属性集质量评估模式①

Evaluation Models for Attribute Set

陈 源 曾德胜 陈海宇 (罗定职业技术学院 电子信息系 广东 罗定 527200)

谢 冲 (深圳华为技术有限公司 广东 深圳 518000)

摘 要: 为满足用户对约简属性集的不同要求，提出了几种属性集质量的评估模式，从属性集的区分准确度、平衡度、强壮度和相似度几个方面对属性集作出客观的评估。实验结果表明：通过几种属性集质量评估模式，能从更多方面去了解属性集。

关键词: 质量评估 属性约简 数据挖掘

1 引言

目前的属性约简研究工作主要集中在如何对给定数据库进行有效的属性约简，约简出来的属性集大多数都只要求能保持与原属性集的区分能力相等就可以了[1,2]。然而，随着数据库日益庞大，属性的个数不断的增加，用户对约简属性集的质量要求不断的提高。只用属性集区分能力不变这一标准来评价一个约简属性集是不够的，我们还需要更多的考虑到属性集中属性之间的联系。

不仅在约简过程中，用户需要与计算机交互通过调整参数来得到自己想要的约简属性集，而且在判断一个约简属性集的好坏时，用户也需要从这个属性集的多个方面来进行观察，仅仅观察属性集的区分能力有时是不够的。一个属性集好差的标准应该是根据用户不同要求而定的，能符合用户需求的属性集，就是一个好的约简集。如果一个约简属性集的各项指标不符合用户的需求，那么即使这个属性集有很高的指标也没有多大的用处。因此，本文给出几种评估属性集质量的模式，并不给出好坏的标准，根据与用户交互的原则，用户可根据各个模式得出的值来自行判断约简属性集是否符合自己的要求。

2 评估属性集质量的几种模式

随着数据库中的数据量越来越大，数据中的噪音也随之增加，有些原来属性值准确的属性也会有可能出现未知和缺失的情况。对一个属性集来说出现了上述情况，它的区分能力有可能会大幅度下降。如果不管出现什么情况都能将属性集的整体区分能力保持在一个较高的水平上，这样的属性集的质量就较好。下面我们就数据库中数据量增大时，可能会出现的几种情况下，各个属性集区分能力的变化，给出几种评估属性集和质量的模式。

2.1 属性集区分准确度

目前现有的属性约简方法都非常重视约简属性集的区分能力[3-7]。我们可以通过计算出属性集的区分准确率来将它的区分能力数字化。一个属性集的区分准确率的大小也就代表了这个属性集的区分能力强弱。

当数据库中两个对象对同一个属性集中的每个属性的属性值都一样，而且两个对象的决策属性值不相等，那么这个属性集是不能将这两个对象区分开的，反之则能分开。那么我们称一个属性集能将数据库中两两对象区分开的概率就是这个属性集的区分准确度。下面给出属性集区分准确度的计算公式。

$$\text{veracity}(c,U)=\frac{\sum_{x,y\in U,\, f_D(x)\neq f_D(y)} \theta}{\sum_{x,y\in U,\, f_D(x)\neq f_D(y)} 1}$$

① 基金项目:国家自然科学基金(60463003)
收稿时间:2008-11-05

$$\theta = 1 - \prod_{i=1}^{k} (1 - tag(f_{c_i}(x), f_{c_i}(y))) \quad (1)$$

其中，k=card (c)

2.2 属性集平衡度

属性集的平衡度反映了属性集中的每个属性区分能力差别的大小。平衡值的范围是在[0,1]内，平衡值越大说明这个属性集里的每个属性的区分能力差别越小。根据不同用户对属性集的不同要求。

属性集平衡度计算如下：

$$healthy(c,U) = \sum_{m \in C} (1 - (1 - \frac{k * \text{veracity}(\{m\},U)}{\sum_{n \in C} \text{veracity}(\{n\},U)})^2) \quad (2)$$

其中，k=card (c)

2.3 属性集强壮度

随着数据库中数据量的增大，有些属性的属性值会因为各种原因出现错误、丢失的现象。许多用户都非常重视在这种情况下原有的约简属性集对这时的数据库还能保持着多大的区分能力。属性集强壮度正是反映出了属性集对出现这种情况时，属性区分能准确率下降的程度。属性集的强壮度越大，那么这个属性集出现属性值错误的时候，区分准确率下降得越小。

$$strong(c,U) = \frac{card(C)}{card(c)} \times \sum_{m \in c} \frac{\text{veracity}(c-\{m\},U)}{\text{veracity}(c,U)} \quad (3)$$

其中，c 为约简属性集，C 为原有属性集。

2.4 属性集内部相似冗余度

给定一个信息系统(数据库)(U,C,D,F)。其中 U 为对象集，即 $U=\{x_1,x_2,\cdots,x_n\}$. U 中的每个 $x_i(i=1,2,3\cdots,n)$,称为一个对象；C 为条件属性集，即 $C=\{c_1,c_2,\cdots,c_k\}$，C 中的每个 $c_j(j=1,2,3\cdots,k)$,称为一个条件属性。D 为决策属性，F 为 U 和 C、D 的关系集，$f_{Cj}(x_i)$ 为对象 x_i 的 c_j 的属性值[8,9]。

现有两个条件属性 C_m，C_n 和两个决策属性值不同的对象 x,y，函数 $tag(x,y) = \begin{cases} 1 & if \quad x \neq y \\ 0 & if \quad x = y \end{cases}$ 如果 $tag(f_{c_m}(x), f_{c_m}(y)) = 0$ 则称属性 C_m 无法区分对象，反之称其为可区分。下面我们讨论到的两两对象是针对决策属性值不同的两个对象，如果两个对象决策属性值相同，不管出现什么约简结果都不会对它们有影响所以不讨论。

定义　属性 C_m，C_n 的区分相似度为：

$$similitude(C_m, C_n) = \frac{\sum_{x,y \in U, f_D(x) \neq f_D(y)} (1-\omega)}{\sum_{x,y \in U, f_D(x) \neq f_D(y)} 1} \quad (4)$$

$$\omega = tag(tag(f_{c_m}(x), f_{c_m}(y)), tag(f_{c_n}(x), f_{c_n}(y))$$

两个属性的区分相似度越大，说明它们在对数据库中两两对象进行区分的时候同时出现不可区分或是可区分的次数就越多。当两个属性对两两对象区分时同时出现不可区分或是可区分，我们称为区分冗余。也就是说属性的区分相似度越高，区分冗余就越多。我们定义属性集的内部相似冗余度如下，就反映出了属性集的区分冗余大小。

$$redundancy(c) = \frac{\sum_{x,y \in c} Similitude(x,y)}{card(c)} \quad (5)$$

3 实验

我们将用这四种评估模式分别对基于聚类的属性约简[4]、基于差距矩阵的属性约简[8]和基于信息熵的属性约简[8]所得到的结果进行评估。实验是在 CPU 为 intel Pentium4，2.4GHZ，内存为 256M，操作系统是 windows XP 的个人计算机上进行的。我们在真实数据集上进行了实验。实验选择的数据及数据处理设置如下：

数据集来源：

http://www.ics.uci.edu/mlearn/MLSummary.html/

选择的数据集：Zoo Database（动物数据库），该数据库中的数据是分类数据。

数据集描述：该数据库中记录了 101 个动物实例以及它们的 18 个属性(animal name，15 boolean attributes，2 numeric attributes)。我们去掉了 animal name 属性。因为每个动物的名字都是唯一的，所以这个属性保留没有什么意义，类别 1-7 代表的动物分别显示在表 1 中。

表 1　数据集中动物的类别

类别号	动物个数	所属类别
Type 1	41	Mammal(哺乳动物)
Type 2	20	Aves(鸟类)
Type 3	5	Creeping animal(爬行动物)
Type 4	13	Fish(鱼类)
Type 5	4	Amphibian(两栖动物)
Type 6	8	Insect(昆虫类)
Type 7	10	Arthropod(节肢动物)

16 个属性的名称分别为：hair、feathers、eggs、milk、airborne、aquatic、predator、toothed、backbone、breathes、venomous、fins、legs、tail、domestic、catsize。我们就用属性的顺序号代替属性名称。

其中，基于聚类的属性约简算法的用户参数为：准确率为 0.99、聚类步数为 4。差别矩阵约简算法用的是二进制差别矩阵约简方法。

实验结果如下：

表 2　不同属性约简方法的各种评估结果

	聚类属性约简	差距矩阵属性约简	信息熵属性约简
约简结果	4，6，13	3，4，6，8，13	3，6，9，13，16
属性集区分准确度	0.992673	1.000000	1.000000
属性集平衡度	0.953102	0.963974	0.921145
属性集强壮度	4.707760	3.015604	3.098614
属性集相似冗余度	0.514588	0.566899	0.493261

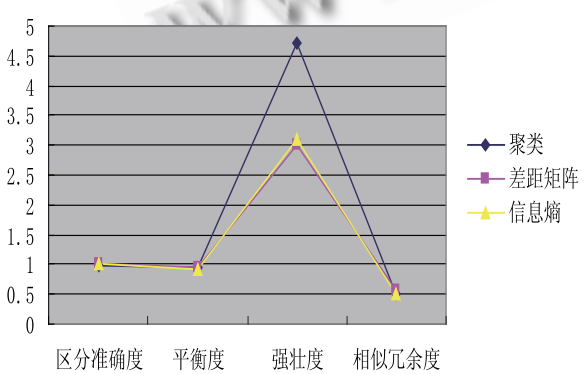

图 1　不同属性约简方法的约简结果比较

从表 2 和图 1 中，可以看到从属性集区分准确度、平衡度、强壮度、相似冗余度几个方面评估分别基于聚类、差距矩阵、信息熵的三种不同种类的属性约简方法。每种属性约简方法得到的结果都各有所长。从属性集的区分准确度来说，传统的两种约简方法得到的准确度都为 1。而根据用户给定准确率 0.99 的聚类属性约简得到的结果的区分准确率为 0.992，这一结果虽然没有前两种方法得到的区分准确率高，但是它符合用户的需要而且得到的属性个数比前两种方法得到的约简属性集里的属性个数都要小很多。从属性集平衡度上来看，差距矩阵属性约简方法得到的平衡度最高，聚类属性约简与它相差不大，信息熵属性约简得到的结果与前两种方法相比就较小了。通过对属性集强壮度的比较我们可以看出，聚类属性约简的约简属性集强壮度最大，差距矩阵与信息熵属性约简得到的约简属性集强壮度都差不多，但不过都比聚类属性约简的小很多。信息熵属性约简的相似冗余度较小，这说明通过信息熵属性约简得到的约简属性集两两属性之间的区分冗余比较少，基于聚类属性约简得到的约简属性集的相似冗余度与信息熵属性约简差距不大，而差别矩阵属性约简方法得到的约简属性集的相似冗余度相对前两种方法是比较大的。由实验结果可以看出这几种度量模式可以从不同的方面反映出属性集的特点。用户可根据自己的需求选择不同的约简属性集进行使用。

4　小结

目前对属性约简方法得到的约简属性集的评估方法都集中在属性集的区分能力的变化，其实从多个方面不同的角度对约简属性集进行评估也是一项有意义的研究工作。一个属性集的好坏是需要用户来评价的，如果能满足用户的要求就是一个好的属性集，所以我们要能让用户多方面的了解约简属性集的特点，然后根据不同用户的不同需求和属性集的评估值，选择满足用户要求的约简属性集。因此我们根据属性集的几个方面提出了几种评估方法，分别对属性集的区分能力、内部属性区分能力的差异，容错能力和内部属性两两区分时出现的区分冗余几个方面进行评估，分别得到属性集区分准确度、平衡度、强壮度、相似冗余度几种度量值。并通过实验用这几种评估模式分别对基于聚类的属性约简方法、基于差距矩阵的属性约简

(下转第 53 页)

(上接第 49 页)

方法和基于信息熵的属性约简方法得到的约简属性集进行评价。实验结果表明，几种评估方法从不同的角度反映出了几种约简方法得到的约简属性集的特点。

参考文献

1 韩家炜,Kamber M.范明,等译.数据挖掘:概念与技术.北京:机械工业出版社, 2001.

2 Miao DQ, Wang J. Analysis on Attribute Reduction Strategies of Rough Set. Journal of Computer Science and Technology, 1998,13(2):189－192.

3 陈彬,洪家荣,王亚东.最优特征矩阵选择问题.计算机学报, 1997,20(2):133－138.

4 夏文克,刘明霄,张志伟.基于属性相似度的属性约简算法.河北工业大学学报,2005,36(4):50－52.

5 孙兴波,杨平先,干树川.基于属性重要度的启发性式特征选取算法.自动化与仪器仪表, 2005,5:40－42.

6 刘靖,陈福生.结合粗糙集和模糊聚类方法的属性约简算法.计算机应用与软件, 2004,11(21):24－25.

7 李家副,李德毅.高纬聚类中的一种特征筛选方法.解放军理工大学学报(自然科学版), 2003,4(6):20－21.

8 Guan JW, Bell DA. Matrix Computation for Information Systems. Information Sciences, 2001,131:129－156.

9 Pawlak Z. Rough Sets. Int. Journal of Computer and Information Sciences, 1982,11:341－356.