ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.10, October 2007, pp.2507−2515
DOI: 10.1360/jos182507

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种本体和上下文知识集成化的数据挖掘方法*

陈 莹[1], 徐 罡[2+], 顾国昌[1]

[1](哈尔滨工程大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

[2](中国科学院 软件研究所 软件工程技术研究开发中心,北京 100080)

A Data Mining Approach Based on the Integration of Ontology and Context Knowledge

CHEN Ying[1], XU Gang[2+], GU Guo-Chang[1]

[1](College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)

[2](Technology Center of Software Engineering, Institute of Software, The Chinese Academy of Sciences, Beijing 100080, China)

+ Corresponding author: Phn: +86-10-62661581 ext 210, E-mail: xugang@otcaix.iscas.ac.cn

Chen Y, Xu G, Gu GC. A data mining approach based on the integration of ontology and context knowledge. ***Journal of Software*, 2007,18(10):2507−2515.** http://www.jos.org.cn/1000-9825/18/2507.htm

Abstract: Using ontology and context knowledge in data mining is one of the effective waies to improve data mining accurateness, which can add general knowledge and certain knowledge in decision factors. How to apply ontology and context knowledge in data mining is discussed in this paper. Firstly, the integration model of ontology and context knowledge is presented, which includes context information categories, context information extended on ontology models and context transformation method. Based on those, using the hierarchy structure of the ontology and context knowledge integration model as an example, the induced learning algorithm is presented in terms of the integration ontology and context knowledge. The experiment of the induced learning is presented and its result is more effective and accurate.

Key words: data mining; ontology; context knowledge; induced learning algorithm

摘 要: 在数据挖掘中使用本体和上下文知识能够将普遍的知识和特定的知识引入数据挖掘的决策因素中,是增进数据挖掘准确性的有效手段,同时也是数据挖掘领域研究的热点和难点之一.针对该问题,首先探讨了本体与上下文知识的集成化表示方法,包括上下文知识分类方法、如何在本体描述方法上扩展上下文知识及上下文知识转化方法.其次,以层次化结构的本体与上下文知识为例,构建了一个依据于本体和上下文知识集成的归纳学习算法并验证了该算法的有效性和准确性.

关键词: 数据挖掘;本体;上下文知识;归纳学习算法

中图法分类号: TP311 **文献标识码**: A

在数据挖掘中使用本体和上下文知识能够将普遍的知识和特定的知识引入数据挖掘的决策因素中,是一

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2007AA04Z148 (国家高技术研究发展计划(863)); the National Natural Science Foundation of China under Grant No.60573126 (国家自然科学基金); the National Basic Research Program of China under Grant No.2002CB312005 (国家重点基础研究发展计划(973))

Received 2006-05-22; Accepted 2006-08-17

种增进数据挖掘准确性的有效手段.但是,目前大多数数据挖掘方法在分析数据时都或多或少地忽略了这两方面的信息.本体提供在特定域内对数据信息的一致性理解,考虑相同信息的不同含义、不同信息相同含义的差异,从而达到目标系统正确处理信息的目的.上下文知识是在对数据信息一致理解的基础上,考虑不同使用者、不同应用场景所导致的对数据及其数值上的理解差异.可以说,本体解决了在数据挖掘中“普通语义”问题,而上下文知识解决了在数据挖掘中“应用语义”的问题.例如一个比较具体的例子,“ThinkPad,100,“05/08/08”,2003”表示 ThinkPad PC 的价格是 100 美元,购买日期是 2005 年 8 月 8 号,生产日期是 2003 年.使用本体,能够知道 ThinkPad 是 PC 的一种,也可对购买日期和生产日期正确地理解,问题是对于不同的应用环境,即便对数据语义有了一致的理解,也存在使用上的差异.如在特定的应用环境中,价格使用人民币来计算,购买日期使用月/日/年,而不是年/月/日.使用本体与上下文相结合的方法能够提供更精确的面向特定应用环境的数据挖掘.

概括地来讲,本体是对共享概念形式化的明确表示,从而使计算机能够解释处理信息的语义.共享和明确是本体的两个基本特征,共享是指本体表达了公认的知识,被一组人所接受;明确意味着这些概念以及概念使用中的限制具有明确的定义.上下文知识是指表达使用者使用数据或操作的特定场景的知识.可以说,本体提供一种泛化的知识,而上下文提供一种特例化的知识;本体是基础,上下文是扩展.本体通过抽取基本的概念元素而建立起来.对于上下文,由于上下文涉及众多场景,描述(特别是形式化描述)上下文始终是一个难题;其次,由于环境的不确定性,不可能穷尽地预见每一个特点应用场景,数据或信息提供者也不会预先提供所有这种不确定性需求.因此,在上下文建立和表示方面,需要提供一种自增量的渐进建模方法.在上下文知识描述上,提供一种抽象层次较高的建模元素,采用渐进建模方法逐步细化完善上下文知识是一种可行而有效的方法.

在数据挖掘过程中集成本体和上下文知识,对数据挖掘效果的提高是显而易见的.目前,国内外还缺少相关的研究成果.针对该问题,本文提出一种在数据挖掘过程中集成本体和上下文知识的数据挖掘方法,该方法兼顾数据语义和数据应用环境两方面,可以进一步提高数据挖掘的准确性.其中,在本体的基础上扩展上下文逻辑并提供一种方法来渐增地集成上下文知识,解决上下文环境的不确定性,维护一个较低成本的上下文建立过程.

本文第 1 节论述上下文知识的表示形态和方法以及与本体的集成化建模方法.第 2 节使用上下文逻辑(context logic)进一步形式化上下文知识,提供可操作的上下文知识表示和建立方法.第 3 节针对层次化结构的本体和上下文集成模型,提出一种依据本体和上下文知识的归纳学习算法.第 4 节比较相关研究.第 5 节总结全文并指出进一步的研究方向.

1 上下文知识表示

集成本体与上下文知识是提高数据挖掘精度的有效手段,可以应用在数据挖掘过程中的多个方面.例如:在确定数据挖掘的搜索空间中,我们可以使用本体知识来精确数据语义,使用上下文知识进一步约束搜索空间;在数据挖掘中,应用集成化的本体和上下文知识的首要问题是如何表示上下文知识以及如何与本体知识相结合.在这里,表示上下文知识主要考虑上下文知识的合理分类以及表示结构;如何与本体知识的结合主要考虑在已有的本体建模方法上如何扩展上下文知识.

1.1 上下文知识分类

在上下文研究领域,特别是在人工智能方面,通常采用组的方法来分类上下文知识,它允许将众多的属性值划分为少量的组,依据组来分类上下文知识,这是一种比较直接的方法.在基于组的上下文知识分类方法中,上下文知识界定为在本体知识基础上,描述那些导致本体知识在不同的上下文场景中存在差异的知识.这里,我们在组分类方法的基础上,进一步探讨采用绑定和约束、概念层次、网络结构 3 种方式来分类和结构化上下文知识.

定义 1. 上下文知识的绑定b和约束l,绑定b定义为$b=[b_{\min},b_{\max}]$,$b_{\min}$和$b_{\max}$分别表示绑定范围的最小值和最大值,相应地,绑定操作定义为$b_{\min}\times b_{\max}\rightarrow\tau$,其中,$\tau$为具有某种单位的确定值.约束$l$定义为规则$l$:antecedent→consequent,antecedent和consequent分别为前条件和后结果,antecedent和consequent可以具有多个值.

绑定用来直接约束目标值的取值范围.例如,在客户销售数据中,可以采用绑定来约束不同应用场景的客户

年龄组、客户登陆时间区及相对应的销售季节等等.而约束可以简单地看作是规则,一般是指属性关系规则表示数据值的限定,用来包含或排除特定的数据值.例如,在分析客户注册信息时,可以建立——包含“.edu”URL 的使用者的职业应该是学生、教师或研究人员——这样的规则.

定义 2. 上下文知识的概念层次h是一个互联的、无环的有向图,定义为$h=(L,E)$,这里,$L=\{l_0,l_1,l_2,l_3,\ldots\}$,$E=\{e_1,e_2,e_3,\ldots\}$.对于每个$e$有如下形式:$e=\langle l_i,l_j\rangle,l_i,l_j\in L,l_0\in L,l_0$的入度为 $0,l_1,\ldots,l_n$的入度为 $1.l_i$是l_j的子概念,即$l_i\subset l_j;l_i$是l_j的超概念,即$l_j\subset l_i$.

概念层次本身就已经包含了分组的思想,在分组的基础上又引入了结构化关系概念,这样可以使上下文知识在同一层达到分组的效果,在同一分支上,组与组之间又是细化、特定化的关系.可以方便而简单地组织特定应用场景的上下文知识.对于某个特定应用场景,我们认为同一层次的组与组之间的交集为空,即存在知识交叉,组与组之间的知识交叉(或是联系)体现在上级结点上.

定义 3. 上下文知识的网络结构w是一个有向、互联的有环图,定义为$w=(N,F)$,这里,$N=\{n_1,n_2,n_3,\ldots\}$,$F=\{f_1,f_2,f_3,\ldots\}$,每个n表示w中的一个顶点,每个f有如下形式:$f=\langle n_i,n_j\rangle,n_i,n_j\in N$.

考虑到在某些应用中,不同分支的组之间也隐含着信息联系,我们在上下文知识概念层次的基础上进一步构造网络结构.在网络结构中,结点代表上下文知识,结点之间的联系也表现为细化、特定化的关系,此外,还表示一种共同生成的关系,即具有多个父结点的结点是由多个父结点共同细化得到.

在数据挖掘中,绑定/约束、概念层次和网络 3 种分类方法分别适合于不同的应用情况,各有优缺点,绑定/约束适合于约束不同上下文环境的数据取值和对数据值转换;概念层次与网络结构相比,可以说是网络结构的子集,两者都可以表示复杂的上下文知识分类和联系.依赖于在数据挖掘中发现模式的不同,上下文知识可以定义为上面提到的一种形式或组合使用几种形式.例如,使用者的喜好也是数据挖掘中的一个重要方面,可以采用绑定和约束在概念层次和网络层次中组合使用的方法,来说明特定挖掘模式的上下限定.

1.2 本体和上下文知识集成化模型

目前,已有许多关于本体的表示方法,如 RDF(resource description framework),DAML(DARPA Agent mark-up language),DAML-S(DAML-service),DAML+OIL(DAML+ontology inference layer)),OWL(Web ontology language)等等.在本体与上下文集成化方面,我们采用在已有的本体表示方法上扩充上下文知识,可以认为上下文知识是本体知识的特定上下文环境的扩展,定义如下:

定义 4. 本体和上下文知识集成化模型 M 是一个三元组,即 $M=\{O,C,R\}$;其中,O 表示本体集合,C 表示上下文集合,R 表示本体间、本体与上下文间、上下文间的关系.

本体集合$O=\{o_1,o_2,o_3,\ldots\}$,其中,o_1,o_2,o_3等表示本体知识.上下文集合 $C=\{c_1,c_2,c_3,\ldots,K_{c_1},K_{c_2},K_{c_3},\ldots\}$,其中,$c_1,c_2,c_3$等表示上下文标识,一个上下文标识对应于一个上下文环境,使用它来识别不同的上下文环境,其本身又对应于多个上下文知识; K_{c_i} 表示属于上下文标识c_i的上下文知识集合.

关系$R=\{r_{oo},r_{oc},r_{ck},r_{kk},r_{cc}\}$,其中,$r_{oo}=\langle o_i,o_j\rangle$表示本体知识间的关系,不同的本体表示方法有不同的关系;$r_{oc}=\langle o_i,c_j\rangle$表示本体知识与上下文标识间的关系,它只表示单一语义“处于”;$r_{ck}=\langle c_i,k_j\rangle$表示上下文标识与构成该上下文标识下的上下文知识间的关系,“细化”、“绑定”和“约束”可作为其关系语义;$r_{kk}=\langle k_i,k_j\rangle$表示上下文知识间的关系,“细化”或“特例化”可作为其关系语义;$r_{cc}=\langle c_i,c_j\rangle$表示上下文标识间的关系,“转化”可作为其关系语义,“转化”表示从一个上下文环境到另一个上下文环境的变迁,通过“上下文公理”可以达到上下文环境转化的目的,我们将在第 2 节中详细论述.

简单来讲,与本体相集成的上下文知识是在本体知识上分配一个额外的上下文标识来表征一个特定的上下文环境,该上下文环境由附加的多个上下文知识构成.上下文知识可以采用层次结构表示,在上下文知识层次结构中,我们也可以包装部分上下文作为子树.图 1 为一个基于 RDF 本体模型扩展的上下文知识表示,该图是依据图 2 中的数据库模式而建立的本体与上下文知识集成化的元模型.其中,product_type 是 product 的子类型,本体知识 cost 连接两个上下文标识,上下文标识 context_1 连接 attribute 和 unit,binding 这 3 个本体知识.

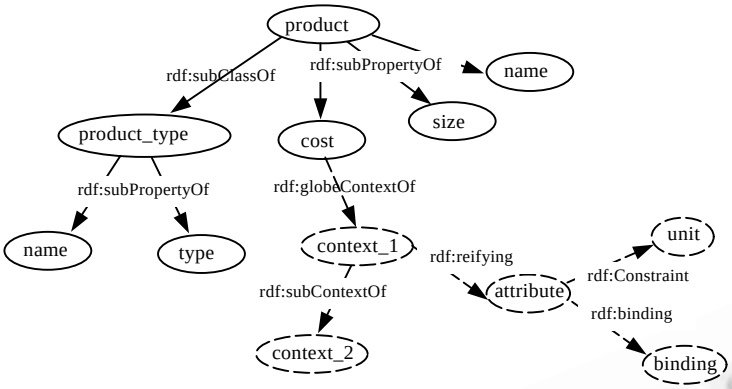

Fig.1 RDF with context knowledge extension

图 1 扩展上下文知识的 RDF

```
create schema    product
key      name:     char
         size:     int
         cost:     int;
creat schema     product_type
key      name:     char
         type:     char;
```

Fig.2 Example of DataBase scheme

图 2 数据库模式举例

2 基于上下文公理的上下文环境转化

由于上下文知识具有不可预见性,因此,需要在上下文知识表示的基础上提供一种上下文知识渐进的转化方法.本文采用上下文公理的方法建立上下文知识.在本体与上下文集成模型中,上下文公理对应于相关上下文标识和上下文知识间的连接,充当相关上下文标识以及附加知识的转化方法.上下文公理定义如下:

定义 5. c':*ist*(c,p).

上下文公理表示一种断言 *ist*,表示在上下文环境 c 中命题 i 为真,其本身在外部上下文环境 c'中.我们以关系数据为基础,研究如何在基于本体的基础上建立上下文知识.初始的本体知识可由数据源的模式自动产生,在此基础上,通过增加上下文公理使隐藏在应用场景中的上下文知识明显化.上下文公理是一阶逻辑的扩展,命题本身是逻辑单元并且可以量化.在此一阶逻辑的基础上,我们定义扩展多个上下文环境的公理,称之为上下文公理.上下文公理提供由一个上下文环境到另一个上下文环境转化方法以及同一个上下环境中的不同上下文知识间的转化,包括执行重命名、改变结构、显示化隐含条件等等.

这里,我们采用如图 2 所示的数据库模式来说明上下文公理的使用,图 3 为对应图 2 的一个实例.那么,针对该数据模式,对应不同上下文应用场景存在以下问题:

- 命名:在单一表中,属性语义存在多样性.例如,在产品数据库中的名称属性 name 包含的产品在语义上就存在很大差异,television_1 表示电视,而 simm_1 表示存储芯片.
- 属性:在单一表中,属性使用多样化.例如,在产品数据库中的尺寸属性存储的尺寸信息,它本身采用多种适合特定产品的尺寸单位,如:电视机的尺寸单位是英寸;存储器的尺寸单位是兆.
- 值:值不需要具有唯一表示,一个单一值可以表示在不同表中、不同列中的不同事物.例如,在产品数据库中,256 同时出现在尺寸列和成本列中,这本身并不存在矛盾.

Table: *product*

name	size	cost
television_1	14	256
simm_1	256	14

Table: *product_type*

name	type
television_3	televison
simm_1	memory_chip

Fig.3 Example of DataBase table

图 3 数据库表举例

针对该问题,我们建立了上下文公理,如图 4 所示.公理 1 转换本体O_1中的*product_type*属性到上下文环境C_1中,从而建立多个本体知识与上下文知识的联系.公理 2 建立*product*表中数组到上下文环境的转化,并明确化表中数据含义,即在本体O_1中的*product*数组对应转化在上下文环境C_1中,该上下文环境限定size使用*natural-size-units*作为单位,cost使用*us-dollar*作为单位.公理 3 和公理 4 给出针对两种情况的*unit*上下文公理,并联系*product_type*本体知识.

Axiom 1: $ist(C_1,product\_type(x,y))\Leftrightarrow ist(O_1,product\_type(x,y))$
Axiom 2: $ist(C_1,\exists y',z'(product(x,y',z')\ \&\ magnitude(y',natural\text{-}size\text{-}units(x))=y\ \&\ magnitude(z',us\text{-}dollar)=z))\Leftrightarrow ist(O_1,product(x,y,z)))$
Axiom 3: $ist(C_1,natural\text{-}size\text{-}units(x)=\text{bit}\times 1024\Leftarrow product\text{-}type(x,memory\text{-}chip))$
Axiom 4: $ist(C_1,natural\text{-}size\text{-}units(x)=\text{inch}\Leftarrow product\text{-}type(x,television))$

Fig.4 Example of context axiom
图 4 上下文公理举例

通过本体与上下文集成方法可以解决上述 3 个问题,即:使用本体可以解决表中属性语义多样性,使用上下文方法可以解决表中属性使用多样化,同时也自然地规避了“值不需要具有唯一表示”的问题.

3 基于本体和上下文知识集成的归纳学习算法

在数据挖掘中,归纳学习算法是一种数据驱动的、无优先级别的数据挖掘算法.目前,绝大多数支持模式分类的学习算法只是支持单一抽象层次的输入模式,并且采用一种有序的属性值排列方法.它们通常假定每一个模式属于一个特定的无交类中的一种,因此忽略不同类之间的任何关系,特别是本体与上下文知识相结合层次化的分类关系.实际上,数据驱动的模式发现依附于本体知识和特定应用场景.知识发现者所运用的本体知识和所处的上下文环境决定了数据选择以及数据间的关系,数据间关系反映了特定兴趣域.特别是在使用多个独立生成和管理的数据源作为数据挖掘的对象时,隐含在数据设计过程中的本体和上下文知识严重地影响不同使用者.例如,某个使用者在分析某种 PC 产品时,需要关注不同 PC 机的类型划分,而其他使用者并不关心 PC 机类型的划分.基于本体和上下文知识的归纳学习过程如图 5 所示,与传统的归纳学习方法相比较,我们可以看出,在数据源的基础上,有扩展本体和上下文知识作为输入量.

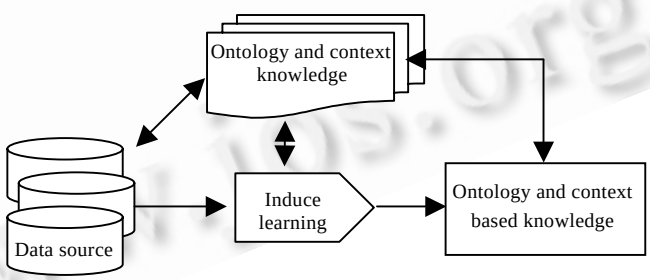

Fig.5 Using ontology and context knowledge in reductive learning
图 5 在归纳学习中使用本体和上下文知识

在基于本体和上下文知识的归纳学习算法中,主要考虑以下几种本体与上下文知识集成化建模情况:

- 本体知识可以附加一个上下文环境,上下文知识以概念层次结构表示.概念层次化的上下文知识是较为简单的形式,典型地可以表示为一棵树,每个结点表示一个上下文知识.两个属性结点间的连接表示一种细化关系.
- 多个上下文环境通过上下文标识互联,表示上下文环境转化.
- 只有一个全局上下文环境与之相对应,该全局上下文环境对应于数据源建立的环境.
- 每个本体知识可以与多个上下文环境直接或间接连接,本体知识处于全局或局部上下文环境中.

3.1 本体和上下文知识集成化决策学习算法

在构造基于本体和上下文知识的决策学习算法时,该算法是一个自顶向下多层的知识引导搜索过程.传统的决策树算法在每一步从多个候选属性中选择符合某种信息获取标准的属性.在我们的方法中,每个属性附带一个层次结构的本体和上下文知识分类.这样,学习算法不仅需要选择特定属性,而且需要选择适合的本体和上下文环境,因此它是一种面向特定使用者的决策树学习算法.

该算法的基本过程如下:

- 依据于本体树,对数据样本进行一次划分,得到基于本体的决策规则;
- 由于上下文知识依附于特定的本体,本体为上下知识的根结点,本体作为属性,得到基于上下文知识的数据样本的二次划分,得到基于上下文知识的决策规则,该决策规则是在本体决策规则上的进一步细化.

定义 6. $A=\{A_1,A_2,\ldots,A_n, A_1', A_2',\ldots, A_n'\}$为描述属性数据集合,其中,$A$为对应本体的属性集合,$A'$为与特定本体相联系的、对应上下文的属性集合;$O=\{O_1,O_2,\ldots,O_m\}$为类标记集合.对于每个数据$A_i$有一个附加本体和上下文知识结合的分类$T_i$,$T=\{T_1,T_2,\ldots,T_n\}$表示知识的集合.

其中,知识的根结点T_i是A_i.$\varphi(c)$表示结点c的子结点,S表示训练数据集合,每个叶子结点附加一个训练数据子集.指针n指向n个知识的n个概念,$P=\{P_1,P_2,\ldots,P_n\}$表示指针向量,其中,p_i表示T_i中的知识.如果每个p_i都是相应的知识T_i中的叶子结点,那么,指针向量称为末端向量.$\phi(p)$=true表示p为末端向量,即$\phi(p)$=true等价于(if and only if)$\forall p_i\in P,\varphi(p_i)=\{\}$.如图 6 所示.

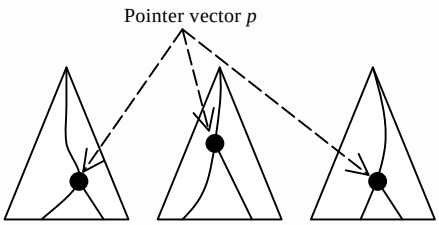

Fig.6 Taxonomies of ontology and context knowledge based on attribute

图 6 基于属性的本体与上下文相结合的知识分类

目前,采用在候选属性集合中选择属性所产生的最大熵的减少作为数据集合的分裂标准.当然,其他分裂标准也同样适用.

OCDT 算法. 基于本体和上下文知识的决策树学习算法.

OCDT(样本 *S*,属性 *A*,知识 *T*,类标示 *O*,指针向量 *P*,缺省值 *Df*,决策树 *T*)

1: if *T*=*null*, then initialize *T*,*S*,$P=\{A_1,A_2,\ldots,A_n, A_1', A_2',\ldots, A_n'\}$,*Df*=*Majority_Class*(*S*);

2: if *S*=*null*, then *node*=*Df*,

if *s* in *S* have same *Df*, then *s*′*Df*=*o*;

if $\phi(p)$=true, then *node*=*Df*;

3: call *Choose_Best*(*P*,*S*,*A*,*O*) //计算最佳属性B_j和最佳知识*b*;

4: *Bvalue*=$\varphi(b)$

5: For V_i in *Bvalue* //使用*Bvalue*的知识划分样本*S*

do

S_i=具有知识V_i的子集

$j=B_j$在*A*中的顺序

通过替换V_i为p_j,更新指针向量*P*为*P*′

调用*ODT*(S_i,*A*,*O*,*P*′,*Majority-Value*(S_i))构建子树

增加新的分支,标记为V_i,建立子树连接

End

6: Return 决策树

Choose_Best(指针向量 *P*,样本 *S*,属性 *A*,类标记 *O*)

(1) 对于每个属性,计算信息获取 *Gain*,并决定最佳属性及属性集合,用来划分数据集

(2) initialize *infogain*=0

(3) For *i*=1 to *n* do

$$Gain(S,p_i)=Entropy(S)-\sum_{v\in\Psi(P_i)}\frac{|S_v|}{|S|}Entropy(S_v)$$

If $Gain(S,p_i)>infoGain$ then $bestA=A_i, bestp=p_i$,

(4) Return *bestA*, *bestp*

基于本体和上下文知识的决策树学习算法可以看作是一种在假定决策树空间内优先搜索最优者的方法,决策树由本体和上下文知识决定.

3.2 算法举例

我们改造文献[1]中简单的客户购买数据库将其作为一个算法举例.数据库中的每个属性有一个对应的分类.该例中有两个 ISA 层次分类,分别为 Beverage 和 Snack.Beverage 分类有 4 个抽象层次,其中 3 个为本体抽象层次,1 个为上下文知识抽象层次,上下文知识抽象层次只包含一个上下文标识 Autumn,具有 Cold 和 Warm 两个上下文知识;Snack 分类有 2 个抽象层次,都为本体抽象层次,不包含上下文知识层次.类有 3 个值,分别为 Yong, Middle-aged,Old,如图 7 所示,表 1 为数据库数据举例.

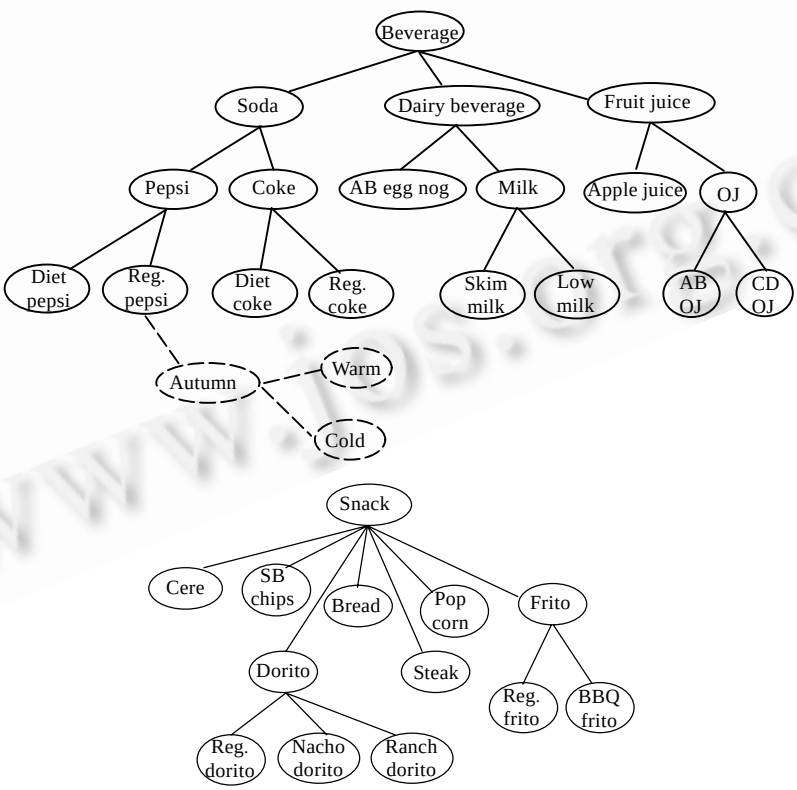

Fig.7 An example of ontology and context knowledge model

图 7 一个本体与上下文知识模型

Table 1 Sample in data

表 1 部分数据举例

Customer	Item 1	Item 2	ContextItem	Class
1	Reg pepsi	Ranch dorito	Cold	Young
2	AB OJ	CD cereal	null	Old
3	Reg pepsi	Reg dorito	Cold	Young
4	Reg coke	SB chips	null	Mid-Aged
5	Diet coke	Nacho dorito	null	Young
6	Reg pepsi	Ranch dorito	Warm	Mid-Aged
7	Reg pepsi	Ranch dorito	Warm	Mid-Aged
8	Skim milk	CD cereal	null	Old
9	Reg pepsi	Reg dorito	Warm	Mid-Aged
10	CD OJ	Bread	null	Old
11	Reg pepsi	Popcorn	Cold	Young
12	AB egg nog	CD steak	null	Old

该算法由指向分类根结点的指针向量开始,Beverage的信息量要高于Snack的信息量,因此,选择Beverage层次根结点的下一层划分数据集合,产生 3 种划分方式.在该数据中,有两个具有old类型的样本对应Dairy Beverage,另外两个具有old类型的样本对应Fruit juice.其他 8 个样本需要进一步划分,指针向量的第 1 个元素p_1改变为Soda,因此,值可以为Pepsi和Coke.指针向量的第 2 个元素p_2仍为Snack分类的根结点,可用的值包括所有下级值,因此,p_2较p_1产生的更大信息获取量,使用Snack分类进一步进行划分.在本体划分的基础上,指针向量的第 3 个元素p_3为对应上下文,该属性依附于p_1,对由p_1产生的规则进一步划分.结果决策树对应规则:IF Soda,Dorito and Cold Then Young;If Soda,Dorito and Warm Then Middle-aged;If Dairy Then Old;If Juice Then Old.与非包含本体知识的决策方法相比较,多出Cold,Warm等信息,因此获得的规则更加准确.

4 相关研究

上下文知识在人工智能、应用集成、移动计算等领域都有广泛的研究,近些年来取得了一些成果,但仍有许多难点需要解决.在数据挖掘领域,目前还很少涉及上下文知识,特别是与本体的综合利用,在数据挖掘中引入本体和上下文知识对数据获取精度的好处是显而易见的.Anand在文献[2]中综合性地分析了域知识的分类以及域知识在数据挖掘模式、约束搜索空间、过滤无用信息方面的使用.Anand所提出的域知识与特定的上下文环境有相近之处,所不同的是域知识较上下文知识更为广泛,而且并没有进一步将域知识划分为本体知识和上下文知识,进而分别处理.文献[3]也建议了一个对域知识的综合性分类,其分类方法中也包含了层次和网络结构,所不同的是缺少支持不同用户喜好的“绑定”和“约束”表示形式.在形式化方面,McCarthy首先探讨了形式化的上下文逻辑,其成果主要集中在对数据集成方面的使用上[4],而没有与本体进一步结合,并应用到数据挖掘中.

在数据挖掘的归纳学习算法方面,文献[5,6]采用预处理的方法处理数型结构属性,并重新编码训练样本.这些算法在一定程度上解决了数据的分类和归纳特定的规则,本文所不同的是引入了本体和上下文知识,并建立了统一的本体和上下文知识模型,依据该模型作为数据进一步分类和归纳特定规则的依据,因此可以依据通用的知识和特定的知识来提供数据挖掘的准确性.

5 结束语

在数据挖掘中使用本体和上下文知识是增进数据挖掘准确性及面向不同使用者和不同应用场景的有效方法,同时也是数据挖掘领域热点和难点之一.本文首先探讨了本体与上下文知识集成表示方法,包括上下文知识的分类方法、在本体描述方法上扩展上下文知识及上下文知识的构建;其次,以层次化结构表示的本体与上下文知识为例,构建了一个依据于本体和上下文知识的决策学习算法,探讨了本体和上下文知识在数据挖掘中的应用.

在进一步的工作中,在本体与上下文知识方面,我们重点研究对应不同的本体表示方法,上下文知识如何表示和建模以及集成机制;在数据挖掘方面,研究在有多种分布的数据源构成的数据仓库过程中,如何使用本体和

上下文知识集成化方法来屏蔽由多种数据源导致的异构性,以支持数据仓库的一致化数据格式.

References:

[1] Taylor M, Stoffel K, Hendler J. Ontology-Based induction of high level classification rules. In: Chaudhuri S, ed. Proc. of the ACM SIGMOD Data Mining and Knowledge Discovery Workshop. New York: ACM Press, 1997. 40−47.

[2] Anand SS, Bell DA, Hughes JG. The role of domain knowledge in data mining. In: Pissinou N, ed. Proc. of the 4th Int'l ACM Conf. on Information and Knowledge Management. New York: ACM Press, 1995. 37−43.

[3] Quinlan JR. Induction of decision tree. Machine Learning, 1986,1(1):81−106.

[4] McCarty J. Notes on formalizing context. In: Kehler T, Rosenschein S, eds. Proc. of the 13th Int'l Joint Conf. on Artificial Intelligence. Morgan Kaufmann Publishers, 1993. 555−560.

[5] Han J, Fu Y. Exploration of the power of attribute-oriented induction in data mining. In: Fayyad U, Shapiro GP, Smyth P, eds. Advances in Knowledge Discover and Data Mining. Cambridge: AAAI/MIT Press, 1996. 399−421.

[6] Han J, Fu Y. Dynamic generation and refinement of concept hierarchies for knowledge discovery in databases. In: Fayyad UM, Uthurusamy R, eds. Proc. of the AAAI'94 Workshop on Knowledge Discovery in Databases. Seattle: AAAI Press, 1994. 157−168.

陈英(1967－),男,北京人,博士生,主要研究领域为数据挖掘,人工智能.

顾国昌(1946－),男,教授,博士生导师,CCF会员,主要研究领域为人工智能,数据挖掘.

徐罡(1973－),男,博士,助理研究员,主要研究领域为分布式计算,中间件,软件集成.

****

中国计算机大会(CNCC2007)开始接受注册

中国计算机大会(China National Computer Conference,CNCC)是由中国计算机学会创立和主办的全国性学术会议，是目前国内计算机科学和技术领域中规模最大、级别最高的学术会议，所涉及的内容涵盖计算机科学技术的各个重要领域。会议将邀请国内、国际计算机领域的权威专家发表独特见解，把握宏观脉络；同时还将展示近期计算机领域最新技术，让学术界和企业界形成良好互动，共同探讨计算机科技的发展趋势。CNCC 已成为中国计算机界最具有影响力的年度盛会。她将为国内计算机界同仁提供一个研究和展示最新成果的交流舞台。本次大会将安排特邀报告、专题报告、企业专题论坛和论文交流，同时将举办 IT 技术展览。届时，大会将公布中国计算机学会王选奖获奖项目和中国计算机学会海外杰出贡献奖获得者名单。热烈欢迎各界人士踊跃参与。本次大会的主题是“网络改变生活，计算创造未来”。

主办单位是中国计算机学会和苏州市人民政府，由苏州市科学技术协会承办。

会议时间：2007 年 10 月 18 日～20 日

会议地址：中国苏州会议中心

大会网站：http://www.cncc2007.cn

联系人：吴树民，E-mail：ccf@ict.ac.cn，电话：010-62562503-17