基于数据挖掘的非单调问题的缺省规则框架*)

郑宏珍[1] 刘 扬[1] 战德臣[2]
(哈尔滨工业大学计算机科学与技术学院 威海 264209)[1]
(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)[2]

摘 要 提出一个求缺省规则的框架,通过合并条件属性所决定的类,生成组合类,可以构造覆盖更多对象的规则,生成从这些组合类映射到占优决策的规则。结果规则比确定规则至少具有两个重要的优点:(1)结构上简单;(2)即使规则相对训练集可能不完全,但是当处理未见的新事例时将表现得更好。系统对未来对象的分类质量,将在很大程度上依据系统一般化知识的能力。

关键词 数据挖掘,缺省规则,关联规则

Default Rules Frame of Non-monotonous Problems Based on Data Mining

ZHENG Hong-Zhen[1] LIU Yang[1] ZHAN De-Chen[2]
(College of Computer Science & Technology, Harbin Institute of Technology ,Weihai 264209)[1]
(College of Computer Science & Technology, Harbin Institute of Technology,Harbin 150001)[2]

Abstract It proposes a frame in finding the default rules. Through the category decided by combining conditions attributes, the new combining category can create rules to cover more objects. These combining categories reflect the advantageous rules. In contrast to the fixed rules, the result rules have at least two main advantages: 1) simplicity in structure; 2) better performance in dealing with new cases despite its incompletion comparing with training set. System 's classification quantity to future object will according as ability of system generalized knowledge to a large extent.

Keywords Data minging,Default rule,Association rule

1 引言

非单调性问题,即指在规则体系中增加新的规则将可能导致原来的规则不成立。一般求解非单调问题的缺省规则(default rule)的形式为:

$$\neg B : C \rightarrow D$$

其中 B,C 为形如 $a=a_v$ 的原子公式的合取,a 为条件属性,a_v 为 a 的取值;D 形如 $(d=d_{v1})$ or $(d=d_{v2})$ or$\cdots(d=d_{vn})$,d 为条件属性,d_u 为 d 的取值,$i=1,\cdots,n$。

目前大多数数据挖掘算法没有涉及求解上述问题的非单调问题的缺省规则,但这样的缺省规则在实际中却有一定的应用背景。

对于非常大的信息系统,算法生成的规则可能非常多、非常复杂[1],也可能使用了无意义或冗余的属性或属性组合。而且,对训练集有效的规则,在一般情况下不一定正确。非常特例化规则不能很好地适应处理新实例,因此这些新实例可能与规则的条件不能精确匹配。覆盖训练线性系统中所有对象的规则,在处理其它输入数据时可能不适用。

Holte[2]认为短简单的规则更有效。事实上,仅有一个条件属性的规则其分类的性能几乎与 ID4 构造的精致决策树一样出色。

严格要求规则在许多实际应用中绝对正确是不够的。计算机系统以及人经常要求在严格的时间约束下做出决策,因此在缺少知识下能够进行推理非常有用。在数据库中发现这些规则的简单、自然的表示在各种分析和预测任务中是有用的。因此,知识发现的目标是获得一个规则集,能够模型化数据的一般特性,并较少受噪音的影响。

学习系统能够处理这种普遍出现的现象非常重要,即使出现不一致,也可能获取许多有意义的信息,特别是反映最一般或最正常情形的规则。因此,当考虑反映数据中非确定关系的规则时,必须进行折中,即牺牲规则的精确性换取规则的简单性。

缺省推理框架非常适合模型化常识推理和不确定决策的制定,即使在缺少知识时也能得出结论。信息系统中的不确定性,导致其本身非常自然地利用缺省规则的模型。缺省规则比确定规则更灵活,因为它对训练集中的错误数据具有较少的脆弱性。此外,缺省知识比确定知识更紧凑,一个适当选择的缺省规则集在应用于新实例时可能具有更好的性能。缺省关系可通过相对较少的不确定规则表示。而信息系统中搜索和表示的信息是不确定的。信息系统中的不确定性起源于数据中的不确定性(由于噪音),或者知识本身的不确定性,即信息系统中搜集和表示的信息的不确定性,这将导致其本身非常自然地适应于缺省规则的模型。缺省推理提供了关于不确定性推理的一个自然框架。缺省规则解释对象及其特性间的普遍依赖关系,允许在没有确定的规则可用时得出关于对象的结论。生成的缺省规则可以覆盖决策系统的不确定部分。下面首先定义一些基本概念。

2 相关知识

2.1 缺省规则

*)基金项目:国家 863/CIMS 主题资助项目(2003AA413021);高等学校博士学科点专项科研基金资助项目(20030213027)。**郑宏珍** 博士研究生,副教授,硕士导师,主要从事人工智能、数据库技术等方面的研究。**战德臣** 教授,博士生导师,主要从事 CIMS、数据库技术等方面的研究。

定义 1 给定决策系统 $A=(U,(C,D))$和阈值 μ_{tr}，$0\leqslant\mu_{tr}\leqslant1$，$A$ 的缺省规则形如：$(\tau\rightarrow\tau')\in F_{-}(C,D,V_C,V_D)$的分类规则，这里 $\mu(\|\tau\|_A,\|\tau'\|_A)\geqslant\mu_{tr}$。

值 $\mu(\|\tau\|_A,\|\tau'\|_A)$称为缺省规则 $\tau\rightarrow\tau'$的可信度。规则的可信度反映了应用规则时对规则的结论的信任程度。缺省规则 $\tau\rightarrow\tau'$是否被接受，取决于 $\|\tau\|_A$ 在类 $\|\tau'\|_A$ 中的隶属度是否大于阈值 μ_{tr}，$0\leqslant\mu_{tr}\leqslant1$。该阈值可以是预定义常量，或参数化变量，反映了用户希望的规则所具有的可信度。如果类 $E\in U/IND(C)$，不存在决策类 $X\in U/IND(D)$，使得 $\mu(E,X\geqslant\mu_{tr})$，则不生成类 E 的任何缺省规则。按照上述标准接受的缺省规则，被加入规则库中，即确定规则集，以适应决策系统的不确定部分。在确定系统中，由公式 τ 描述的类整个在由 $\|\tau'\|$ 描述的类中，所以 $\mu(\|\tau\|_A,\|\tau'\|_A)=1$。由定义知，任何确定规则也是缺省规则，如果阈值大于 0.5，则每个类最多生成一个规则，而小于或等于 0.5 时会生成相互不一致的规则。

2.2 信息系统的投影

定义 2 给定信息系统 $A=(U,A)$，对任何 $A'\subseteq A$，记表达式 $\pi(A,A')$为 A 到属性 $A-A'$的投影。因此，$\pi((U,A),A')\equiv(U,A-A')$。

符号“—”标记为集合差运算。因此投影返回从 A 中移去属性集后的结果信息系统，“新”关系定义于属性 $A-A'$。删除条件属性运算 $remC$ 是特殊投影操作，$remC((U,(C,D))C')=(U,(C-C',D))$，即从原始决策系统 A 中除去一个属性子集。结果变体 A'具有较少的条件属性，因此具有很大的类。从组合类得出结论，比覆盖孤立实例的类更具有一般性，这样的规则更具有意义。

记 C_{del}标记在 $remC$ 操作中删除的条件属性集，使用 C_{rem} 表示剩余的条件属性，即 $C_{rem}=C-C_{cut}$。

3 生成缺省规则的框架

假设决策系统的一个变体 $A'=(U,(C_{rem},D))$已通过去除属性 C_{del}的投影而得到，即 $C_{rem}=C-C_{cut}$。根据定义 1，如果给定条件，结论的条件概率高于阈值，则规则是可接受的缺省规则，即如果类 $E_{(C_{rem})}=U/IND(C_{rem})$在任意 $X\in U/IND(D)$中的隶属度大于等于阈值 μ_{tr}（即 $\mu(E_{(C_{rem})},X)\geqslant\mu_{tr}$），则就可以安全地生成缺省规则。新对象类 $E_{(C_{rem})}$可以被极小描述：$Des(E_{(C_{rem})},MinDes)$，其中 $MinDes$ 为差别函数 $f_{[C_{rem},D]}(E_{(C_{rem})})$的主蕴涵。类 E(定义于条件属性 C)到决策类 X 的候选缺省规则集标识如下：

给定决策系统 $A=(U,(C,D))$，对任意 $E\in U/IND(C)$，$X\in U/IND(D)$，从 E 到 X 的候选缺省规则集为：

$DBClass(E,X,C,D)=\{Des(E,redE)\rightarrow Des(X,D):redE\in red(E,C)\}$。

缺省分类的构造如下：

给定决策系统 $A=(U,(C,D))$，设 $C_{cut}\subseteq C,C_{rem}=C-C_{cut},A'=remC(A,C_{cut})=U(U,(C_{rem},D))$，则对 $0<\mu_{tr}\leqslant1$：

$Rules(A')=\bigcup\{DRClass(E_{C_{rem}},X,C_{rem},D):E_{(C_{rem})}\in U/IND(C_{rem}),X\in U/IND(d),\mu(E_{C_{rem}},X)\geqslant\mu_{tr}\}$。

下面给出分类规则与缺省规则生成算法的粗略描述。

Input：(1) 决策系统 $A=(U,(C,D))$

(2)控制规则可信度的阈值 μ_{tr}

Output：A 的结构化变体 A'的缺省分类规则

步骤如下：

(1) 选择基于属性的投影 $C_{rem}=C-C_{cut}$，使得每个投影为不确定粘合。

(2) 对每个投影 C_{cut}，生成 A 的变体，$A'=remC(A,C_{cut})$，以及新的复合类 $E_{(K,C_{rem})}\in U/IND(C_{rem})$，$k=1,\cdots,|U/IND(C_{rem})|$生成每个复合类 $E_{(k,C_{rem})}$的规则，检查下列公式是否成立，保证有特定的决策 X 占优，

$$\mu(E_{(k,C_{rem})},X)=\frac{|E_{(k,C_{rem})}\cap X|}{|E_{(k,C_{rem})}|}\geqslant\mu_{tr}。$$

(3) 生成类的规则的例外。这些类 $E\in E_{(k,C_{rem})}/IND(C_{cut})$，但映射到非占优决策类，即 $E\cap X=\phi$。

(4) 重复每个变体 $A'=remC(A,C_{cut})$。

通过计算各层的隶属值，可以限制构造新的结点，防止进一步的无效搜索。

影响算法的复杂度有两个因素：首先在格中的每个结点，生成一个缺省决策算法，定义结点的类的个数和属性的个数，也都是影响规则生成阶段算法复杂性的因素。可使用智能的启发式方法限制在每个结点上约简和规则的计算。动态约简[3,4](Dynamic ruduct)方法用于根据某种特殊准则选择最重要的属性子集。其次是搜索过程开始与条件属性 C 的子集，该集合的大小是关键的。对大条件属性集，可以假定搜索问题是影响在特定结点生成规则的复杂度的主要因素。属性合并[5](attribute joining)可用于限制 C 的大小，生成新的组合属性。

结论 本文提出一个求缺省规则的框架。基于缺省逻辑的缺省规则的求解是求解非单调问题的重要方法，缺省规则对训练集中的错误数据具有较少的脆弱性，因此，缺省规则比确定性规则更灵活、更紧凑。一个适当选择的缺省规则集在应用于新实例时具有更好的性能。该框架通过合并条件属性所决定的类，生成组合类，可以构造覆盖更多对象的规则，生成从这些组合类映射到占优决策的规则。结果规则比确定规则至少具有两个重要的优点：在结构上简单，因为覆盖更大的类，定义于更少的合取项；其次，即使规则相对训练集可能不完全，当处理未见的新事例时将表现更好。系统对未来对象的分类质量，将在很大程度上依据系统一般化知识的能力。

参 考 文 献

1 Piatetsky-Shapiro G, Frawley W J. Konwledge Discovery in Database. AAAI/MIT, 1991

2 Holte R C. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Machine Learning, 1993(11): 66～91

3 Bazan J G, Skowron A, Synak P. Discovery of Decision Rules from Experimental Data: [Technical Report]. University of Warsaw, Poland, 1994

4 Bazan J G, Skowron A, Synak P. Dynamic Reduce as a Tool for Extrating Laws from Decision Tables: [Technical Report]. University of Warsaw, Poland, 1994

5 Synak P. Rough Set Expert System User's Guide-Wersion1. 0. University of Warswaw. Polan, 1995

6 Wang Jue, Wang Ju. Reduction algo rithms based on discernbility matrix: the ordered attributes method. Journal of Computer Science & Technology, 2001, 16(6): 489～504

7 Ziarko W, Yao Y Y. Rough Sets and Current Trends in Computing. SCTC. Berlin: Springer-verlag, 2001

8 Kumar A. New Techniques for Data Reduction in a Database System for Knowledge Discovery Applications. Journal of Intelli gent Information Systems, 1998(10): 31～48

9 Dash M, Liu H. Feature Selection for Classification. Intelligent Data Analysis, 1997(3)

10 Kyrszkiewicz M. Mining Association Rules. PFKDD, 1998. 198～209