计算机科学2000Vol. 27№. 11

一个混合特征属性选择算法*)

A Mixing Algorithm for Feature Attribute Selection

刘明吉　王秀峰　饶一梅

(南开大学计算机与系统科学系　天津300071)

Abstract　The feature attribute selection is a very interesting problem. With the development of Rough Set theory(RS)during these years, many researchers and scholars proposed the attribute selection based on RS. But with the increasement of the attribute number, the efficiency declines rapidly. In this paper, we combine the RS theory with GA and propose a mixing heuristic algorithm for attribute selection. The experiment result shows that it can get better result and higher efficiency especially for settling the problem of large attribute number.

Keywords　Rough Set, Featrue attribute selection(FAS), Data mining, Reduct, GA

1. 引言

特征属性选择(feature attribute selection, FAS)是机器学习和模式识别中比较困难而又非常有意义的一个问题[1,2,7]。FAS 问题是从一个大的侯选属性集合中选择一个较好的、有代表性的属性子集。由于在实际应用中，过多的属性会严重影响归纳学习的质量，一些不必要的属性会加大训练数据量，影响学习速度，损害所生成规则的精度[7]，因此 FAS 是一个有实际意义的问题。

目前 FAS 问题的处理方法主要有基于信息熵的特征子集选择启发式算法[7]、基于正例集 PE 和反例集 NE 的连接扩张矩阵算法[7]、Skowron 提出的基于辨识矩阵的属性约简算法[5,9]等等。近年来随着粗糙集理论的兴起，人们开始以 RS 理论的属性约简方法来解决 FAS 问题[4,7]。RS 理论是一种较新的软计算方法[6]，可以有效地分析和处理不完备信息。如何有效地求解属性集合的最优约简和属性集合的核(core)是粗糙集理论研究的中心问题之一。

然而，现已证明最优特征属性的选择问题是一个典型的 NP 完全问题[3]。特征属性选择的复杂性随着数据表中属性数目的增大呈指数增长。虽然 RS 提供了一个较好的约简搜索方法，但是目前的研究基本上是根据属性的不同排列组合以及属性的重要性来选取属性，根据约简的定义求解属性集的约简，计算复杂度随属性维数呈指数增长，效率直线下降。然而，在实际应用中一般没有必要求出所有的约简。因此，很自然的想法是寻找一个较好的启发式算法，找出其最优或次优的约简。本文给出了一种新的基于粗糙集理论的属性集优劣评判方法，并把它与遗传算法相结合，提出了一种混合的特征属性选择新算法，最后通过实例验证表明该算法在处理较多属性的 FAS 问题方面具有较好的效果。

2. 粗糙集理论简介

粗糙集把客观世界或对象世界抽象为一个信息系统，也称属性-值系统。一个信息系统 S 是一个四元组：

$$S=\langle U, A, V, f\rangle$$

其中，U 是一组对象(或事例)的有限集合，称论域；设有 n 个对象，则 U 可表示为：$U=\{X_1, X_2, \cdots, X_n\}$。A 是有限个属性的有限集合，设有 m 个属性，则其可表示为：$A=\{A_1, A_2, \cdots, A_m\}$。V 是属性的值域集，$V=\{V_1, V_2, \cdots, V_m\}$，其中 V_i 是属性 A_i 的值域；A 又可进一步划分为两个不相交的集合：条件属性集 C 和决策属性集 D，C 和 D 满足 $A=C\cup D$ 且 $C\cap D=\varnothing$，D 一般只有一个属性。函数 $f: U\times C\cup D\rightarrow V$，定义对象的属性。

1)**下近似和上近似**：对任意一个概念(或集合)X，R 是 A 的一个子集，X 的下近似定义为：

$$R_-(X)=\{x\mid x\in U: [x]_R\subseteq X\}$$

*)本论文的研究工作得到国家自然科学基金项目(79790130)和天津市自然科学基金项目(993600811)资助。刘明吉　博士生，研究领域为数据仓库、数据挖掘以及金融信息系统；王秀峰　博导，研究领域为遗传算法、计算智能技术；饶一梅　博士生，研究领域为金融信息系统。

其中$[x]_R$表示x在R(X)上的等价类。X的上近似定义为：

$$R^-(X)=\{x|x\in U:[x]_R\cap X\neq\emptyset\}$$

我们也把$POS_R(X)=R_-(X)$称为X的R正域，把$Neg_R(X)=U-R_-(X)$称为X的R负域。

2)**约简和核**(reduct and core)：在S中，各个条件属性之间往往存在着某些程度上的依赖或关联。若在条件属性集合A中存在$A'\subset A$，由A'中的属性可以确定属性D的取值，这种关系称为约简关系。约简可以定义为不含多余属性并能保证分类正确的最小属性集。具体地说：

对于一个给定的信息系统S，A的约简是A的一个非空子集A'，并且满足

(1)$IND(A',D)=IND(A,D)$

(2)对任意$A''\subset A'$，$IND(A'',D)\neq IND(A,D)$

IND表示不可分辨关系。A的约简记为RED(A)

核：全体约简的交集称为A相对于D的核(CORE)，$CORE(A)=\cap RED(A)$

3)**属性依赖度**：设有两个属性集P和Q，则P对Q的依赖度定义为：$\gamma_P(Q)=\frac{|POS_P(Q)|}{|U|}$，其中$POS_P(Q)=\bigcup_{X\in U|Q}P_-X$，$P_-X$表示集合$X$在属性集上的下近似，U/Q表示论域U根据属性集Q所得到的划分类。用‖表示集合中的元素个数。因此$\gamma_P(Q)$表示了根据P能被准确分类的对象在系统中所占的比例，或称为属性集P区分划分类(U/Q)的能力。

4)**属性重要度**：属性的重要度就是该属性会对于决策属性或数据分类问题的影响程度。属性集度量的方法很多，如信息熵增益、Gini、χ^2等方法，它们都基于广义差别矩阵，因而计算量较大。基于粗糙集理论，为了评测某些属性或属性集的重要性，需要从属性集中去掉某些属性，再来考察没有该属性后分类会怎样变化。若去掉该属性改变相应的分类，则说明该属性的强度大，即重要性高；反之说明该属性强度小，即重要性低。这里我们用属性重要度来描述属性的重要性。

属性集评价函数：设G(X)为属性集度量函数(如上述各种方法)，对于任意一个属性集X，令$G(X)=\gamma_X(D)$，则$SIG(X)=G(A)-G(A-X)$。在本文中，我们以粗糙集理论的属性依赖度作为属性度量函数，则$SIG(B)=\gamma_A(D)-\gamma_{A-B}(D)$，$SIG(B)$表示了在A中由于缺少属性集B而导致不能被准确分类的对象在整个系统中所占的比例。容易证明如下的性质：

①$SIG(B)\in[0,1]$；

②若$SIG(B)=0$，则表示属性B对于D来说是可以省的。

基于属性依赖度的属性集评价函数给出了一个评价属性集重要性的一个标准，也给我们处理FAS问题提供了一个新的思路，众所周知，导致FAS是一个NP-难问题的主要原因是属性的组合爆炸，根据RS理论约简的定义进行的约简求解方法，需要多次访问全部数据，需要多次计算属性组合的重要性，其计算效率将随着属性维数的增加而呈指数下降。如果我们把属性集评价函数作为一种启发算子，使它与一般的搜索算法(如遗传算法)结合起来，就可以逐步缩减搜索空间，搜索效率一定会大大提高。

3. 基于遗传操作和粗糙集理论的属性选择方法

遗传算法(Genetic Algorithm，GA)是一种集效率与效果于一身的优化搜索方法。它利用结构化的随机信息交换技术组合群体中各个结构中最好的生存因素，从而复制出最佳代码串，并使之一代一代地进化，最终获得满意的优化结果。在这里，我们把特征属性的选择问题转化成为属性组合的寻优问题，首先对属性空间进行遗传编码，以属性编码构成染色体，用基于RS理论的评价函数来启发并指导进化，逐步得到较优的能基本体现属性空间分类特性(达到用户满意度)的属性组合代码串，从而发现描述该属性空间的特征属性集。下面我们将详细讨论基于遗传操作和粗糙集理论的属性选择算法。

3.1 遗传编码

本算法采用传统的二进制遗传编码方式。染色体的长度取决于属性的个数，每个基因位对应于每个属性，表示其选择与否。其中"1"表示选中，"0"表示没有选中。

例如：染色体串1001010001表示由10个属性构成的染色体串，并在10个属性中选中了第1、4、6、10个属性。

3.2 遗传操作

·交叉算子　交叉算子在遗传算法中非常重要。首先按照一定的方法，随机地从交配池中取出要交配的一对染色体，然后进行交叉，产生一对新的位串。交叉算子可以分为单点交叉和多点交叉，本文采用两点交叉。交叉的方法是先根据位串长度L，随机产生两个交叉位置，即[1，L-1]上的两个整数，然后进行交叉，例如：

染色体1　10100△01010△10010　交叉后代　10100△00010△10010

染色体2　10101△00010△10100　交叉后代　10101△01010△10100

·变异算子　变异算子根据一定的变异率P_m，在染色体上随机选择一个基因，然后改变该基因的特征。变异不仅可以保证引入有用的遗传物质，保持种群的差异性，而且能适当地提高GA的搜索效率。在本算法

中，就利用变异算子来随机改变属性的选择与否，即"0"变为"1"或"1"变为"0"，从而产生新的特征属性选择。

3.3 评价函数

评价函数定义为：

$$eval(X)=\begin{cases} a\times A(X)+b\times SIG(X) & (SIG(X)\geq l) \\ 0 & (SIG(X)<l) \end{cases}$$

l 表示可信度阈值

其中，SIG(X)表示 X 中含有属性集的重要程度。A(X)表示用户设定的属性个数与满意度的一个对应函数。我们在评价函数中引入 A(X)的目的是为了加入用户的领域知识，从而使特征属性集的选择更加符合实际情况。因为在某一领域的实际应用中，用户可能会对特征属性的选择有一定的经验知识。例如，用户凭借领域经验所获得的约简比例或属性个数与效果的一个期望函数。图1列举了两个 A(X)的函数示意图。这两个函数都表示用户认为约简后属性个数为 m 时最为满意。参数 a，b 是两个权重比例调节参数，用以表示用户是关心属性的缩减还是关心分类的准确度。

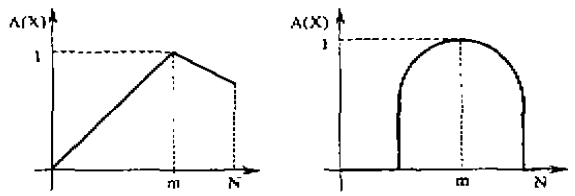

图1 两个 A(X)函数的示意图

3.4 算法流程

本算法采用简单遗传算法 SGA，编码十分简单，关键在于属性集评价函数的设计。其主要算法流程如下：

①确定遗传算法的有关参数值，比如种群大小、迭代次数、停机条件；

②初始化种群，随机生成表示属性选择与否的染色体；

③根据评价函数的定义，计算各染色体的适合度函数值：

$eval(x_j^t)\quad j=1,\cdots p\_size$

④根据各染色体适应值的比例信息

$$p_j^t=eval(x_j^t)/\sum_{i=0}^{p\_size} eval(x_i^t)$$

由轮盘赌的方式产生下一代染色体 $U^t=\{x^t\}$，其中 t 表示代数。

显然，适应值较大的个体参与产生下一代的概率较大，为了加快算法的收敛速度，可以将每代中适应值较大的一些个体强行传到下一代，而不受选择过程的限制；

⑤交叉。在 U^t 中以一定的概率 P_c 随机选择个体 x_i、x_j 进行两点交叉操作；

⑥变异。从交叉操作后得到的个体中，以一定的概率 P_m 随机选取某个个体，按照前面的变异算子定义，进行变异操作。至此我们得到了经过遗传操作后获得的下一代种群 $U^{t+1}=\{x^{t+1}\}$；

⑦如果满足停机条件，则退出；否则转向③；

在本算法中，选择算子体现了适者生存的原则；交叉算子组合父代种群中有价值的信息，产生新的后代，具有遗传功能；变异算子的作用是保持群体中基因的多样性。

编号	数据库名称	属性个数	实例个数	特征属性个数		时间(平均)	
				RS	GA	RS	GA
1	Ballon Database(1)	4	20	2	2	20多秒	瞬间
2	BUPA Liver Disorders	6	200	3	3	2分钟左右	3分钟左右
3	Voting database	16	435	10	9	12分钟	10分钟
4	Solar flare database	10	1066	8	8	半小时左右	10多分钟
5	German Credit Data	24	999	7	6	50分钟	30分钟
6	Mushroom database	22	5128	5	5	1小时左右	40分钟
7	Lung cancer	98	252	43	41	2小时左右	1小时
8	Thyroid disease	78	2128	37	39	2个半小时	1小时稍多

4. 实例验证

实验的目的主要是为了验证本算法的有效性，并从运行效率，约简率等方面来验证本算法是否优于其它的特征属性选择算法。我们从 UCI 机器学习数据集中挑选几个有典型意义的数据集，并选择 RS 做为参照来验证算法的优良。另外，我们为了科学地验证本算法的有效性，防止由于先验知识所带来的干扰，设评价函数中 $A(X)=0$。

在实验结果中，我们发现两种算法所获得的特征属性个数差不多，可见 GA 的特征属性选择能力并不明显强于 RS 的属性约简能力，这主要是由于我们采用的属性集评价标准一致，也没有加入用户的经验知识。然而由于其内在的并行性，其效率明显优于 RS，特别是对于属性数目较多、实例个数较多的 FAS 问题，GA 的运算效率明显优于单纯的 RS 算法。

结束语 本文以粗糙集理论为基础，给出了一个基于属性依赖程度的属性重要度标准，并结合遗传算法，提出了一个特征属性选择算法。该算法具有编码简单，运行效率高的特点，特别对于解决巨维数据的特征属性选择问题取得了一定的效果，克服了粗糙集约简方法计算量大、效率低的不足。本算法不仅可以应用于人工智能、模式识别，而且在数据库知识发现的数据预处理中也有广泛的应用。

但是，本算法受粗集属性重要度的影响，只能处理离散属性。如何基于粗糙集理论构造能够评测连续属性的测度标准，还留待以后的工作进一步研究。另外，如何运用并行遗传算法来进一步提高运算效率，也是特征属性选择的一个可以进一步研究的问题。

参 考 文 献

1 Hu X H, Cercone N. Learning in relation database: a Rough Set approach. Computational Intelligence, 1995, 11(2): 323～338

2 Jelonek J, et al. Rough Set reduction of attributes and their domains for neural networks. Computational Intelligence, 1995, 11(2): 339～347

3 韩祯祥，张琦，等. 粗糙集理论及其应用综述. 控制理论与应用，1999，16(2)：153～157

4 王珏，王任，等. 基于 Rough Set 理论的"数据浓缩". 计算机学报，1998，21(5)：393～399

5 王珏，苗夺谦，等. 关于 Rough Set 理论与应用的综述. 模式识别与人工智能，1996，9(4)：337～344

6 曾黄麟. 粗集理论及其应用. 重庆：重庆大学出版社，1998

7 王志海，胡可云，等. 基于粗糙集合理论的知识发现综述 模式识别和人工智能，1998，11(2)：346～351

8 刘明吉，王秀峰. 数据挖掘中的数据预处理. 计算机科学（已录用）

9 赛英，陈文伟 从数据库中发现知识的方法研究与应用. 管理科学学报，1999，2(3)：92～96

10 常犁云，王国胤，等. 一种基于 Rough Set 理论的属性约简及规则提取方法. 软件学报，1999，10(11)：1206～1211

（上接第81页）

参 考 文 献

1 Chuchro M. On Rough Sets in Topological Boolean Algebras. In: Rough sets, Fuzzy Sets and Knowledge Discovery. Springer-Verlag, 1994

2 Rosiowa H, Sikorski R. The Mathematics of the Metamathematics. PWN, Warsaw, 1968

3 Bonikowski Z. A Certain Conception of Calculus of Rough Sets. Notre Dame Journal of Formal Logic, 1992, 33(3)

4 Pawlak Z. Rough Sets: An Algebraic and Topological Approach: [ICS PAS Report] Pol. Ac. Sc., Warsaw, 1982

5 Pawlak Z. Rough Sets. Theoretical Aspects of Reasoning About Data. Kluwer, Dordrect, 1992

6 Pawlak Z, et al. Rough Sets, Comm. ACM, 1995, 38(11)

7 何华灿，等. 经验性思维中的泛逻辑. 北京：中国科学（E辑），1996，1

8 何华灿，等. 人工智能导论. 西安：西北工业大学出版社，1988

9 祝峰，何华灿. 粗集的公理化. 计算机学报（已录用）

10 祝峰，何华灿 Logical Properties of Rough Sets, HPC-Asia2000（已录用），Beijing, May, 2000

11 祝峰，何华灿，等. 粗集中粗元的结构及其拓广. 计算机科学，2000，27(6)