容忍噪音的最大复合问题启发式算法*)

Noise-Tolerated Heuristic Algorithm for Most General Complex Problem

王兴起　孔繁胜

(浙江大学人工智能研究所　杭州310027)

Abstract In this paper, the concept, Extension Matrix Set is proposed, which is derived from Extension Matrix. A new algorithm based on Extension Matrix Set, Noise-Tolerated Heuristic Algorithm for Most General Complex (NMGC), is designed and implemented. In order to induce most general complexes, information entropy and Mexico cap function are used as attribute selection criterion and terminate function respectively. The experimental results in the real-world databases show that more general rules can be achieved; high precision can be also obtained. This implies that NMGC can be applied to real-world databases effectively.

Keywords Inductive learning, Most general complex, Extension matrix, Noise

1. 引言

从给定的数据集中发现有用的知识一直是示例学习和数据库知识发现等领域研究的重要内容[1,2]。一般地说[3]:规则越简单,归纳概括能力就越强,分类精度越高。因此,近几年来,从给定示例中归纳简单而概括的规则,即最大复合问题的算法研究逐渐成为上述诸领域的一个热点。然而,现有的规则归纳算法多为建立在不含噪音的理想数据基础上的,而在实际的应用领域中不可避免地存在噪音数据[4,5],这样致使现有的算法一直得不到令人满意的结果,甚至很难应用于实际领域,从而给实际领域规则的获取带来了一定难度。噪音数据一般可以分为如下三种形式[6],即个别属性值错误型噪音、未知属性值型噪音和冗余属性值型噪音。规则归纳算法能否有效地解决上述三种情况的噪音、是其能否成功应用于实际领域的关键。

针对现有算法的不足,本文对扩张矩阵理论进行扩充,提出了扩张矩阵集的概念,并依此设计与实现了一个容忍噪音的最大复合问题启发式算法(Noise-tolerated Most General Complex algorithm, NMGC)。NMGC 算法将加权信息熵和墨西哥帽函数分别引入到最大复合选择子属性的选择和噪音处理,从而较好地解决了噪音问题。实际领域的实验结果表明:与同类算法相比,NMGC 算法具有生成规则简单、概括能力强、分类精度高的优点,并且能够较好地应用于实际领域。

2. 基本概念

下面我们首先引用文[7]和[8]中的相关概念,然后给出本文的扩充定义。

设 E 是一个 n 维离散符号的有穷向量空间,即 $E=D_1\times D_2\times\cdots\times D_n$,其中 D_j 是有穷离散符号集,$j\in N$,$N=\{1,2,\cdots,n\}$ 为变元下标集。PE 和 NE 是 E 的子集并分别称为正例集和反例集。

定义1 选择子是形如$[x_j\#A_j]$的关系语句,其中 x_j 是第 j 个变元,$A_j\subseteq D_j$,关系 $\#\in\{=,\neq,<,\leqslant,>,\geqslant\}$。公式或复合为选择子的合取式,记为 $\wedge_{j\in J}[x_j=A_j]$,或补形集合 $\wedge_{j\in J}[x_j\neq A_j]$。

定义2 已知例子 $e=\langle v_1,\cdots,v_n\rangle$,选择子 $S=[x_j\neq A_j]$及公式 $L=\wedge_{j\in J}[x_j\neq A_j]$。$e$ 满足选择子 S 当且仅当 $x_j\notin A_j$;e 满足公式 L 当且仅当 e 满足 L 的每一个选择子,即对所有 $j\in J$,$x_j\notin A_j$,e 满足 $S(L)$也叫做 $S(L)$覆盖 e,否则,称 $S(L)$排斥 e。

定义3 已知正例 $e^+=\langle v_1^+,\cdots,v_n^+\rangle$及反例矩阵 NE。对于每个 $j\in N$,用"死元素"*对 v_j^+ 在 NE 中第 j 列的所有出现做代换,这样得到的矩阵叫做 e^+在反例集 NE 背景下的扩张矩阵,记为 $EM(e^+)$,e^+叫做该扩张矩阵的种子。

定义4 在扩张矩阵 $EM(e^+)$中,由分别来自不同行的 m 个非死元素组成的集合叫做一条路。在两个以上的扩张矩阵中,具有相同值的对应的非死元素叫做它们的公共元素。只有公共元素组成的路叫做它们的公共路,具有公共路的扩张矩阵叫做相交的,否则叫做不相交的。

定理1 扩张矩阵 $EM(e^+)$中的路同 e^+在 NE 背景下所满足的公式一一对应。

定义5 由最多的一组相交扩张矩阵所具有的公共路叫做最大公共路,由最大公共路形成的公式叫做最大复合(MGC)。

定理2 寻找最大复合是 NP-难题。

定义3中的扩张矩阵只适用于两个类别的示例学习问题,对于具有 $C(C\geqslant 3)$个类别的示例学习问题,该扩张矩阵只能够将当前讨论的类别看作正例集,把全部的其余类别看作反例集,再在此基础上进行规则归纳,这样它并没有考虑到各个类别的类内相似性和类间差异,从而导致建立在该扩张矩阵基础上的规则归纳算法得到的结果不够理想,为了较好地解决该问题,这里我们给出一个扩充定义。

定义6 已知 $e'_j=\langle v'_{j1},\cdots,v'_{jn}\rangle$为第 i 类的第 j 个示例,NE_k 为第 k 个类别的示例矩阵,$k=1,2,\cdots,C$,且 $k\neq i$,C 为论域的类别数。对于每个 $l\in N$,用"死元素"*对 v'_{jl}在每一个 NE_k 中第 l 列的所有出现做代换,这样得到的矩阵集合叫做第 i 类的第 j 个示例 e'_j 在其余类背景下的扩张矩阵集,记为 $EMS(e'_j)$,其中的每一个矩阵称为第 k 类的扩张矩阵,记为

*)本文研究得到973项目基金资助。王兴起　博士研究生,研究方向为机器学习,数据挖掘。孔繁胜　教授,博士生导师,研究方向为数据库知识发现,机器学习,智能决策支持系统等。

EM_k。

显然，当 C 取值为2时，定义5的扩张矩阵即为定义3中的形式，因此定义3的扩张矩阵是定义5的一个特例。

定理3 示例 e'_i 的扩张矩阵集中每一扩张矩阵的路同 e'_i 在其余类背景下所满足的部分公式一一对应。

证明类似于定理1(参见文[7])，这里从略。

定义7 称一个扩张矩阵为空矩阵，如果该扩张矩阵的所有行均被删除；称一个扩张矩阵集为空扩张矩阵集，如果组成该扩张矩阵集的每一个扩张矩阵均为空矩阵。

3. 最大复合问题启发式算法

我们首先给出 NMGC 算法的噪音处理方法，然后依次给出其属性选择标准和属性选择终止条件，最后给出具体的 NMGC 算法。

3.1 噪音处理方法

这一节所处理的噪音指未知属性值型噪音和冗余属性值型噪音，个别属性值错误型噪音将在属性选择终止条件中得到解决。

对于未知属性值型噪音，NMGC 算法将利用统计的方法把未知属性值型噪音根据相应属性的各已知属性值所占的比例转换成已知属性值。转换方法为：根据当前类当前属性的各个属性值所占的比例，将当前示例的未知属性值型噪音转换为所占比例最大的已知属性值，然后重新计算各已知属性值在当前类中所占的比例，当再次遇到该属性的未知属性时，则根据重新计算的已知属性值比例进行转换。这种转换方法确保了属性转换误差最小。当全部完成转换后，取某一属性值的示例个数为：

$$N'_i = N_i + N_u * ratio, \tag{1}$$

其中，N'_i 为当前类的当前属性取第 i 个属性值的总示例数，N_i 为当前类的当前属性第 i 个属性值为已知属性值的示例数，N_u 为当前类的当前属性为未知属性的示例数，$ratio_i$ 为当前类的当前属性第 i 个属性值所占的比例，该比例通过该属性的已知属性值计算。这种未知属性处理方法是基于如下考虑：训练集中的例子基本上反映了数据的变化规律，因此对于每一个例子的未知属性值型噪音可以根据已知属性值分布来计算，即对于所占比例越大的属性值，未知属性值取该属性值的几率就越大，反映到取值个数上也就越多，这种方法与人对事物的认识过程是相当类似的。

对于冗余属性值，NMGC 算法将使用全部非当前论域类的例子集中相应属性取值概率最小的属性值取代冗余属性值，例如当前种子的第 i 个属性取值为冗余属性值，那么转换后的取值为全部非当前论域类中第 i 个属性取值个数最少的属性值。

3.2 属性选择准则

由定理2可知：求解最大复合问题不存在一个通用的精确算法，这样就必须寻找一个有效的启发式算法，以求得近似最优解。因此，如何选择组成最大复合的选择子的属性便相当重要，本文把加权信息熵引入到扩张矩阵集路的选取，用以实现属性选择。

设 C 为当前论域中示例的类别数，i 为当前要生成最大复合的类别，EM_j 为第 j 类示例的扩张矩阵，$j=1,2,\cdots,C$，$j\neq i$。定义属性 A 的选择函数为：

$$E(A)=\sum_{j=1,j\neq i}^{C}\frac{R(EM_j)}{\sum_{k=1,k\neq i}^{C}R(EM_k)}\times\left(-\frac{n_j}{n_j+d_j}\log_2\frac{n_j}{n_j+d_j}-\frac{d_j}{n_j+d_j}\log_2\frac{d_j}{n_j+d_j}\right)/n_j \tag{2}$$

这里，n_j 和 d_j 分别为当前扩张矩阵集中第 j 个扩张矩阵相对于属性 A 的列中非死元素和死元素的个数，$R(EM_k)$ 为第 k 类扩张矩阵当前的行数。

很容易证明，当 n_j 和 d_j 的差异量越大时，$E(A)$ 的取值越小，这时属性 A 的归纳能力就越强。因此，在最大复合生成过程中，我们可以通过该属性选择函数对属性进行评价，选取使得该函数取值最小的属性组成选择子合取到当前的部分最大复合中。

3.3 属性选择终止条件

现有规则归纳算法的属性选择终止条件多为建立在示例集一致基础上的，即排斥全部其它类别中的示例。而在实际的应用领域中，由于噪音，特别是个别属性值错误型噪音数据的存在，导致这些规则归纳算法所得到的规则相当复杂，因为噪音数据往往导致产生冗余选择子。为了消除噪音的影响，必须寻求一个容忍噪音的属性选择终止条件。

在规则归纳的过程中我们发现：随着属性选择的进行噪音数据的影响越来越大。这一点我们可以这样理解，数据集中的非噪音数据往往呈现一定的规律性，而噪音数据却不具有这种规律性，随着属性选择的进行，非噪音数据基本上被当前的部分公式所覆盖(当前示例集中的示例)或排斥(其它示例集中的示例)，而噪音数据多数被保留下来，致使当前的例子集中包含的噪音数据比例相对增大，这样对属性选择的影响随之增大。为了减少这种噪音的影响，我们引入墨西哥帽函数来构造属性选择终止函数。

定义属性选择终止函数为：

$$M(E)<\alpha\, M(E_{avg}) \tag{3}$$

其中，$0.2\leqslant\alpha\leqslant1$，称为收敛系数，该参数用以控制算法的收敛速度，其值越大算法收敛速度越快。$M(E)=(1-x^2)\times e^{-x^2/2}$，$x=(E-\frac{E_{max}+E_{min}}{2})/(E_{max}+E_{min})$，$E_{avg}=\sum_{j=1,j\neq i}^{C}\frac{R^2(EM_j)}{\sum_{k=1,k\neq i}^{C}R(EM_k)}$。这里 E_{max} 和 E_{min} 分别为没有被当前部分公式所排斥的单个类别的最大示例数和最小非零示例数，E_{avg} 为没有被当前部分公式所排斥的所有类别的加权平均示例数，$R(EM_k)$ 为第 k 个类别扩张矩阵的行数，即没有被当前的部分公式所排斥的示例数。

该属性选择终止函数构造是基于如下一个事实：对于给定的训练示例，从聚类的角度来看，一般满足如下规律，即同一类别的数据一般具有较强的内聚性，而不同类的数据相似性则较差，表现在单个属性上则为：同一类别数据的各个属性集中取某一或几个属性值，不同类别数据间的属性取值较为分散，但是噪音数据却不具有这种规律性，从而导致其很难被当前生成的部分公式所覆盖(当前论域类而言)或排斥(非当前论域类而言)。随着规则归纳的进行，有些类别呈现两种极端的情况：一是有少量的示例很难被当前所生成的部分公式排斥，二是有大量的示例很难被当前所生成的部分公式排斥。它们分别对应于类别中存在少量噪音和被选择的种子当前所选择属性的属性值为噪音的情况，因此，我们可以根据这两种情况终止相应类别的规则归纳，从而减少噪音的影响。

图1是该属性选择终止函数图例解释，当某一类没有被排斥的示例数小于 E_0 或大于 E_1 时，它们分别对应于上述的第一种和第二种情况。在算法的实现中，我们只需计算属性选择终

止函数值即可，而无需具体求出 E_0 和 E_1 的值。

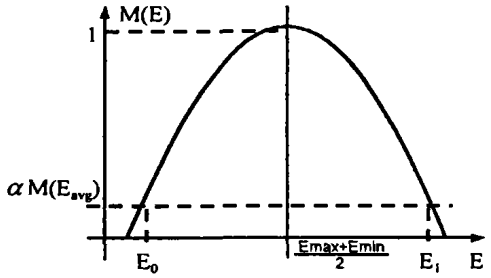

图1 属性选择终止条件图例解释示意图

3.4 启发式算法 NMGC

这里给出具体的容忍噪音的最大复合问题启发式算法 NMGC。

```
算法：容忍噪音的最大复合问题启发式算法 NMGC
输入：训练示例集 T，描述示例的属性数 n，示例类别数 C。
输出：各类的最大复合集 R_i，i=1,2,…,C。
Begin
(1)FOR 每个类别 i，i=1,2,…,C {
(2)   WHILE 第 i 类示例集 T_i≠Φ DO {
(3)      从第 i 类中选取一示例 e；
(4)      置扩张矩阵集 EMS=Φ，置第 i 类规则集 R_i=Φ；
(5)      FOR 第 j 个类别，j=1,2,…,C，j≠i
(6)         构造第 j 类的扩张矩阵 EM_j，处理未知、冗余属性，并加
            入到 EMS 中；
(7)      WHILE 扩张矩阵集 EMS≠Φ DO {
(8)         FOR 每一个没有被选择的属性 j
(9)            根据公式(2)评价属性 j，并记录评价值最小的属性，
               记为 A_min；
(10)        构造属性 A_min 的选择子，并合取到当前的部分最大复
            合 L 中；
(11)        计算未被当前部分公式排斥的类平均示例数 E_avg；
(12)        FOR 每一个非空扩张矩阵 EM_j，j=1,2,…,C，j≠i
(13)           IF M(E_j)<α M(E_avg)
(14)              置扩张矩阵 EM_j 为空；
(15)           ELSE
(16)              从第 j 类中删除能够被当前部分公式所排斥的示
例；
(17)     }
(18)     删除第 i 类扩张矩阵中被 L 覆盖的示例；
(19)     R_i=R_i∪{L}；
(20)  }
(21)  输出第 i 类最大复合集 R_i；
(22)}
End
```

NMGC 算法通过每次选取归纳能力最强的属性组成选择子来生成最大复合。其具体过程为：首先构造各类的扩张矩阵，并组成扩张矩阵集，然后根据属性选择准则选择归纳能力最强的属性组成选择子合取到当前的部分最大复合中，对噪音影响较大的类别 NMGC 算法将利用属性选择终止函数提前终止该类的最大复合的生成过程，以求减少噪音的影响。

4. 实验比较

这里给出 NMGC 算法在一些实际领域的运行结果，并与文[8]中的 FCV 算法和文[7]中的 AE5算法进行比较，实验数据取自 UCI 机器学习数据库。

从数据集中随机选取70%的示例构成训练例子集，其余示例构成测试例子集，在训练例子集上分别运行上述三种算法求得规则集，然后在测试例子集上测试所得规则的分类精度。三种算法在每一数据集上分别如上重复运行10次，取10次的平均值作为最终运行结果，由于篇幅所限，这里仅列举部分例子的运行结果于表1中。

表1 NMGC 算法实际领域运行结果

Database No.	Data description					NMGC		FCV		AE5	
	Database Name	Size	CN	AN	UF	Rules	Pre.	Rules	Pre.	Rules	Pre.
1	Annealing	898	6	39	Y	11.50	85.16	0.00	74.76	0.00	74.76
2	Audiology	226	24	71	Y	71.20	62.28	2.70	1.40	2.40	5.96
3	Letter Recognition	20000	26	17	N	2780.40	72.90	4863.20	65.71	4902.60	63.68
4	Mushroom	8124	2	23	Y	14.60	81.30	17.90	78.06	17.80	78.35
5	tic-tac-toe	958	2	10	N	160.90	81.46	172.30	80.21	179.59	78.16
6	Zoo	101	7	18	N	11.60	94.23	12.20	91.54	12.60	91.47

其中，表1中 CN 为示例集类别数，AN 为描述示例的属性数，Size 为示例集的示例总数，Rules 为算法所得到的平均规则数，Pre. 为算法所得到的规则在测试示例集上的分类精度百分比，UF 为未知属性标识，Y 表示该数据库中含有未知属性值型噪音，N 表示不含有未知属性值型噪音。

从实验结果中我们可以看出：较其它算法相比，NMGC 算法能够得到较为简单的分类规则，同时所得到的规则对测试数据的分类精度也较高，这说明 NMGC 算法所得到的规则是较优的。

另外，我们就上述三个算法在一个农业数据库上进行测试，该数据库包含有3000多个示例，描述示例的特征数为64。实验结果表明：无论是规则的简单程度还是规则的分类精度，NMGC 算法都明显优于其它两种算法。

结束语 本文提出了一个容忍噪音的最大复合问题启发式算法 NMGC，它利用信息熵和墨西哥帽函数实现最大复合选择子的生成，从而较好地解决了最大复合中属性选择和噪音的难题。在几个实际领域的实验结果表明：NMGC 算法能够得到简单、精确而概括的规则，并能够有效地应用于实际领域。

参 考 文 献

1 Srikant R. Agrawal R. Mining generalized association rules. In: Proc. of the 21st Intl. Conf. on Very Large Data Bases, Sept. 1995. 407～419

2 Shi D M, et al. Recognition rule acquisition by an advanced extension matrix algorithm. Engineering intelligent systems for electrical engineering and communications. 2000, 8(2): 97～101

3 Michalski R S, et al. Multi-purpose incremental learning system AQ15 and its testing application to three medical domains. In: Proc. of the Fifth AAAI, 1986. 1041～1045

4 Niblett T. Constructing decision trees in noisy domains. In: Mitchell T M, ed. Proc. of the 2nd European Working Session on Learning. UK: Sigma Press, 1987. 67～78

5 Norton S W, Hirsh H. Classifier learning from noisy data as probabilistic evidence combination. In: Proc. the Tenth National Conf. on Artificial Intelligence, CA, 1992. 141～146

6 Wu X D. Rule induction with extension matrices. Journal of the American Society for Information Science, 1998, 49 (5): 435～454

7 洪家荣. 示例学习及多功能学习系统 AE5. 计算机学报, 1989, 12 (2): 89～105

8 陈彬, 洪家荣. 示例学习的最大复合问题及算法. 计算机学报, 1997, 20(2): 139～144