计算机科学 2002Vol. 29№. 9

基于属性值分类的特征规则的知识发现

The KDD of Characteristic Rules Based on Attribute Values Classification

尹世群　张为群

（西南师范大学计算机与信息科学学院　重庆 400715）

Abstract This paper presents a new classification technique of data mining and knowledge discovery in any scale relational database. Based on set theory, it divides relational table into several equal value categorys based on attribute values, calculates information capacity in decision factor of the every condition attribution, eliminates surplus attributions, and erases repeat units. Then it gets strong equal value categorys of the relational table being reduced data, gets strong characteristics on the each strong equal value category, and gathers characteristic knowledge rules. It overcomes the redundancy nature, complicated nature and unfit nature to big capacity data of some classification technique at present. It is better efficiency and widespread application perspective of the most large relational databases. The discovery algorithm and an example are discussed in details.

Keywords KDD, Data mining, Attribute values classification, Data reduction, Characteristic rules

1 引言

知识发现(Knowledge discovery in database, KDD)是应用一系列技术从大型数据库或数据仓库中提取隐含的、未知的、非平凡的对决策有潜在应用价值的知识和信息的过程，提取的知识表示为概念、规则、规律、模式等形式。知识发现过程一般由三个主要的阶段组成：数据准备，数据开采，结果表达和解释[2]。

知识发现的主要方法是数据总结、分类发现、聚类分析和关联规则的发现。其中分类是知识发现中非常重要的方法[5]。现在从统计学和机器学习的角度提出了较多的分类技术，其中以 ID3(Iterative Dichotomizer 3)系列算法(包括 ID3、C4.5、C5.0)为代表，就是将分类结果以分类树的形式给出，树的内部节点是一个决策，而叶节点代表一个类，而不是实际的规则[2]。进行分类时要查找树，而且算法产生时比较复杂，要求将分类树放入内存，进行查找。在产生的分类树转换成规则过程中，考虑了所有的因素，对主要的因素和次要的因素都要作出选择，不利于规则的生成和化简。以属性为分类节点的 ID3 为代表的一类算法的效率对于较少的数据而言是适当的，但是随着数据量和决策属性的增加，则效率会大幅度下降，而且不容易形成规则[2]。基于此，本文提出了一种新型分类技术，运用集合论把关系数据库按属性值分成若干等价类，缩减多余属性及重复元组，然后对数据缩减后的目标关系表求取分类支持度大于阈值的强类和特征置信度大于阈值的强特征，从而获取强类中的强特征的知识规则。它能够有效地克服 ID3 系列算法的冗余性和复杂性。这一分类技术的实现方法是以数据库中关系表为基础，而且在原始数据增加的情况下，可以通过缩减来压缩数据规模，使之只与属性值有关系，而与原始的数据量无关。而现在的数据存放中，几乎所有的数据都是用关系表存放，这为基于属性值分类发现特征规则提供了极大的方便，并可方便发现属性间的联系形成决策规则。

2 基于属性值分类

基于属性值分类是对数据库关系表的数据准备过程，大体分为数据集成、数据选择、数据预处理、数据缩减几个阶段。准备结果形式上可以是一个表也可以是多个表，关键是要便于进行数据开采，内容上要与数据开采的目标一致。由于数据库技术的广泛使用，使得现在大量的数据是使用表的形式存放，因此数据的准备体现在对表中的数据进行处理。数据缩减是数据准备过程的最后一个关键环节，它主要是对表中数据离散化(将一些连续的值按区间变成离散的值)和剔除重复项，减少数据量[6]。

关系表中的每一列都是某一属性的值集，列名即属性。在关系表中，依据决策规则的需要可以将属性分成两大类：一类是条件属性，称为 C 类；另一类是决策属性，称为 D 类。条件属性的值对决策属性的值有影响，甚至某一属性值的改变使得决策属性的值发生改变[1]。

在数据预处理之后，形成了一个信息系统，其中的信息是以表格的形式存放。此时的信息还不能提供决策上的支持，不能发现各个条件属性与决策属性的关系。

定义 1 信息系统 S 定义为 4 元组(U,A,V,f)的关系表，其中 $U=\{r_1,r_2,\cdots,r_n\}$是整个论域的对象的有限集；$A=\{a_1,a_2,\cdots,a_m\}$是属性的有限集合，有 m 个属性，有 $A=C\cup D$ 且 $C\cap D=\Phi$；V 为属性域的值，即 $V=\{v_{11},v_{12},\cdots,v_{nm}\}$；定义 $f:U\times A\to V$，它是形如 $f(r_i,a_j)=v_{ij}\in V$ 的全函数($a_j\in A, r_i\in U$)。

定义 2 不可区分等价关系：令 B 是属性有限集 A 的子集，$ind(B)=\{(r_i,r_j)\in U\times U \mid a\in B, f(r_i,a)=f(r_j,a), r_i\neq r_j\}$，$r_i$、$r_j$ 对属性集 B 而言具有相同属性值，则称 r_i 和 r_j 在 S 中属性 B 的集合是不可区分等价的。

定理 1 当属性有限集 A 有定义时，存在一个等价关系 $\approx$，比如，$r_i\approx r_j$ 若对任意的 k，v_k 是 r_i 的一个属性值，当且仅当 v_k 也是 r_j 的一个属性值时，则 r_i 和 r_j 对于所给定的 A 是不可区分等价的。

因 r_i 和 r_j 是论域对象，在用关系表表示的知识系统中认为是唯一标识的元组号，从而有：

$ind(D)$：对决策属性 D 的值进行分类产生的等价类$\{D_1,D_2,\cdots,D_n\}$。

$Ind(C)$：对条件属性 C 的值进行分类产生的等价类$\{C_1,C_2,\cdots,C_m\}$。

因 ind(C)中的每个元素都包含在 ind(D)中的每个元素中，故有 C 的条件，就有 D 的结果，即 C==)D。

虽然这样的产生式规则是正确的，但规则过分复杂没有任何的实际应用意义，而且考虑了所有的因素，是不合理的，这就有必要进行数据缩减。

数据缩减就是要消去规则中的不重要的属性，减少规则中的约束条件。但缩减后具有更少的条件属性的信息系统 S 具有缩减前系统 S 的功能。

数据缩减之前要先分析 S 中是否存在条件属性值相同的条件下，其决策属性可能有两个或两个以上的值，这时需进行数据离散化，使得决策属性值是比较相近的相容性数据。

然后就可以进行 S 表的缩减，缩减方法是：消除信息量小的条件属性(一般信息量小，则处于决定因素的可能性就小)，即从 S 表中消去某些列，进行条件属性的简化；再按属性值分类，把不可区分等价的重复元组删除；这样重复进行直到没有多余的属性存在。

实现 S 表缩减的具体算法 1 为：

输入：按上述要求经过数据预处理及离散化后的一个表，

名为 S_table

条件属性 C 有 $X_1, X_2, \cdots, X_k$ 属性

决策属性 D 有 $Y_1, Y_2, \cdots, Y_l$ 属性。

输出：一个缩减后的表。

算法

```
Step1  计算每个条件属性 Xi(1≤i≤k)的信息量，并将属性按信息量
       从小到大排序，记为 XP1,XP2,…,XPk，先消除信息量小的属
       性。
Step2  For i=1 to k Do //对每个条件属性 XP1,XP2,…,XPk 执行
         C_table=select condition attribute C from S_table
         P =count from C_table
           //从 S 表中，选择条件属性组成的元组，按属性值分类，
           去掉其中的重复元组，得出记录数 P
         C-XPi_table = select condition attribute C-XPi from S_
           table
         Q =count from C-XPi_table
           //在 S 表中，从条件属性中去除"侯选消去"属性，按新
           的条件属性选择出元组，并按属性值分类，去掉其中的
           重复元组，得出记录数 Q
         CD-XPi table =select attribute C+D-XPi from S_table
         W =count from CD_XPi_table
           //从 S 表中，去除"侯选消去"属性，按新的属性选择出
           元组，并按属性值分类，去掉其中的重复元组，得出记
           录数 W
Step3    If P=Q or W=Q Then
           消除属性 XPi
           C=C-XPi//从条件属性 C 中去掉 XPi 属性
           S_table=select attribute C+D from S_table //按消去属
                   性 XPi 及其值后组成新的 S 表
         Else
           保留属性 XPi
         Endif
         Next i
           //对每个条件属性处理完毕后，得到最简的决策信息表
```

例：一个学生评奖项目训练数据集，是上次评估的结果或经过评委公认典型代表，包含正反例子，作为训练数据集表，如表1所示。

表1

姓名	性别	年龄	思想品德	平均成绩	体育	获奖量化值	发表论文	加权总分	评奖等级
r1	男	21	A	98	优	90	2	91	1
r2	女	23	A	95	良	95	2	92	1
r3	男	21	A	90	优	80	1	85	2
r4	女	23	A	90	优	60	0	70	3
r5	男	23	B	90	优	60	0	70	无

条件属性 C={性别，年龄，思想品德，平均成绩，体育，获奖量化值，发表论文，加权总分}

决策属性 D={评奖等级}

执行上述算法：

把条件属性按信息量从小到大排序为：

性别，年龄，思想品德，体育，平均成绩，获奖量化值，发表论文，加权总分

首先针对"性别"属性计算

P=5　Q=5　W=5

这里 P=Q=W，所以条件属性"性别"可消去。

同样，"年龄"属性可消去。

针对"思想品德"属性，此时 S 表如下表2所示。

继续执行算法：

ind(C)={{r1},{r2},{r3},{r4},{r5}}则 P=5

ind(C-思想品德)}={{r1},{r2},{r3},{r4,r5}}

r4和 r5是不可区分等价的。去掉重复，保留一个元组，如去掉 r5，则　Q=4

ind(CD-思想品德)={{r1},{r2},{r3},{r4},{r5}}　则 W=5

那么 P≠Q 且 W≠Q

因此，"思想品德"属性保留，不能消去。

继续执行算法：

可消去属性："平均成绩"、"体育"、"获奖量化值"、"发表论文"

最后的 S 表如下表3所示。

表2

姓名	思想品德	平均成绩	体育	获奖量化值	发表论文	加权总分	评奖等级
r1	A	98	优	90	2	91	1
r2	A	95	良	95	2	92	1
r3	A	90	优	80	1	85	2
r4	A	90	优	60	0	70	3
r5	B	90	优	60	0	70	无

表3

姓名	思想品德	加权总分	评奖等级
r1	A	91	1
r2	A	92	1
r3	A	85	2
r4	A	70	3
r5	B	70	无

3　获取特征规则

获取特征规则是对基于属性值分类缩减后的目标关系表的数据开采，找出元组所具有的特征，然后以规则式来描述开采出的知识。

定义3　设 M 是一个4元组(U,A,V,f)信息系统的关系表，Sr 是总记录数，B 是 A 的子集，X 是关于 B 的一个等价类，Sx 是 X 的记录个数，则称 S=Sx/Sr * 100%是等价类 X 的分类支持度。

对于关系表 M，按照某个属性集 B 进行分类分成等价类 $B_1, B_2, \cdots\cdots$，表示为

ind(B)={$B_1, B_2, \cdots\cdots$}，可以求出每个等价类对应的分类支持度{$S_1, S_2, \cdots\cdots$}。

定义4　设 Sp 是一个给定的阈值，0≤Sp≤1，对于分类

支持度 S≥Sp 的等价类，称为强类，反之则称为弱类。

在大量的数据中，常常关心和感兴趣的是分类支持度较大的强类。强类所包含的知识更具有代表性，需对强类作进一步的特征分析，然后发现其隐含的规则。

定义5 设 X 是基于条件属性 C 的一个等价类，G 是决策属性 D 的子集，基于 G 的等价类 Y 称为 X 分类中的特征域，G 中各属性取值{g_1，g_2，……}称为 X 分类中的特征。

定义6 设 Xc 是等价类 X 中的记录数，Yc 是特征域 Y 中的记录数，则称 C＝Yc/Xc * 100%为特征置信度。

定义7 设 Cp 是一个给定的阈值，0≤Cp≤1，特征置信度 C≥Cp 的特征域，称为强特征域，反之则称为弱特征域，强特征域中属性取值称为强特征，弱特征域中属性的取值称为弱特征。

强类 X 中的强特征 y 是具有代表性的知识，表达为规则 X＝＝〉y。

获取特征规则的关键是求强类及其上的强特征。其具体实现算法2为：

```
step1  在缩减后的关系表中，按条件属性 C 进行分类
       ind(C)＝{X1，X2，…，Xn}
step2  For i＝1 to n Do //对每个等价类 Xi，Xi 若是强类就求取其强
                        特征
       Si＝Xi 的分类支持度
       If Si≥Sp Then
          求所有 Xi 分类中的特征域 Y1，Y2，…，Yk
          Cj＝Yj 的特征置信度 其中 1≤j≤k
step3        For j＝1 to k Do
              If Cj ≥ Cp Then
                 Yj 为强特征域
                 y＝取 Yj 的强特征
step4            规则 X＝＝〉y 放入结果库
             Endif
           Next j
         Endif
    Next i
```

例 对上述表3奖学金评奖，获取特征规则：

取 Sp＝20%，Cp＝80%

条件属性 C＝{思想品德，加权总分}　决策属性 D＝{评奖等级}

ind (C)＝{{r1，r2}，{r3}，{r4}，{r5}}

分别计算分类支持度 S 和特征置信度 C，有

{r1，r2}分类的特征为“1”；{r3}分类的特征为“2”；{r4}分类的特征为“3”；{r5}分类的特征为“无”。

因此规则库含有4条规则：

思想品德为 A、加权总分≥90＝＝〉评奖等级为“1”；

思想品德为 A、加权总分80～90 ＝＝〉评奖等级为“2”；

思想品德为 A、加权总分70～80 ＝＝〉评奖等级为“3”；

思想品德为 B、加权总分70～80 ＝＝〉评奖等级为“无”。

从这个人们都熟习易懂的实例可知，算法的执行和处理是简单易行的，产生的规则也是准确的。只要输入预处理后的数据关系表，就可以直接应用算法1和算法2发现知识，获得特征规则。领域决策者就可以直接应用这些规则知识进一步对同类问题进行分类和决策。

结束语 本文从集合论的概念出发，提出了基于属性值分类，进行数据缩减及获取特征规则的知识发现的过程和算法。能够有效地克服 ID3系列算法的冗余性、复杂性和对大数据量的不适应性。这一分类技术的实现方法是以数据库中关系表为基础，可以充分利用数据库中的有力工具，而且在原始数据增加的情况下，可以通过缩减来压缩数据规模，使之只与属性值有关，而与原始的数据量无关。在各种评估工作、医疗诊断、交通信息、案件信息等大型数据库的数据挖掘中有广泛的应用前景及实用价值。

参 考 文 献

1 洪家荣. 归纳学习——算法理论和应用. 北京：科学出版社，1997

2 Han Jiawei，Kamber M. Data Mining：Concepts and Techniques. Simon Fraser University，2000

3 Chen M -S，et al. Data mining：an overview from database perspective. IEEE Transactions on knowledge and data engineering，1996，8(6)：866～833

4 Pawlak Z. Rough Sets. Theoretical aspects of reasoning about data. Now ow jeska 15/19，Warsaw . Poland. 1990

5 Fayyad. From data mining to knowledge discovery：An Overview. [J] Advances in Knowledge Discovery and Data Mining，AAAI/MIT press . 1996

6 Siberchatz. What makes patterns interesting in knowledge discovery systems. [J] IEEE transactions on Knowledge and data Engineering，1996，8(6)：870～874

7 周欣，朱扬勇，施伯乐. 兴趣度——关联规则的又一个阈值. 计算机研究与发展，2000．5

8 朱定华，等. 面向属性的 RST 在数据挖掘中的应用. 华中理工大学学报，2000．2

9 孟波，等. 一种新型的有损数据简约算法 LDR. 计算机研究与发展，2000. 12

（上接第72页）

2 Dainty J C，Shaw R. Image Science. Academic Press：London，1974

3 Born M，Wolf E. Principle of Optics. 2 nd edition. Pergamon Press：New York，1980

4 Pratt W K. Digital Image Processing. 2 nd edition. John Wiley：New York，1991

5 Gonzeiez R C，Wintz P. Digital Image Processing，2nd edition. Addison-Westey：Reading，1987

6 Martin M. 利普舒茨，杨正清等译. 微分几何的理论和习题. 上海科学出版社，1989