ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.1, January 2008, pp.73−81
DOI: 10.3724/SP.J.1001.2008.00073
© 2008 by *Journal of Software*. All rights reserved.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种从不完备关系数据中学习 PRM 的方法*

李小琳, 周志华+

(南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093)

An Approach to Learning PRM from Incomplete Relational Data

LI Xiao-Lin, ZHOU Zhi-Hua+

(National Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)

+ Corresponding author: Phn: +86-25-83686268, Fax: +86-25-83686268, E-mail: zhouzh@nju.edu.cn, http://cs.nju.edu.cn/people/zhouzh/

Li XL, Zhou ZH. An approach to learning PRM from incomplete relational data. *Journal of Software*, 2008,19(1):73−81. http://www.jos.org.cn/1000-9825/19/73.htm

Abstract: Existing relational learning approaches usually work on complete relational data. However, in real-world applications, data are often incomplete. This paper proposes the MLTEC (maximum likelihood tree and evolutionary computing method) method to learn structures of the probabilistic relational models (PRMs) from incomplete relational data. The incomplete relational data are filled randomly at first, and a maximum likelihood tree (MLT) is generated from each completed data sample. This population of MLTs is then evolved through an evolutionary computing process, and the incomplete data are modified by using the best evolved structure in each generation. As a result, the probabilistic structure is learned. Experimental results show that the MLTEC method can learn good structures from incomplete relational data.

Key words: machine learning; relational learning; incomplete data; probabilistic relational model; maximum likelihood tree; evolutionary computing

摘 要: 现有的关系学习研究都是基于完备数据进行的,而现实问题中,数据通常是不完备的.提出一种从不完备关系数据中学习概率关系模型(probabilistic relational models,简称 PRMs)的方法——MLTEC(maximum likelihood tree and evolutionary computing method).首先,随机填充不完备关系数据得到完备关系数据.然后从每个随机填充后的数据样本中分别生成最大似然树并作为初始 PRM 网络,再利用进化过程中最好的网络结构反复修正不完备数据集,最后得到概率关系模型.实验结果显示,MLTEC 方法能够从不完备关系数据中学习到较好的概率关系模型.

关键词: 机器学习;关系学习;不完备数据;概率关系模型;最大似然树;进化计算

中图法分类号: TP181 **文献标识码**: A

在传统的机器学习范式中,数据通常以“属性-值”方式存在,即表示为单表形式.但在现实世界中,许多数据都存在着内部关系,即表示为多表形式的关系数据.例如,由顾客的性别、年龄、收入、消费来判断该顾客是否是 *bigspender*,从而发现潜在的顾客群.但是,一个人是否是 *bigspender* 还取决于其他人,如配偶、朋友.因此,该问

* Supported by the National Natural Science Foundation of China under Grant Nos.60635030, 60473046 (国家自然科学基金); the China Postdoctoral Science Foundation under Grant No.20060390921 (中国博士后科学基金); the Jiangsu Planned Projects for Postdoctoral Research Funds of China under Grant No.0601017B (江苏省博士后科研资助计划)

Received 2006-10-08; Accepted 2006-12-19

© 中国科学院软件研究所 http://www.jos.org.cn

题不满足传统机器学习中普遍要求的独立同分布假设.在此类数据的样本之间或者样本的属性之间,往往存在内在的关系或结构.由于关系数据的表示形式与“属性-值”形式截然不同,传统的基于“属性-值”表示的机器学习技术难以用于解决这类问题.因此,关系学习应运而生并受到了极大的重视[1].

关系学习最初比较成功的算法大都来自 ILP(inductive logic programming)[2]领域.近年来,最受重视的领域之一是统计关系学习(statistical relational learning)技术.由于概率模型自身的特点,将概率模型,尤其是 Bayesian 网及其扩展[3,4]与一阶逻辑结合起来的统计关系学习方法能够有效地处理 ILP 所不能处理的噪声和不确定性问题[5],并且学习到的模型能够很直观地看出变量间的依赖性.近年来,已经有许多统计关系学习方面的模型被提了出来,例如,Probabilistic Relational Models[6],Stochastic Logic Models[7],Bayesian Logic Programs[8],Relational Markov Models[9],Fist-Order Bayesian Classifiers[10,11],Markov Logic Networks[12],Directed Acyclic Probabilistic Entity-Relationship Model[13]等.

现有的关系学习研究大多是基于完备数据进行的,而现实问题中,数据通常是不完备的.在传统的机器学习领域中,从不完备数据中学习的问题已经得到了研究[14,15],但不完备的关系数据问题非常复杂,因此,几乎没有任何一项技术可以直接被扩展到关系学习领域.传统的机器学习算法可以被看成是数据集中仅有的一个表,并且不存在关系的学习算法.例如,Bayesian 网络可以看成是仅包含一个属性类,并且不存在关系的 PRM.因此,PRM 结构学习的复杂度至少相当于 Bayesian 网络学习的复杂度.由于具有多个局部极值,如果将传统的机器学习中处理不完备数据问题的算法直接扩展到关系学习中,学习的复杂度将会明显提高,并且会得到较差的结果.因此,从不完备的关系数据中学习是关系学习领域中一个重要的、有待解决的问题.

本文提出一种从不完备关系数据中学习概率关系模型 PRM(probabilistic relational model)的方法——MLTEC(maximum likelihood tree and evolutionary computing method).首先,随机填充不完备关系数据得到完备关系数据,然后从每个随机填充后的数据样本中分别生成最大似然树(maximum likelihood tree,简称 MLT)[16],并将这些 MLT 作为初始 PRM 网络群体,最后通过进化计算,利用进化过程中最好的网络结构反复修正不完备数据集,从而得到 PRM 的结构.第 1 节介绍相关的研究背景,然后详细描述 MLTEC 方法,最后进行实验并总结全文.

1 背景知识

PRM[6]是在传统“属性-值”方式的 Bayesian 网络基础上,将其扩展成复杂的关系结构,是从关系数据库中学习到的概率模型.给定一个实例集合以及这些实例之间的关系,PRM 表示这些实例属性的联合概率分布.

1.1 关系框架

关系框架包含属性类集合$\chi=\{X_1,\ldots,X_n\}$和关系集合 $R=\{R_1,\ldots,R_m\}$.每个属性类中包含若干属性.属性类 X 的属性集合记作 $A(X)$,X 中的属性 A 记作 $X.A$.

关系框架的骨架结构σ是框架的一个实例表示.对于每个属性类,$0^{\sigma}(X_i)$表示属性类中固定属性的值以及它们之间的关系,但没有描述概率属性的值.完备实例 L 将骨架结构σ进行扩展,对概率属性的值也进行了描述.

1.2 概率模型

概率关系模型 PRM 由以下两部分组成:依赖结构 S 和与之相关的参数θ_S.依赖结构 S 是通过关联属性 $X.A$ 及其父节点集 $Pa(X.A)$定义的.属性 $X.A$ 可以依赖于属性类 X 的其他属性 B,也可以通过关系链 τ依赖于其他属性类的属性 $X.\tau.B$.当关系不一一对应时,模型引入数据库理论中集合的概念来解决这一问题.

定义 1[6]. 关系模式下,对每个属性类 X 及其属性 $X.A$ 的概率关系模型 PRM 定义为:

- 父节点集 $Pa(X.A)=\{Pa_1,Pa_2,\ldots,Pa_n\}$,其中,每个 Pa_i 具有 $X.B$ 或$\gamma(X.\tau.B)$两种形式,τ和$\gamma()$分别为关系链和集合函数.
- 条件概率模型 $P(X.A|Pa(X.A))$.

一个 PRM 的联合概率分布如公式(1)所示,为

$$P(L\mid\sigma,S,\theta_S)=\prod_{X_i}\prod_{A\in A(X_i)}\prod_{x\in 0^{\sigma}(X_i)}P(L_{x_i.a}\mid L_{Pa(x_i.a)}) \tag{1}$$

© 中国科学院软件研究所 http://www.jos.org.cn

1.3 相关工作

正如文献[6]中指出的那样,从完备数据中学习 PRM 有两个方面的任务:参数估计和结构学习.在参数学习中,首先假设模型结构已知.也就是说,算法的输入必须包括依赖结构 S 和训练数据集.而结构学习不需要额外的输入,学习的目的是从训练数据集中自动得到一个完整的 PRM 结构.显然,结构学习是 PRM 学习的核心内容.

结构学习是一个富有挑战性的问题,其主要困难在于如何从众多可能的结构中寻找到最适合的依赖结构.大多数结构学习算法主要关注这样 3 个问题:假设空间、打分函数、搜索算法.对于 Bayesian 网络来说,找到具有最高打分的网络结构是 NP 问题[17].而 PRM 学习的难度至少等同于 Bayesian 网络学习.因此,需要引入一些技术(例如,启发式搜索)来寻找具有最高打分的结构.这类方法成功的关键在于潜在父亲节点集的选择.显然,错误的初值将会使最终的结构较差.

近年来,对于 PRM 的应用及扩展方面的研究越来越多.Getoor 和 Sahami[18],Newton 和 Greiner[19]将 PRM 应用到 collaborative filtering.Getoor 等人[20]将 PRM 应用到超文本分类.Tasker 等人[21]提出一个基于 PRM 的分类和聚类模型,在文献[21]中也提到对于不完备关系数据的参数学习问题,但并未涉及从不完备关系数据中学习模型结构的问题.Getoor 等人[22]通过建立属性和与其相关的结构之间相互关系的模型将 PRM 进行扩展.Sanghai 等人[23]将动态 Bayesian 网络加以扩展,将每个时间片表示成一个 PRM 等等.但是,现有的关系学习研究大多是基于完备数据进行的,几乎没有其他算法能够解决从不完备关系数据中学习 PRM 结构的问题.

通常,从不完备数据中学习 Bayesian 网络需要对不完备数据进行估计.一类是基于蒙特卡洛或采样[15]的方法.使用这种方法能够得到非常准确的结果,但计算复杂度随变量增加呈指数集级增长,并且变量较多时很难实现.另一类是基于 EM(expectation-maximization)算法的方法[14].对于从不完备数据中进行参数估计来说,标准的 EM 算法具有很强大的能力.结构学习中潜在、可能的结构数量巨大,对于 EM 算法中的 E 步来说,如何有效地确定当前结构是否适合非常困难.文献[24]中指出,由于搜索空间巨大且具有多个局部极值,此类方法通常易于陷入局部最优.

2 MLTEC 算法

能够有效地评价不同的网络结构以找到与数据样本匹配程度最高的依赖关系模型,是从完备数据中学习依赖关系模型的关键问题之一.我们可以利用打分函数来选择网络结构,例如,MDL(minimum description length)标准[25].MDL 标准源于信息论中的交叉熵.MDL 标准综合考虑网络结构和数据样本的描述长度,试图找到一个既简洁又精确的网络结构.与其他打分函数一样,对于完备数据来说,MDL 标准是可以分解的.一个 PRM 模型的 MDL 打分是模型中每个属性 $X.A$ 的父亲结点集 $Pa(X.A)$MDL 打分的总和.PRM 模型的 MDL 打分函数如公式(2)所示:

$$MDL(S;L)=\sum_i MDL(X.A,Pa(X.A)) \tag{2}$$

根据 MDL 标准的可分解性,公式(2)可以分解为关于每个节点的父节点集的独立因式,如公式(3)所示:

$$\begin{aligned} MDL(S;L)=N\sum_{X_i A\in A(X_i)}\sum_{x\in 0^{\sigma}(X_i)}\sum P(X_i.A,Pa(X_i.A))\log P(X_i.A,Pa(X_i.A))- \\ \sum_{X_i A\in A(X_i)}\sum\frac{\log N}{2}\|Pa(X_i.A)\|(\|X_i.A\|-1) \end{aligned} \tag{3}$$

其中,N 是数据样本的大小,$\|X_i.A\|$表示 $X_i.A$ 所有可能取值的个数,$\|Pa(X_i.A)\|$是节点 $X_i.A$ 的所有可能父亲节点集取值的个数.

在数据不完备的情况下,MDL 标准无法分解成像公式(3)那样只与局部结构相关的因式,因此,结构学习问题变得更为复杂.

如果能够将缺失数据准确填充,那么,MDL 标准就可以评价不同的网络结构以找到和数据样本匹配程度最高的网络结构.传统的机器学习算法在处理不完备数据集时,通常是通过对不完备数据进行预测,使不完备数据完备化.然而,由于 PRM 结构复杂,直接将传统机器学习处理不完备数据的方法直接扩展到关系学习中,计算复

© 中国科学院软件研究所 http://www.jos.org.cn

杂度高并且学习到的结果差.

MLTEC 利用进化计算的方法填充缺失数据,简化了计算复杂度,从而可以从不完备关系数据中学习到 PRMs 网络结构,此方法适用于大规模数据集.首先,随机填充不完备关系数据并得到完备的关系数据,然后从每个随机填充后的数据样本中分别生成最大似然树,并将这些最大似然树作为进化计算的初始 PRM 网络进化.算法选择 MLT 作为初始 PRM 网络是由于 MLT 是与 Bayesian 网络具有最好拟合结构的树状结构,并且具有结构简单、不确定性推理效率高等特点.

Chow 和 Liu[16]提出一种著名的学习树状 Bayesian 网络的方法.他们将建立 MLT 的过程简化为建立一个最大权重生成树的过程.文中将 Chow 和 Liu 的算法扩展到关系学习中,见表 1.

Table 1 The procedure for constructing an MLT

表 1 建立 MLT 的过程

1. Compute $I(X_i.A;X_j.B)$ between each pair of attributes, $A \neq B$, where $I(X;Y)=\sum_{x,y}P(x,y)\log\frac{P(x,y)}{P(x)P(y)}$ is the mutual information function.
2. Build a complete undirected graph according to the weight of an edge connecting $X_i.A$ to $X_j.B$ by $I(X_i.A;X_j.B)$.
3. Build a maximum weighted spanning tree.
4. Transform the resulting undirected tree to a directed one by choosing a root attribute and setting the direction of all edges to be outward from it.

然后将这些 MLT 作为初始群体,再通过进化计算,利用进化过程中最好的网络结构反复修正不完备数据集,最后学习得到打分最高的 PRM 结构.

为了避免陷入局部极值,算法采用了扩展的进化规划(evolutionary programming,简称 EP)方法作为搜索算法.算法采用 3 种变异算子(增加边、删除边、转向边)产生后代,每次执行变异操作时,3 种变异操作以相同的概率被选择.由于自适应机制,传统的 EP 易于陷入局部最优值[26].因此,为了防止早熟收敛现象的发生,算法将重开始策略引入到 EP 中.

重开始策略的主要过程是:在进化过程中,动态地监控群体的多样性,当群体的多样性降到事先规定的界限之下时,就认为进化过程中出现了早熟收敛的趋势,然后对当前群体进行重新初始化,以恢复群体的多样性,使进化有效地继续进行.

MLTEC 仅重新初始化群体的一部分,这样,引入的重开始策略不仅能够较好地保留已获得的有效信息,同时又能够有效地避免早熟收敛,为下一轮进化奠定良好的基础.MLTEC 方法见表 2.

Table 2 The MLTEC approach

表 2 MLTEC 方法

1. While i is not bigger than the initial population size PS
   a) Initialize the incomplete relational data randomly and obtain complete relational data.
   b) Generate a maximum likelihood tree from the complete relational data.
   c) Increase i by 1.
2. Set to 0.
3. Create an initial population with the maximum likelihood trees, $Pop(0)$, of PS PRMs.
4. Each PRM in the population $Pop(0)$ is evaluated by using the MDL metric.
5. While t is smaller than the maximum number of generations G
   a) Each PRM in $Pop(t)$ produces one offspring by performing mutation operations. If the offspring has cycles, delete the set of edges that violate the PRM condition. If choices of set of edges exist, we randomly pick one choice.
   b) The PRM in $Pop(t)$ and all new offspring are stored in the intermediate population $Pop'(t)$. The size of $Pop'(t)$ is $2\times PS$.
   c) Conduct a number of pair-wise competitions over all PRMs in $Pop'(t)$. Let S_i be the PRM being conditioned upon, q opponents are selected randomly from $Pop'(t)$ with equal probability. Let S_{ij}, $1\le j\le q$, be the randomly selected opponent PRMs. The S_i gets one more score if $D_i(S_i)\le D_i(S_{ij})$, $1\le j\le q$. Thus, the maximum score of a PRM is q.
   d) Select PS PRMs with the highest scores from $Pop'(t)$ and store them in the new population $Pop(t+1)$.
   e) Re-Initialize parts of the population if the population variety decreases to a certain finitude.
   f) Modify dataset by utilizing the best evolved PRM structure in $Pop(t+1)$.
   g) Increase t by 1.
6. Return the PRM with the highest score found in any generations of a run as the result of the algorithm.

© 中国科学院软件研究所 http://www.jos.org.cn

3 实验测试

实验选择 3 个问题域对 MLTEC 进行测试,包括两个真实问题域和一个模拟问题域.图 1 是 School 问题域的结构图.根据图 1 所示结构生成用于实验的模拟数据.算法仅以数据集作为输入.

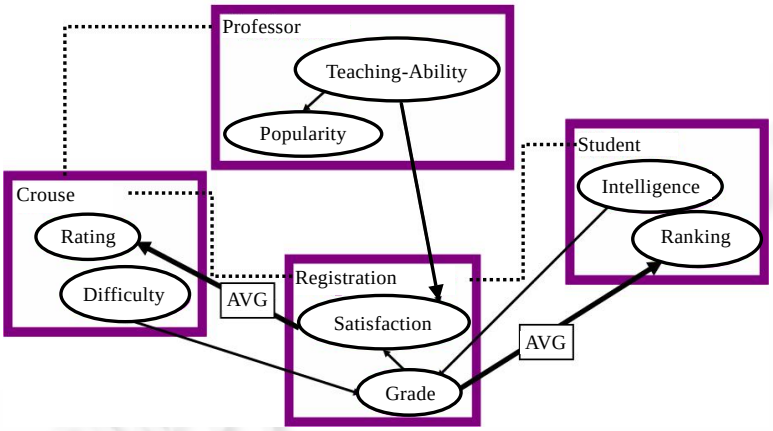


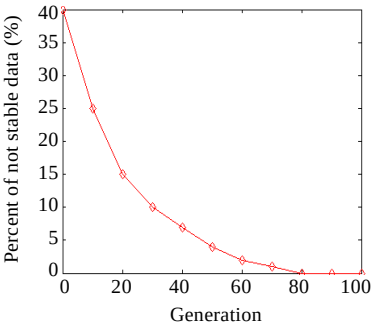
Fig.1 A PRM structure for the school domain

图 1 School 模型结构

生成具有 5 000 个数据样本的 4 个数据集(这里,样本数量是指学生数量).算法分别在具有 10%,20%,30%和 40%的丢失数据这 4 个数据集上进行测试,每个数据集上测试 10 次.丢失数据是分别从初始数据集中随机去除 10%,20%,30%和 40%属性值得到的.进化计算中,由于适应度函数的计算复杂度,进化计算中种群规模不宜过大,而为了预防早熟收敛现象的发生,种群规模又不宜过小.本文将种群规模 PS 设为 30,q 值设为 5.在进化过程中,动态地监控群体的多样性.由于本文的重点在于 PRM 的结构学习,因此,文中没有对重开始策略的参数选择进行过多的讨论,仅当种群中大部分个体收敛到几个点的时候对部分个体进行初始化.

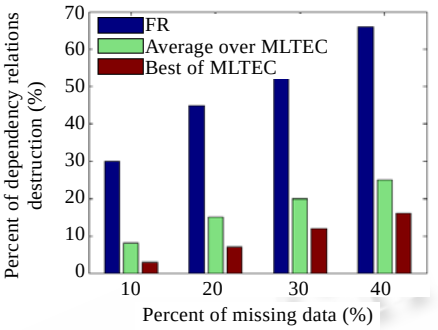
由于现存的方法中没有从不完备数据中学习 PRM 结构的方法,因此,文中用于比较的方法是先随机填充不完备数据,然后从得到的完备数据中学习 PRM 结构的方法(称之为 FR(fill randomly)).图 2(a)所示为不稳定的缺失数据平均百分比,也就是算法对具有 40%缺失数据的数据集进行 10 次实验,每代要修正的缺失数据平均百分比.可以看出,当算法开始时,随机填充不完备关系数据并得到完备的关系数据,这些填充数据带来很多噪声,需要修复的数据量很大.然而,算法通过将进化过程中最好的网络结构嵌入到不完备数据集中,有效地修复噪声数据.随着进化的进行,修正的缺失数据越来越少,数据趋于稳定并最终收敛.图 2(b)所示为分别使用 MLTEC 和 FR 方法学习的结果对依赖关系破坏的比较.对于 MLTEC 方法,图中分别给出了 10 次实验的平均和最优结果.从图中可以看出,用 MLTEC 实验的结果明显优于使用 FR 的结果.这并不难理解,丢失数据被随机初始化后,这些数据变成了噪声数据,变量之间的依赖关系遭到很大程度的破坏.但通过进化过程的反复修正以后,由于充分利用了原有数据所蕴含的信息,很大程度地把噪声数据变成了非噪声数据,因此,被破坏的依赖关系能够被有效地修复.图 2(c)所示为 MLTEC 在 4 个数据集运行 10 次的平均精度.精度计算是以图 1 为标准的,从增加和丢失边两方面进行比较.真实网络的精度设为 1.从图中可以看出,随着进化的进行,MLTEC 学习到的网络精度不断提高.通过实验发现,即使在丢失数据量很大的情况下,MLTEC 方法也并不会产生大量的冗余边.这是由于使用 MLT 作为进化计算的初始群体,因此,学习到的结构倾向于简单化.这同时也说明,MLTEC 方法可以通过引入一些策略来提高其在减少边的丢失方面的能力,也是将来要研究的课题.

© 中国科学院软件研究所 http://www.jos.org.cn



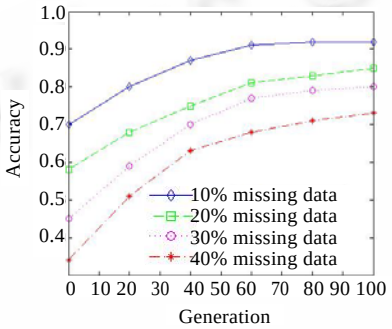
(a) Average percent of unstable missing data on the database with 40% missing data

(a) 具有 40%缺失数据的数据集中不稳定的缺失数据百分比



(b) Comparison of the destruction of dependency relations in structures learned from datasets

(b) 学习到的模型结构依赖关系破坏比较



(c) Average accuracy of the structures learned by MLTEC

(c) MLTEC 学习到的模型结构平均精度

Fig.2

图 2

然后,MLTEC 对两个真实问题域进行实验.第 1 个问题域是具有丢失数据问题的电影数据库(http://www-db.stanford.edu/pub/movies/doc.html),数据库中包含 4 个关系:Movie,Actor,Director 和 Role.我们从数据库中选择出一个含有 5 000 个 movie、3 000 个 actor 和 1 500 个 director 的子集,同时,此操作也会产生丢失数据问题.分别使用 MLTEC 方法和 FR 方法从子集中学习,其中,MLTEC 方法学习 10 次.如图 3(a)所示,MLTEC 学习 10 次中最优的结果为属性类 movie 的 Genre 依赖于 movie 的 Year 和 Process 以及属性类 director 的 Name、属性类 movie 的 Process 依赖于 movie 的 Year,同时也学习到一个能够关联所有表的重要依赖关系:在电影中演员扮演角色的 Role-Type 依赖于 actor 的 Gender 和 Rank 以及 movie 的 Genre.这个结果明显合理.图 3(b)所示为 MLTEC 学习 10 次的平均结果.与最优结果比较发现,丢失了属性类 director 的 Name 和属性类 movie 的 Genre 之间的依赖关系,并且增加了属性类 director 的 Name 和属性类 movie 的 Year 之间的依赖关系.图 3(c)所示为使用 FR 方法学习到的结构.与前两个结构相比,由于丢失数据的影响,FR 方法学习到的网络具有较多的冗余依赖关系.

第 2 个真实问题域来自 PKDD 2000 的竞赛[26].这是一个由 Czech 银行提供的数据得到的经济问题数据库,它描述的是 5 369 个客户持有 4 500 个账户的操作.银行希望从数据中发现有意义的客户群体以便提高他们的服务(例如,如何区分信用较高的客户和信用较差的客户).数据库中所包含的 8 个表分别为 account,client,disposition,permanent,order,transaction,loan,credit card 和 demographic data.本文主要研究客户信用问题.从数据库中选择一个包含 4 个关系的子集:account,client,loan 和 credit.使用与电影问题域同样的方法进行测试.如图 4(a)所示,MLTEC 学习 10 次中最优的结果为属性类 loan 的 Payment 依赖于 loan 的 Date,Amount 和 Duration、属性类 account 的 Balance 以及属性类 client 的 Credit cards owner or not.同时,也学习到一个能够关联所有表的

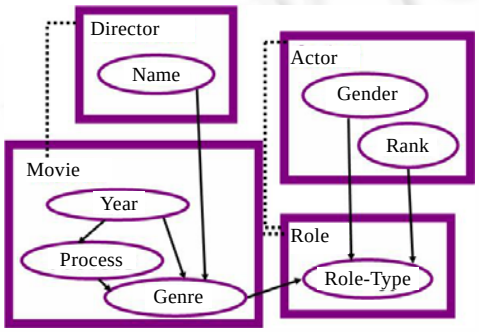
© 中国科学院软件研究所 http://www.jos.org.cn

依赖关系:属性类 client 的 Rank 依赖于 loan 的 Payment.图 4(b)所示为 MLTEC 学习 10 次的平均结果.与最优结果比较发现,丢失了属性类 loan 的 Date 和 Payment 之间的依赖关系,并且将属性类 loan 的 Payment 和属性类 client 的 Credit cards owner or not 之间的依赖关系反向.图 4(c)所示为使用 FR 方法学习到的结构.与前两个结构相比,这种丢失了属性类 account 的 Balance 和属性类 loan 的 Payment 之间重要的依赖关系,并且学习到的网络具有较多的冗余依赖关系.

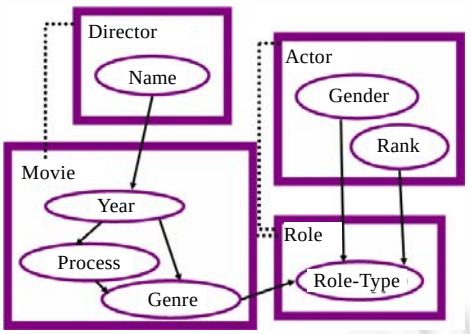
4 结 论

关系学习算法处理的是存在于关系数据库中的多表及它们之间的关系问题,而传统的基于“属性-值”表示的机器学习技术难以用于解决这类问题.现实问题中,数据通常是不完备的,现有的关系学习研究大多是基于完备数据进行的.本文提出一种从不完备关系数据中学习 PRM 结构的算法——MLTEC 方法.实验结果表明,MLTEC 方法能够从不完备关系数据中学习到较好的模型结构.

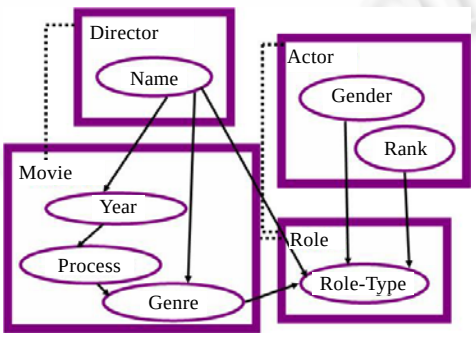
正如前面第 3 节曾经提到的那样,即使在丢失数据量很大的情况下,MLTEC 也不会产生很多冗余边.因此,今后一个有趣的工作是,从通过嵌入某些策略来减少丢失边的数量入手,以提高算法的有效性.另外,重开始策略各参数对 MLTEC 方法进化过程的影响,也将作为今后的研究任务.



(a) The best PRM structure of the movie domain learned by MLTEC
(a) MLTEC 方法学习到的 movie 主域最优 PRM 结构



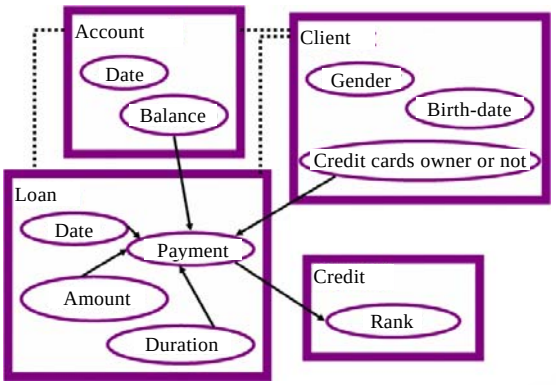
(b) The median PRM structure of the movie domain learned by MLTEC
(b) MLTEC 方法学习到的 movie 主域平均 PRM 结构



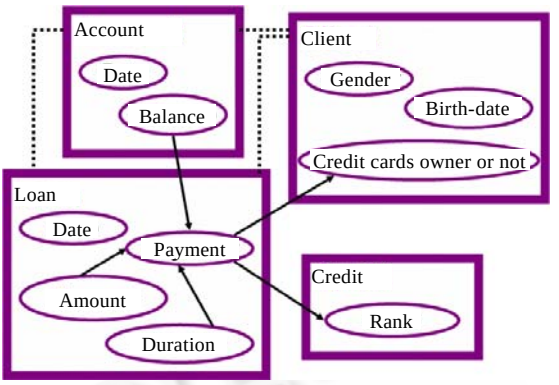
(c) The PRM structure of the movie domain learned by FR
(c) FR 方法学习到的 movie 主域 PRM 结构

Fig.3
图 3

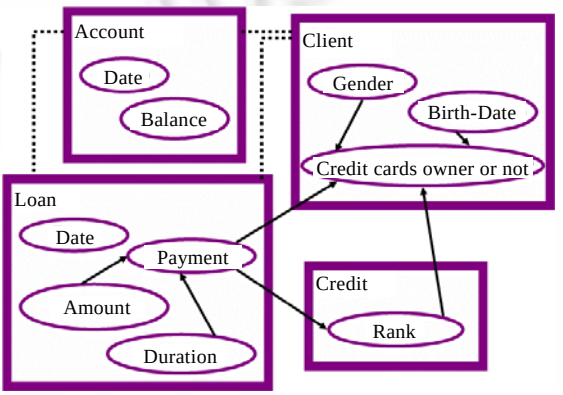
© 中国科学院软件研究所 http://www.jos.org.cn



(a) The best PRM structure of the financial domain learned by MLTEC
(a) MLTEC 方法学习到的 financial 主域最优 PRM 结构



(b) The median PRM structure of the financial domain learned by MLTEC
(b) MLTEC 方法学习到的 financial 主域平均 PRM 结构



(c) The PRM structure of the financial domain learned by FR
(c) FR 方法学习到的 financial 主域 PRM 结构

Fig.4
图 4

References:

[1] Džeroski S. Multi-Relational data mining: An introduction. SIGKDD Explorations, 2003,5(1):1−16.

[2] Muggleton S, ed. Inductive Logic Programming. London: Academic Press, 1992. 3−27.

[3] Heckerman D. Bayesian networks for data mining. Data Mining and Knowledge Discovery, 1997,1:79−119.

[4] Mu CD, Dai JB, Ye J. Bayesian network for data mining. Journal of Software, 2000,11(5):660−666 (in Chinese with English abstract).

[5] Blockeel H, Sebag M. Scalability and efficiency in multi-relational data mining. SIGKDD Explorations, 2003,5(1):17−30.

[6] Friedman N, Getoor L, Koller D, Pfeffer A. Learning probabilistic relational models. In: Proc. of the 16th Int'l Joint Conf. on Artificial Intelligence. Stockholm, 1999. 1300−1307. http://www.eecs.harvard.edu/~avi/Papers/lprm-ch.ps

[7] Muggleton S. Stochastic logic programs. In: De Raedt L, ed. Advances in Inductive Logic Programming. Amsterdam: IOS Press, 1996. 254−264.

[8] Kersting K, de Raedt L. Basic principles of learning Bayesian logic programs. Technical Report, 174, Freiburg: University of Freiburg, 2002.

© 中国科学院软件研究所 http://www.jos.org.cn

[9] Anderson C, Domingos P, Weld D. Relational Markov models and their application to adaptive Web navigation. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Edmonton, 2002. 143−152. http://www.cs.washington.edu/homes/pedrod/kdd02a.pdf.gz

[10] Flach P, Lachiche N. 1BC: A first-order Bayesian classifier. In: Proc. of the 9th Int'l Workshop on Inductive Logic Programming. Bled, 1999. 92−103. http://citeseer.ist.psu.edu/cache/papers/cs/8565/

[11] Flach P, Lachiche N. Native Bayesian classification of structured data. Machine Learning, 2004,57:233−269.

[12] Richardson M, Domingos P. Markov logic networks. Technology Report, Seattle: University of Washington, 2005. http://www-lrn.cs.umass.edu/lab-lunch/papers/markov-logic-nets.pdf

[13] Heckerman D, Meek C, Koller D. Probabilistic model for relational data. Technical Report, MSR-TR-2004-30, Microsoft, 2004.

[14] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 1977,B39:1−38.

[15] Geman S, Geman D. Stochastic relaxation Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1984,6:721−742.

[16] Chow CK, Liu CN. Approximating discrete probability distributions with dependence trees. IEEE Trans. on Information Theory, 1968,14:462−467.

[17] Chickering DM. Learning Bayesian networks is NP-complete. In: Fisher D, Lenz HJ, eds. Learning from Data: Artificial Intelligence and Statistics V. Berlin: Springer-Verlag, 1996. 121−130.

[18] Getoor L, Sahami M. Using probabilistic relational models for collaborative filtering. In: Proc. of the Workshop on Web Usage Analysis and User Profiling under KDD'99 (WEBKDD'99). San Diego, 1999. http://www.cindoc.csic.es/cybermetrics/pdf/201.pdf

[19] Newton J, Greiner R. Hierarchical probabilistic relational models for collaborative filtering. In: Proc. of the ICML Workshop on Statistical Relational Learning and Its Connections to Other Fields. 2004. http://www.cs.umd.edu/projects/srl2004/Papers/newton.pdf

[20] Getoor L, Segal E, Taskar B, Koller D. Probabilistic models of text and link structure for hypertext classification. In: Proc. of the IJCAI-01 Workshop on Text Learning. Beyond Supervision, 2001. 24−29. http://www.seas.upenn.edu/~taskar/pubs/ijcai01-ws.ps

[21] Taskar B, Segal E, Koller D. Probabilistic classification and clustering in relational data. In: Proc. of the 17th Int'l Joint Conf. on Artificial Intelligence. Seattle, 2001. 870−876. http://genie.weizmann.ac.il/pubs/conference/ijcai01.pdf

[22] Getoor L, Friedman N, Koller D, Taskar B. Learning probabilistic models of link structure. Journal of Machine Learning Research, 2002,3:679−707.

[23] Sanghai S, Domingos P, Weld D. Relational dynamic Bayesian networks. Journal of Artificial Intelligence Research, 2005,24:1−39.

[24] Friedman N. The Bayesian structural EM algorithm. In: Proc. of the 14th Conf. on Uncertainty in Artificial Intelligence. Madison: Morgan Kaufmann Publishers, 1998. 129−138. http://www.cs.huji.ac.il/~nir/Papers/Fr2.pdf

[25] Heckerman D. A tutorial on learning with Bayesian networks. In: Jordan MI, ed. Learning in Graphical Models. Cambridge: MIT Press, 1998. 301−354.

[26] Siebes A, Berka P. Discovery challenge. Notes of the Workshop Held at the 4th European Conf. on Principles of Data Mining and Knowledge and Knowledge Discovery. Lyon, 2000. http://www.cwi.nl/events/conferences/pkdd2000/

附中文参考文献:

[4] 慕春棣,戴剑彬,叶俊.用于数据挖掘的贝叶斯网络.软件学报,2000,11(5):660−666.

李小琳(1978－),女,吉林长春人,博士后,主要研究领域为机器学习,数据挖掘.

周志华(1973－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为机器学习,数据挖掘,信息检索,模式识别,进化计算,神经计算.

© 中国科学院软件研究所 http://www.jos.org.cn