高阶逻辑下知识表示与聚类方法的研究 *

许 晟[1a], 杨 [1b,2], 唐志刚[2], 李琳娜[2], 杨炳儒[2]
(1. 江西农业大学 a. 职业技术师范学院; b. 计算机与信息工程学院, 南昌 330045; 2. 北京科技大学 信息工程学院, 北京 100083)

摘 要: 针对一阶逻辑在复杂结构数据环境中存在模式搜索空间庞大和不能发明新谓词的缺点,提出了使用类型化的高阶逻辑知识表示语言 Escher 去表示各种复杂结构的数据,利用其强类型语法有效地约束知识发现过程中模式的搜索空间和高阶的特点去解决新谓词构造的问题。设计了以 Escher 为基础的复杂结构数据中的知识发现过程和基于复杂结构数据的聚类算法,并以实验验证了其有效性。

关键词: 复杂结构数据; 一阶逻辑; 高阶逻辑; 知识发现
中图分类号: TP301 **文献标志码**: A **文章编号**: 1001-3695(2010)08-2878-04
doi:10.3969/j.issn.1001-3695.2010.08.017

Higher-order logic-based knowledge representation and clustering algorithm

XU Sheng[1a], YANG Jun[1b,2], TANG Zhi-gang[2], LI Lin-na[2], YANG Bing-ru[2]
(1. *a. Vocational & Technical Normal College, b. College of Computer & Information Engineering, Jiangxi Agricultural University, Nanchang* 330045, *China*; 2. *School of Information Engineering, University of Science & Technology Beijing, Beijing* 100083, *China*)

Abstract: The problems of predicate invention and utility are also difficult to solve and remain open problems in knowledge discovery based on first-order logic. Typed, higher-order logic knowledge representation formalism, Escher can express all kinds of complex structured data. It not only can provide strong guidance on the search for frequent patterns with its strong typed syntax, but also can resolve the problem of the invention of new predicates with its higher-order characteristic. It is fit for knowledge discovery in complex structured data. This paper investigated the knowledge discovery in complex structured data by employing Escher as knowledge representation formalism. In the case of algorithms, clustering of complex structured data was studied in it and experimental verification of its effectiveness.

Key words: complex structured data; first-order logic; higher-order logic; knowledge discovery

0 引言

随着机器学习与知识发现在诸多领域应用深度和广度的拓展,如计算生物学、医学、病毒营销、反恐、语义 Web、社会网络分析、普适计算等复杂结构数据领域,复杂结构数据中的知识发现已成为知识发现领域的核心问题。但是迄今为止,学术界对复杂结构数据还没有一个公认的定义。这里对复杂结构数据给出一个描述性的定义。

定义 1 复杂结构数据不能用属性—值语言描述,无法用传统数据挖掘技术处理的结构化数据称为复杂结构数据。

另一个比较显著的复杂结构领域就是面向对象数据库,如 CAD、CAE、CASE、CAM 系统,基于知识的系统,多媒体处理系统等领域都需要面向对象数据库模型提供关系数据库模型所不能提供的复杂数据结构和运算。本文研究涉及的复杂结构数据主要来自上述两个领域。

面向对象数据模型和面向对象数据库系统体现了丰富的数据结构和语义信息,如复杂的数据对象、类层次结构、类的继承与组合关系。一个对象的属性可以是任意复杂的类型。这些使得面向对象数据挖掘的知识表示机制不仅能表示复杂的对象,还要能体现对象之间的关系。而属性—值语言不能表示复杂类型的属性,一阶逻辑语言不能够表示对象之间的继承关系。

1 一阶逻辑

归纳逻辑编程 ILP 使用逻辑程序作为知识表示方式,这种知识表示语言是一阶谓词逻辑的子集,逻辑程序基本概念与关系数据库基本概念间有明确的对应关系。因此,归纳逻辑编程能够灵活方便地表示关系学习与多关系数据挖掘过程中的多关系数据、背景知识以及涉及多关系的复杂模式。

这是归纳逻辑编程技术与思想成为关系学习与多关系数据挖掘主要方法的直接原因。

随着归纳逻辑编程技术在更多领域应用的拓展,其自身的缺陷也逐渐地暴露出来:

a)由于一阶逻辑知识表示机制的表示能力强,挖掘算法的模式搜索空间非常巨大,算法的可扩展性是其应用的一个巨大弊端[1,2]。

收稿日期: 2010-02-02; **修回日期**: 2010-03-05 **基金项目**: 国家自然科学基金重点资助项目(69835001,60875029)

作者简介:许晟(1980-),江西南昌人,讲师,主要研究方向为数据挖掘;杨 (1970-),江西南昌人,副教授,博士研究生,主要研究方向为数据挖掘、知识工程、柔性建模(ycjun515@163.com);唐志刚(1976-),湖南永州人,博士研究生,主要研究方向为数据挖掘;李琳娜(1981-),河南郑州人,博士研究生,主要研究方向为数据挖掘;杨炳儒(1943-),天津人,教授,博导,主要研究方向为知识工程、柔性建模.

b)为了缩小挖掘算法的搜索空间,人们提出了各种偏置约束学习到的模式格式,这有可能会剔除那些非常有价值的模式[3]。

c)虽然一阶逻辑能够利用领域的背景知识指导挖掘过程,但在复杂结构化领域,一阶逻辑很难有效利用结构化信息引导挖掘过程[4]。

2　高阶逻辑

针对基于一阶逻辑数据挖掘技术的上述缺点,研究者们分别从知识表示机制、学习策略等角度出发,提出了各种解决方案。这方面最为成功的就是 Lloyd[5~7] 提出的类型化的高阶逻辑知识表示语言 Escher。Escher 语言是一个强类型的函数逻辑编程语言,其解决了一阶逻辑在复杂结构领域知识发现的弊端,非常适用于研究复杂结构领域的知识发现。基于高阶逻辑的知识表示语言 Escher 在复杂结构数据决策树学习[8~10]、核学习[11]、遗传编程[12, 13]方面都已取得了不错的效果。

与基于命题逻辑与基于一阶逻辑的知识表达方式相比,其具有如下特征:

a)高阶逻辑能详细地声明假设语言,这一点对学习过程非常有用。

b)高阶逻辑消除了变量带来的麻烦——变量只需用来定义重写,以后便不再出现。

c)可设计直接处理具有复杂结构数据的归纳学习算法,使其以更加直接、自然、简洁的方式应用到复杂结构领域中。

d)支持各种数据类型,如集合、多集及图等任意复杂的类型,能描述复杂结构的样例。

e)学习到的归纳定义具有统一的表示方式,并且更易于理解,能捕捉数据的本质特征。

f)类型系统能够自然地包含类型信息及其语义信息。

g)样例空间中的每一个样例都用一个封闭的项进行描述,从而将样例的所有信息集中在一个位置,有利于学习过程中这些信息的使用。

h)类型系统能够阻止逻辑变量被无意义地归一,限制变量的实例化。更重要的是,每一个类型都决定了可对其进行的操作,可利用这些操作构造归纳假设,因此从一定程度上约束了假设语言,缩小了归纳假设的搜索空间。

i)能以统一的方式处理函数和谓词。

笔者发现在 Escher 基本概念与面向对象数据库基本概念之间也存在着对应关系,从而提出了将归纳逻辑编程中使用的一阶逻辑语言替换为 Escher 语言,得到的高阶归纳逻辑编程技术非常适合于研究面向对象数据库中的知识发现。表 1 展示了这种对应关系。

表 1　面向对象数据库术语与 Escher 术语之间的对应关系

面向对象数据库术语	Escher 术语
类	类型
属性	类型
方法	函数
对象	Escher 项定义的个体
类:对象的集合	类型:个体的集合
类间的继承/组合关系	构造更加复杂类型的数据构造器

从表 1 可以看出,面向对象数据库中的类概念是一个基本的类型,如 Nat,它对应于 Escher 中的基本类型。当面向对象数据库中的类概念是复杂类型时,其对应于 Escher 中由数据构造器构造的复杂类型。由于一些类型可以构成更复杂的类型,合成的类型可以用来表示面向对象数据库中类之间的继承和合成关系。由于 Escher 的每个类型都有对应的操作,这些操作对应面向对象数据库中类的方法。在面向对象数据库中,一个对象是某一个类的实例;在 Escher 中,一个项是某一个类型的实例,因此面向对象数据库中的对象对应于 Escher 的项。在面向对象数据库中,类是具有相同属性和行为的对象的抽象;在 Escher 中,某一个类型的所有个体都具有相同的属性和可操作的函数。本文通过 Mutagenesis 数据集[14] 来详细论述这种对应关系。

例如,Mutagenesis 数据集的面向对象数据模型与其 Escher 表示之间的对应。Mutagenesis 数据集是一个描述分子结构的数据集,每一个分子由原子及原子之间的结合关系构成。

在面向对象数据模型中,可将该数据集表示为三个类:

```
ClassMolecule{
classAtom}
Ames:Float,
Element: String,
Mutagenicity: String,
Type: Int,
Atoms:setofAtom}
Charge: Float,
Bonds: set of Bond }
class Bond {
Type: String };
```

在 Escher 表示机制下,该数据集合用如下类型表示:

```
Bond-type = String; Bond = Bond-type; Element = String; Atom-type = Int;
Charge = Float; Bonds = {Bond};
Atom = Element × Atom-type × Charge × Bonds;
Ames = Float; Mutagenicity = String;
Atoms = {Atom};
Molecule = Ames × Mutagenicity × Atoms
```

可以看出,Molecule、Atom 和 Bond 在面向对象数据模式中是类,在 Escher 中是类型。在面向对象数据模式中,每个对象的属性在 Escher 中也是类型。类 Atom 与 Bond 之间的组合关系是通过类型 Atom 与{Bond}之间的数据构造器来体现的。因此,Escher 能表示任意复杂的类型,它为属性—值学习和归纳逻辑编程提供了一个统一的框架。

3　高阶逻辑环境下的距离计算

在高阶逻辑的知识描述方式下,一个实例用一个基本项进行描述,故实例之间的距离是基本项之间的距离。文献[7]根据高阶逻辑知识表示方式的特点,提出了距离计算方法。根据距离计算的方法,得到如下所示的距离计算算法。

算法　高阶逻辑下的距离计算

输入:利用 Escher 的知识描述方式所表达的基本项 s、t 以及构成 s、t 的类型构造器的类型。

输出:实数值 d, $d \in [0, 1]$。

if ¬ ∃(类型 α),使得 s, $t \in \beta_\alpha$, then

　$d = 0$

else

　if ∃(类型 T)使得 $\alpha = T_{\alpha 1 \cdots \alpha k}$ ∧ ∃(类型构造器 C)使得

　　$s = Cs_1 \cdots s_n$ ∧ ∃(类型构造器 D)使得 $t = Dt_1 \cdots t_m$ then if

$C \neq D$ then $d = 1$
else
$d' = 0$
for $i = 1$ to $\max(m, n)$ do
$d' = d' + 1/2 \times d(s_i, t_i)$
if ∃(类型 β,γ)使得 $\alpha = \beta \rightarrow \gamma$, then
$$d(s,t) = \frac{\Sigma_{r \in B_\beta} d(V(sr), V(tr))}{1 + \Sigma_{r \in B_\beta} d(V(sr), V(tr))}$$
if ∃(类型 $\alpha_1, \cdots, \alpha_n$)使得 $\alpha = \alpha_1 \times \cdots \times \alpha_n$, then
$$d(s,t) = 1/n \sum_{i=1}^{n} d(s_i, t_i)$$

4　高阶逻辑环境下的聚类算法设计

在高阶逻辑的知识表示方式和相应的距离计算方式下,本文分别讨论了复杂结构数据的 PAM(partitioning around medoids)算法。

PAM 算法是改进的 k-medoid 方法。该方法不容易被噪声和孤立点数据影响,但该方法仍然需要预先确定需要聚类的簇的个数。PAM 算法的基本思想为:首先找到 k 个 medoids,然后将每个非 medoid 点赋值为距它最近的 medoid 所属的簇。k 个 medoids 的评估使用:对象与它被分到的簇的质心的平均距离最小。为了搜索 k 个 medoids,该算法首先任选 k 个点作为初始点,然后在迭代过程的每一步交换选中的点 O_i 和没有被选中的点 O_h,交换的条件是导致聚类质量的提高。PAM 计算所有没被选中的点 O_j 的 C_{jih},根据 O_j 所属簇的不同情况,C_{jih} 定义如下:

a)O_j 目前属于 O_i 所代表的簇,$d(O_j, O_h) \geqslant d(O_j, O_{j,2})$,$O_{j,2}$ 是 O_j 的第二个最相似的质心。因此,若 O_h 代替 O_i 成为质心,那么 O_j 将被赋值为 $O_{j,2}$ 所代表的簇,则

$$C_{jih} = d(O_j, O_{j,2}) - d(O_j, O_i) \tag{1}$$

该值为一个非负值。

b)O_j 目前属于 O_i 所代表的簇,$d(O_j, O_h) < d(O_j, O_{j,2})$,因此,若 O_h 代替 O_i 成为质心,那么 O_j 将被赋值为 O_h 所代表的簇,则

$$C_{jih} = d(O_j, O_h) - d(O_j, O_i) \tag{2}$$

该值可能是正值,也可能是负值。

c)O_j 目前不属于 O_i 所代表的簇,$O_{j,2}$ 是该簇的质心,$d(O_j, O_h) \geqslant d(O_j, O_{j,2})$,若 O_h 代替 O_i 成为 medoid,那么 O_j 将被赋值为 $O_{j,2}$ 所代表的簇,则

$$C_{jih} = 0 \tag{3}$$

d)O_j 目前属于 $O_{j,2}$ 所代表的簇,$d(O_j, O_h) < d(O_j, O_{j,2})$,若 O_h 代替 O_i 成为质心,那么 O_j 将从被赋值为 $O_{j,2}$ 所代表的簇,变为被 O_h 所代表的簇,则

$$C_{jih} = d(O_j, O_h) - d(O_j, O_{j,2}) \tag{4}$$

该值总是负值。

综合上述四种情况,O_h 代替 O_i 成为 medoid 的总代价为

$$TC_{ih} = \sum_j C_{jih} \tag{5}$$

具体的 PAM 算法如下所示:

a)任选 k 个中心。

b)所有的 O_i、O_h 对象对,计算 TC_{ih}、O_i 目前是中心,而 O_h 不是。

c)选择 $\min_{O_i, O_h}$ 的 O_i, O_h,若 $TC_{ih} < 0$,O_h 代替 O_i 成为中心;否则,回到 b)。

d)否则,对所有没被选中为中心的点,找到最相似的代表点,停止。

实验结果表明,PAM 算法在小规模的数据集上运行效果非常好。它执行一次迭代的时间复杂度为 $O(k(N-k)^2)$,故其时间复杂度较高。由于本文实验所用的数据集的规模都比较小,采用该算法。若进行大规模的实验,只需采用该算法的改进算法 CLARANS,这也是笔者下一步工作的重点。

5　实验

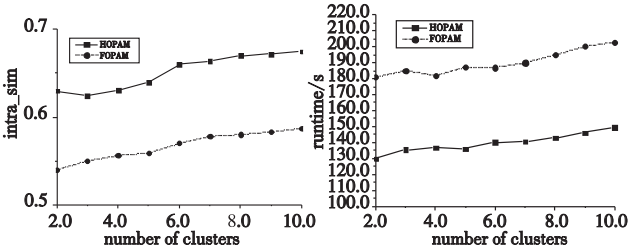
本文在复杂结构领域的人工数据集和真实数据集上,对上述介绍的聚类算法进行了实验。实验数据集和评价指标下面详细介绍,所有的实验运行平台是具有以下参数的 PC:赛扬 2.3 GHz CPU,2 GB RAM,Windows XP。实验采用了 Mutagenesis 数据集,该数据集中共有 188 个样例。这里利用数据集的固有特征和量值来评价一个聚类算法的聚类结果,数据集的结构未知。常用的内部度量方法有 Cophenetic 相关系数、簇内部相似度(intra-cluster similarity, intra_sim)等。本文采用簇内部相似度方法。

假设聚类的结果簇集为 C,其中的每个簇为 $C_i(1 \leqslant i \leqslant k)$,那么簇内部相似度定义为

$$\text{intra\_sim}(C) = \frac{\sum_{i=1}^{k} \sum_{x_i, x_j \in C_i, i \neq j} \text{similarity}(x_i, x_j)}{\sum_{i=1}^{h} |C_i| (|C_i| - 1)}$$

其中:$|C_i|$ 是簇 C_i 中包含的实例个数。

此处的实验主要是一阶逻辑表示方式下和 Escher 表示方式下的聚类结果的比较。在一阶逻辑知识表示方式下的距离计算采用 RIBL 系统中的距离计算方法。为了实验展示的方便, Escher 知识表示方式及对应的距离计算方式下执行的 PAM 聚类算法记为 HOPAM (higher-order PAM),一阶逻辑知识表示方式及对应的距离计算方式下执行的 PAM 聚类算法记为 FOPAM (first-order PAM)。图 1 是算法 PAM 在数据集 Mutagenesis 上的运行结果。



(a)intra_sim比较　(b)运行时间比较

图1　数据集在算法下的运行结果

从算法 PAM 在数据集 Mutagenesis 上的运行结果可看出:

a)在 Escher 知识表示方式及对应的距离计算下,聚类算法在执行效率和准确率上优于一阶逻辑知识表示方式下的执行结果。这是因为 Escher 的强类型语法能够自然而准确地体现结构化数据的语义信息,在 Escher 表示机制基础上定义的距离计算方法能够充分考虑数据的结构化信息,利用数据的语义信息;而一阶逻辑知识的表示机制不能够很好地体现数据的

类型化信息,其下的距离计算方法仅仅是对计算的复杂结构化数据之间的距离进行递归计算,直至最终转换为属性—值下的距离计算。

b)无论在哪一种知识表示方式下,PAM 算法的执行准确率都受初始 k 值的影响。这些都与属性—值知识表示方式下的执行结果相一致。

c)虽然实验中的数据集的规模比较小,但验证了 Escher 知识表示方式表示研究复杂结构领域数据聚类的可行性,拓宽了 Escher 的应用领域。

6　结束语

众所周知,一阶逻辑的知识表示方式虽然在机器学习的各个领域都取得了非常好的效果,但可扩展性是其在很多领域应用的一个弊端。Escher 作为一个比一阶逻辑知识表示方式表达能力更强的语言,虽然初步实验结果显示 Escher 知识表示方式下学习方法的效率优于一阶逻辑知识表示方式下学习方法的效率,但开发相应的可扩展性强的 Escher 知识表示方式下对应的学习方法仍是其在大规模领域应用的一项迫切需要。

参考文献:

[1] BLOCKEEL H, SEBAG M. Scalability and efficiency in multi-relational data mining[J]. ACM SIGKDD Explorations Newsletter, 2003, 5(1):17-30.

[2] DOMINGOS P. Prospects and challenges for multi-relational data mining[J]. ACM SIGKDD Explorations Newsletter, 2003, 5(1):80-83.

[3] BOWERS A F, GIRAUD-CARRIER C, LLOYD J W. A unifying view of knowledge representation for inductive learning [EB/OL]. (2005). http://users.rsise.anu.edu.au/~jwl/.

[4] FLACH P A, GIRAUD-CARRIER C, LLOYD J W. Strongly typed inductive concept learning [C]//Proc of the 8th International Conference on Inductive Logic Programming. [S.l.]: Springer-Verlag, 1998:185-194.

[5] LLOYD J W. Declarative programming in Escher, CSTR-95-013 [R]. [S.l.]: Department of Computer Science, University of Bristol, 1995.

[6] LLOYD J W. Programming in an integrated functional and logic language[J]. Journal of Functional and Logic Programming, 1999, 3(1):1-49.

[7] LLOYD J W. Knowledge representation, computation and learning in higher-order logic[EB/OL]. (2001). http://csl.anu.edu.au/~jwl.

[8] BOWERS A F, GIRAUD-CARRIER C. KENNEDY C, *et al*. A framework for higher-order inductive machine learning[C]//Proc of COMPULOGNet Area Meeting on Representation Issues in Reasoning and Learning. 1997.

[9] BOWERS A F. Early experiments with a higher-order decision-tree learner[C]//Proc of COMPULOGNet Area Meeting on Computational Logic and Machine Learning. 1998:42-48.

[10] BOWERS A F, GIRAUD-CARRIER C, LLOYD J W. Classification of individuals with complex structure[C]//Proc of the 7th International Conference of Machine Learning. [S.l.]: Morgan Kaufmann, 2000: 81-88.

[11] MONTANA D J. Strongly typed genetic programming[J]. Evolutionary Computation, 1995, 3(2):199-230.

[12] KENNEDY C J. Evolutionary higher-order concept learning[C]// Proc of the Genetic Programming Conference. 1998:22-25.

[13] KENNEDY C J, GIRAUD-CARRIER C. An evolutionary approach to concept learning with structured data[C]//Proc of the 4th International Conference on Artificial Neural Networks and Genetic Algorithms. 1999: 331-336.

[14] KING R D, MUGGLETON S, SRINIVASAN A. Structure-activity relationships derived by machine learning: the use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming[C]//Proc of National Academy of Sciences. 1996:438-442.

(上接第 2874 页)

4　结束语

本文针对区际救援物资中转网点定位—配给问题,建立了一种非线性混合整数规划模型,并设计了一种参照费用矩阵标杆、实行模块化变异操作的遗传算法。算例分析结果表明,该模型和算法能够较好地解决区际救援物资中转网点定位—配给问题。进一步的研究将不断改善该算法的性能,并尝试解决区际救援物资中转的多目标优化问题。

参考文献:

[1] 沈荣华.国外防灾救灾应急管理体制[M].北京:中国社会出版社,2008:134.

[2] TZENG G H, CHENG H J, HUANG T D. Multi-objective optimal planning for designing relief delivery systems[J]. Transportation Research Part E, 2007,43(6):673-686.

[3] SHEU J B. An emergency logistics distribution approach for quick response to urgent relief demand in disasters[J]. Transportation Research Part E, 2007,43(6):687-709.

[4] YI Wei, KUMAR A. Ant colony optimization for disaster relief operations[J]. Transportation Research Part E, 2007,43(6):660-672.

[5] MICHALEWICZ Z. Genetic algorithm + data structure = evolution programs[M]. New York:Springer-Verlag,1996.

[6] MICHALEWICZ Z, VIGNAUX G A, HOBBS M. A non-standard genetic algorithm for the nonlinear transportation problem[J]. ORSA Journal of Computing, 1991, 3(4):307-316.

[7] GEN M, LI Yin-zhen. Spanning tree-based genetic algorithm for bicriteria fixed charge transportation problem[C]//Proc of Congress on Evolutionary Computation. 1999:2265-2271.

[8] LI Yin-zhen, GEN M. Spanning tree-based genetic algorithm for bicriteria transportation problem with fuzzy coefficients[J]. Australian Journal of Intelligent Information Processing Systems, 1998, 4(3):220-229.

[9] 玄光男,程润伟.遗传算法与工程优化[M].于歆杰,周根贵,译.北京:清华大学出版社,2004.