为[0,18]。

从图 3 中可以看到，除了之前分析过的情况，*alpha* 参数对算法效果的影响较为有限。 在 *alpha* 参数增加时，在一定程度上将会提升算法的效果，但有时也会导致算法效果下降。研究中可以在 IMDB 向 UWCSE 和 WebKB 向 UWCSE 迁移时发现较大的波动，说明在 Data+Rule 的 *AUC － PR* 值不高时，*alpha* 参数影响相对较大。当 Data+Rule 的 *AUC － PR* 值较大时，*alpha* 参数的影响相对较小，正如研究中看到当 UWCSE 向 WebKB 迁移时未见任何波动。由此也可以推断得出，作为本算法的关键部分—基于规则的公式迁移策略—起着关键的作用，用于覆盖谓词的公式起到一个补充的作用，而 *alpha* 参数却只是用于最后的微调。

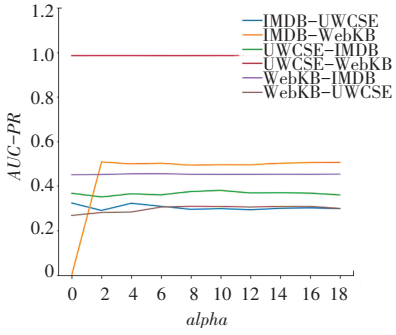

图 3　*alpha* 参数对算法的影响

Fig. 3　The effect of *alpha* parameters on the algorithm

研究至此，又统计了 RFPC 算法（*alpha* 参数设置为 4）在所有迁移场景下生成的 MLN 模型中不同来源的公式数量。并且推导计算了每类公式为推理效果做出的贡献，用于详尽评估每种迁移策略的效果。计算公式贡献的方法为每类公式带来的 *AUC － PR* 的提升除以该类公式的数量。

图 4 展示的是 *alpha* 参数设置为 4 时，RFPC 迁移算法生成的模型中不同来源的公式的分布情况。从图 4 中可以看出，不同的迁移场景，不同来源的公式分布情况是不同的。除了 2 种极端情况，也就是 IMDB 向 UWCSE 迁移时数据验证的公式占主导和 IMDB 向 WebKB 迁移时只有谓词覆盖的公式以外，研究发现基于规则的迁移公式构成了 MLN 迁移模型公式的主要部分。

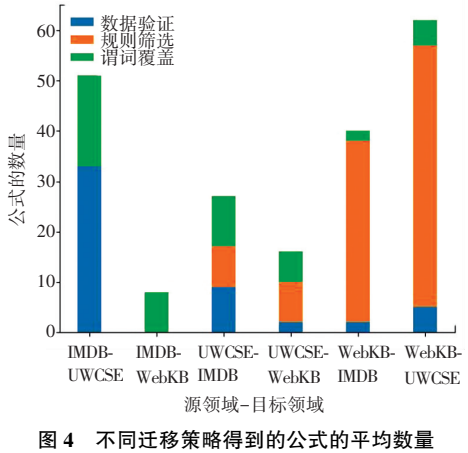

图 4　不同迁移策略得到的公式的平均数量

Fig. 4　The average number of formulas from different sources

图 5 展示的是不同类别的公式为最终的推理效果做出的贡献。由图 5 可以看到，除了 IMDB 向 WebKB 迁移的时候（因为该迁移场景下 MLN 模型

中只有谓词覆盖的公式),数据验证迁移的公式做出了最大的贡献。正如研究所希望的,基于规则的公式贡献总体而言仅居次席,并且相对来说是明显大于基于谓词覆盖的公式,这说明本文提出的规则发挥了应有效果,规则迁移的公式要远远胜过根据权重进行谓词覆盖得到的迁移公式。某些情况下,基于规则的公式贡献比基于数据验证的公式贡献明显要小,这是因为真实的数据是检验映射公式的最佳规则。需要特别注意的是,基于数据验证的公式中也会存在部分公式符合规则,这也说明满足规则的公式所做出的实际贡献会高于图 5 中所显示的贡献。

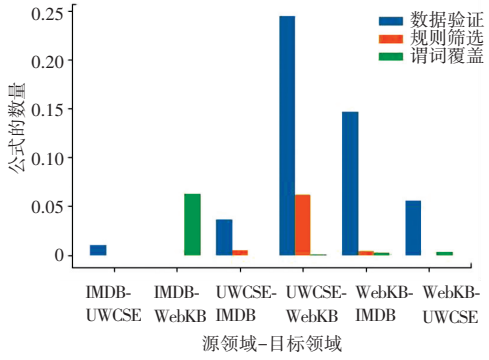

图 5 不同迁移策略得到的公式的平均贡献

Fig. 5 Average contribution of formulas from different sources

4 结束语

本文研究提出了一种针对极为有限的目标域数据的 MLN 迁移算法,通过提出一种符合逻辑的公式迁移规则,在没有目标域数据支撑的情况下迁移映射出来的目标域公式,并根据迁移公式的权重,尽量覆盖所有目标域谓词。通过在 6 种迁移场景下的实验验证,可以看到本文的算法超过现有的 MLN 迁移算法,同时也验证了本文提出的规则的有效性,以及基于权重覆盖目标域谓词的价值。

在后续的工作中,研究将尝试把 MLN 迁移学习算法运用到更多领域中去,考虑在多个源领域向一个目标域迁移的场景中运用本文的算法以及提出更多迁移公式的规则或算法使得在没有更多目标域数据支撑的情况下迁移更好的公式到目标域。

参考文献

[1] WEISS K, KHOSHGOFTAAR T M, WANG Dingding. A survey of transfer learning[J]. Journal of Big Data, 2016, 3(1):9.

[2] LI Shoushan, ZONG Chengqing. Multi-domain adaptation for sentiment classification: Using multiple classifier combining methods[C]// International Conference on Natural Language Processing and Knowledge Engineering. Beijing, China: IEEE, 2008:1-8.

[3] RICHARDSON M, DOMINGOS P. Markov logic networks (vol 62, pg 107, 2006)[J]. Machine Learning, 2006, 63(2):207.

[4] DAVIS J, DOMINGOS P. Deep transfer via second-order Markov logic[C]//ICML '09 Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Quebec, Canada: ACM, 2009:217-224.

[5] HAAREN J V, KOLOBOV A, DAVIS J. TODTLER: Two-order-deep transfer learning[C]// Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin, Texas: AAAI Press, 2015:3007-3015.

[6] MIHALKOVA L, HUYNH T N, MOONEY R J. Mapping and revising Markov logic networks for transfer learning[C]// Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence. British Columbia, Canada: AAAI Press, 2007:608-614.

[7] LILYANA M, MOONEY R J. Transfer learning from minimal target data by mapping across relational domains[C]// Proceedings of the 21st International Joint Conference on Artificial Intelligence. Pasadena, California, USA: Morgan Kaufmann Publishers Inc, 2009:1163-1168.

[8] KOK S, DOMINGOS P. Learning the structure of Markov logic networks[C]// Proceedings of the Twenty-Second International Conference (ICML 2005) on Machine Learning. Bonn, Germany: dblp, 2005:441-448.

[9] SANGHAI S, DOMINGOS P, WELD D. Learning models of relational stochastic processes[M]// GAMA J, CAMACHO R, BRAZDIL P B, et al. Machine learning: ECML 2005. Lecture Notes in Computer Science. Berlin/ Heidelberg: Springer, 2005, 3720:715-723.

[10] PEARL J. Probabilistic reasoning in intelligent systems: Networks of plausible inference[M]. San Francisco: Morgan Kaufmann, 1988.

(上接第 264 页)

3 结束语

本文通过运用利益博弈理论分析方法对农村居民医疗保险制度在实施过程中参保农民、定点医疗服务机构、医疗保险经办机构三个参与主体之间的利益博弈以及做出的各种选择分析医疗保险骗保问题产生的原因,利用当前大数据的广泛应用,从而提出相应的对策措施,能够对医疗保险骗保问题的解决提供一定的路径借鉴。

参考文献

[1] 周坚,周志凯,何敏. 基本医疗保险减轻了农村老年人口贫困吗-从新农合到城乡居民医保[J]. 社会保障研究,2019(3):33-45.

[2] 阳义南,肖建华. 医疗保险基金欺诈骗保及反欺诈研究[J]. 北京航空航天大学学报(社会科学版),2019,32(2):41-51.

[3] 单苗苗. 构建多元共治的医疗保险监管体系[J]. 中国人力资源社会保障,2018(6):39-40.

[4] 谭思然,蒲川. 实现医保对医疗行为监管模式转变的路径思考[J]. 中国卫生事业管理,2018,35(7):507-508,524.

[5] 支济祥. 大数据背景下我国医保基金风险防控研究[J]. 中国卫生产业,2017,14(8):12-13.