关系数据库中的模糊知识发现*

杨雪南 李德毅

(中国电子设备系统工程公司研究所,北京 100039)

摘要 本文提出用部分量词和模糊谓词来表示模糊知识这一方法,在简要介绍数据抽象这一关系数据库中知识发现的方法后,详细描述了该方法中对模糊性问题的处理方法.

关键词 关系数据库,模糊集,知识发现.

专家系统是人工智能中获得最成功的领域,它的出现标志着人工智能进入了知识处理的时代.然而,它却遇到了很大的困难,其中最突出的是知识获取这一"瓶颈"问题.而关系数据库在80年代作为成熟的技术得到了相当广泛的应用,这些数据库隐含了大量的未知信息和规则,成为一个巨大的知识源.遗憾的是,这些知识难以通过数据库查询或统计展现在人们面前.因此,研究关系数据库中的知识发现,对于解决知识获取问题具有广阔的前景.

人类具有模糊思维的能力,能很好地把握对象的模糊性.我们知道,大部分领域中的许多知识是不十分清晰,即模糊的,这些知识对于人们决策(特别是宏观决策)具有非常重要的意义.因此,模糊知识的发现自然成为人们研究的重要内容之一.

1 数据抽象方法简介

1.1 基本概念

概念树是概念的一个层次结构.在概念树上,有的概念是模糊的,"上概念"的外延比"下概念"更广,更一般,树根相应于关系数据库的一个属性,取值为 *ANY*,即可取该属性域的任何值,而树叶的集合则构成属性的域,非叶结点称为抽象概念.例如,对于表1这个数据库,年龄的概念树如图1所示,其中,[1,…,100]是年龄的域,年轻,中年,老年是抽象概念,并且是模糊概念.同样,受教育程度的概念树如图2所示.

表1

姓名	性别	年龄	受教育程度	年收入
小王	男	28	大学	2050
…	…	…	…	…

* 本文1991-12-02收到,1992-09-13定稿

作者**杨雪南**,1966年生,工程师,主要研究领域为机器学习,数据库,管理信息系统.**李德毅**,1944年生,高级工程师,主要研究领域为智能数据库,模糊逻辑程序设计,专家系统.

本文通讯联系人:杨雪南,北京100039,北京丰台区郑常庄307号院软件中心

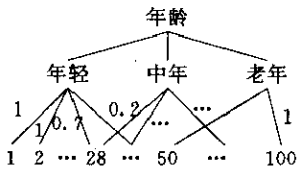

图1 “年龄”概念树

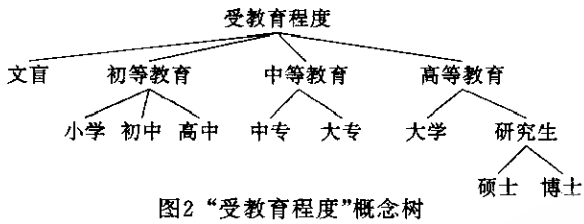

图2 “受教育程度”概念树

一个“元组”各属性的取值均取自于域,则称为原始元组;若其中某些属性值取为抽象概念,则称为抽象元组.例如,〈小王,男,28,大学,2050〉是一原始元组,而〈年轻,高等教育,2050〉则是一个抽象元组.

一个元组的覆盖度为 n,则该元组覆盖了数据库中的 n 个原始元组,显然,原始元组的覆盖度为1,因为它仅覆盖了它本身.

部分量词用来抽象描述从统计角度考虑的一些自然语言变量,是对统计结果在两个极端及中间状态的抽取.例如,极少部分、少部分、一些、基本上、大部分、绝大部分、全部等.原则上可以有无穷多个这种量词,但实际上选取有限几个就足够了.

我们用部分量词和模糊谓词来表示一条模糊知识.例如,“绝大部分年轻人收入较低”.

1.2 数据抽象方法简介

从数据库中进行知识发现的问题可以抽象为一个五元组:$\langle T,D,C,S,K\rangle$[1],其中 T 表示知识发现的任务;D 表示与 T 相关的数据;C 表示一组有助于发现特定知识的基本概念或背景知识;S 表示从数据库中发现知识的策略;K 表示发现的知识.

数据抽象方法根据知识发现的任务取得原始数据库,然后针对该数据库选择某一属性作为抽象属性,如果满足一定的条件,则对元组在该属性上进行抽象,用上概念取代当前的值,形成新的抽象元组,同时计算该抽象元组的覆盖度.这一过程在各属性上循环进行,直到满足结束条件.最后,根据所得的抽象元组形成知识.

2 模糊知识发现

发现更富有表现力的模糊知识需解决两个问题:

1.单值到模糊概念的抽象过程,它是模糊概念的单值模拟的逆过程.

2.部分量词的形成.

第一个问题仅限于第一类语言量,即具有精确定义的基本变量.我们知道,一个元素隶属于某个模糊概念不是是与否的关系,而是隶属的程度问题,并且同一元素对几个模糊概念可同时具有隶属关系.因此,当我们对数据元素进行抽象、形成抽象元组时需正确反映这种隶属关系.

下面给出这一过程的描述:(设属性 A 是进行模糊抽象的属性)

1.取下一待需抽象的元组 T,根据属性 A 及 $T.A$ 的值,从背景知识库中取得其隶属度大于0的概念集$\{X_1,X_2,\cdots,X_n\}$及相应的隶属度$\{U_1,U_2,\cdots,U_n\}$;

2.for ($k=0;k<T.N;k++$) {3,4,5};

3.for ($i=1;i<=n;i++$) {4,5};

4.产生一个[0,1]区间均匀分布的随机数 r;

5. 若 $r<=U_i$ 则形成新的元组 T'，$T'.A=X_i$，$T'.N=1$，其它属性的取值与元组 T 相同；

6. 若还有元组，则返回 1；

7. 所有元组抽象后，在不考虑属性 N 的情况下，对抽象后的元组（在新的关系库中）进行合并（因为此时有许多相同的元组），得到新的元组，这些元组其属性 N 的取值等于原来各与之相同的元组的该域取值之和.

（注：$T.A$ 表示元组 T 属性 A 的取值，属性 N 不是原来数据库的属性，而是新增加的，用来表示量的信息即覆盖度.）

从上述过程可知，对某一确定的数值元素，其抽象后的取值是不确定的. 存在三种情况：一是不作任何抽象，该元素所在的元组被丢弃；二是恰好用一模糊概念来抽象；最后一种情形是用几个模糊概念来替换. 因此，元组与抽象后的元组之间不是多对一而是多对多的关系. 从结果上看，一个数值元素如何抽象是不确定的，但在整体上与该元素的隶属度是符合的，反映隶属度这一量的关系. 例如，28 岁这一元素是“年轻”的隶属度为 0.72386，是“中年”的隶属度为 0.21322，则所有年龄为 28 岁的元组中，大约有 72.386% 的元组用“年轻”抽象，大约有 21.322% 的元组用“中年”抽象. 当数据库元组越多时，则符合得更好，而数据库中的知识发现其特点之一正是数据库常常具有相当大的规模.

覆盖度是能否进一步抽象的判定依据之一，也是发现最终反映给人的知识时形成部分量词的根据. 下面给出部分量词的形成过程：

1. 计算各元组的量的总和 $X=SUM(N)$；

2. 根据问题的角度，从数据库中选出某一类元组 $T_1,T_2,\cdots,T_n$. 计算各元组的统计量：$S_1=T_1.N/S$，$S_2=T_2.N/S,\cdots,S_n=T_n.N/S$，其 $S=T_1.N+T_2.N+\cdots+T_n.N$ 计算 $S_0=S/X$；

3. for ($i=0;i<=n;i++$) {4,5,6,7,8}；

4. 根据 S_i，取出其隶属度不为 0 的部分量词集合 $\{P_1,P_2,\cdots,P_m\}$ 及对应隶属度 $\{U_1,U_2,\cdots,U_m\}$；

5. 对 $\{U_1,U_2,\cdots,U_m\}$ 进行归一化处理，形成 $\{Q_1,Q_2,\cdots,Q_m\}$；

6. 形成区间 $I_1:(0,Q_1]$，$I_2:(Q_1,Q_1+Q_2]$，$\cdots$，$I_m:(\sum_{i=1}^{m-1}Q_i,1]$；

7. 产生一在[0,1]区间均匀分布的随机数 r；

8. 若 r 在区间 I_j，则形成的部分量词为 P_j.

（S_0 用来形成该类占总体的统计量信息.）

其中对第一步需加以说明，由于对同一抽象库，从不同的角度看，能得出不同的结论，因此，在第一步中应根据问题的角度来处理. 例如，下列抽象库：

受教育程度	年收入（元）	N
初等教育	1200 左右	100
初等教育	1500 左右	600
高等教育	1800 左右	80
高等教育	2000 左右	520

从受教育程度这个角度出发，分成初等教育和高等教育两类. 我们得到：

少部分受过初等教育的年收入在1200元左右；
大部分受过初等教育的年收入在1500元左右；

少部分受过高等教育的年收入在1800元左右；
大部分受过高等教育的年收入在2000元左右；

而从年收入角度出发，我们得到：

年收入在1200元左右的全部是受过初等教育者；

年收入在1500元左右的全部是受过初等教育者；

年收入在1800元左右的全部是受过高等教育者；

年收入在2000元左右的全部是受过高等教育者；

关于这些问题，在此不详述.

3 结束语

从大型数据库中获取知识是一个十分吸引人的课题. KDDB是我们在386机上用C语言，基于ORACLE关系数据库开发的一个知识发现系统. 本文主要介绍了数据抽象方法中模糊知识的获取方法，包括部分量词及模糊谓词的形成.

参考文献

1 李德毅，杨雪南. 关系数据库中的知识发现研究. 小型微型计算机系统，1992，13(4)：40—44.

2 Shapiro G P. Discovery of strong rules in databases. Proceedings of IJCAI—89, Workshop on Knowledge Discovery in Databases, 1989：264—274.

3 Cai Y, Cercone N, Han J. Learning characteristic rules from relational data—bases. Proceedings of the International Symposium Computational Intelligence, 1989：187—196.

4 Michalski R S. A theory and methodology of inductive of learning. Machine Learning：an Artificial Intelligence Approach, 1983.

FUZZY KNOWLEDGE DISCOVERY IN RELATIONAL DATABASES

Yang Xuenan Li Deyi

(*Institute of China Electronic System Engineering Company, Beijing* 100039)

Abstract In this paper, a method of fuzzy knowledge representation using partial quantifies and fuzzy predicates is proposed, a data—abstraction method of discovering knowledge from relational databases is briefly introduced followed by a presentation of an approach to deal with fuzziness in discovering.

Key words Relational databases, fuzzy set, knowledge discovery.