基于粗糙集理论和覆盖算法的模式分类方法*)

王伦文[1,2] 张 铃[2]
(解放军电子工程学院204研究室 合肥230037)[1]
(国家教育部安徽大学人工智能与信号处理重点实验室 合肥230039)[2]

A Method of Pattern Classification Based on Rough Set and Neighborhood Covering Algorithm

WANG Lun-Wen[1,2] ZHANG Ling[2]
(204 Research Room of Electronic Engineering Institute of PLA, Hefei 230037)[1]
(Key Laboratory of Artificial Intelligence & Signal Processing of Education Ministry of Anhui University, Hefei 230039)[2]

Abstract In this paper, the advantages and disadvantages of Rough Set and Neighborhood Covering Algorithm (NCA) are analyzed. A classification method which combines Rough Set and NCA is pointed out. And then, the reasonableness, feasibility and necessity of the method are demonstrated. The recognition of wireless communication signals is given as an example to show that the method can be used practically.

Keywords Rough set, Neighborhood covering algorithm, Pattern classification

1. 引言

模式分类是模式识别和人工智能研究最基本也是最重要的课题之一。现实世界事物纷繁复杂,尤其是海量数据库、互联网出现,这些信息的处理加工,对分类的要求更加迫切。

事物可形式化为三元组{U,C,V},U是论域集合,C是事物特征集合,V是特征值集合。u∈U是事物的标识,c∈C是事物特定的属性,v∈V是该属性的特定值。对于|U|=N,具有n类的事物进行分类,其分类器一般可用图1表示,假定$y_i\neq\phi$,(i=1,2,…,n),理想的分类器应满足条件:

$$y_i\cap y_j=\phi,(i,j=1,2,\cdots,n\ i\neq j) \quad (1)$$

$$y_1\cup y_2\cup\cdots\cup y_n=U \quad (2)$$

(1)要求分类器不误分,(2)说明U中没有一个拒分。

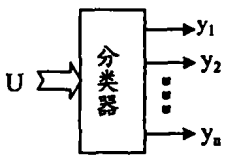

图1 分类器示意图

然而事物的复杂性主要表现在C和V上。对C而言,不同的事物有不同的特征,同一事物针对不同方面,取不同特征;V表现为数值类型不一,分为数值型和非数值型[1]。数值型包括连续的、离散的和模糊的;非数值型包括字符型、逻辑型、日期型、时间型、枚举型等。国内外科学家研究的许多成果,大多针对某一类或某些类分类很成功,用于其他方面则逊色。

近年来利用神经元网络和粗糙集理论构造分类器备受关注,本文尝试将二者相结合。

2. 基于M-P覆盖算法构造的分类器的分析

如图2,神经元M-P模型,是n输入,单个输出的处理单元。其输入输出关系如下:

$$y=\mathrm{sgn}(w^Tx-\theta)$$

其中$x=(x_1,x_2,\cdots,x_n)^T$为输入向量,$w=(w_1,w_2,\cdots,w_n)^T$为权向量,θ为阈值$\mathrm{sgn}(v)=\begin{cases}1 & v\geqslant 0\\ -1 & v<0\end{cases}$,所以M-P模型判决边界表达式为:

$$w^Tx-\theta=0 \quad (3)$$

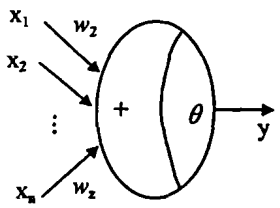

图2 神经元M-P模型

从几何意义上讲,式(3)对应于n维空间的一个超平面,权向量W就是该超平面的法向量,当为单位向量时,θ的绝对值为坐标原点到该超平面的距离。这就是M-P模型的几何意义。式(3)对应n维空间中的一个超平面,它将n维空间分为两个部分,当$w^Tx-\theta\geqslant 0$时,输入向量x落在正的半空间内,此时M-P模型输出为1。而$w^Tx-\theta<0$时,输入向量x落在负的半空间内,M-P模型输出为-1。因此,从几何角度看一个M-P模型可将模式空间分为两类,分别对应于超平面$w^Tx-\theta=0$分成的两个半空间,要进行更复杂的分类,只需引入更多参数不同的神经元,即用多个不同的超平面对样本空间进行划分即可。

以上这种超平面分类,虽然用到了几何意义,但对多维向量的分类还是不够直观。张铃教授巧妙地将n维向量投影到$n+1$维球面上,分类问题变得很直观,易于理解,引起了广泛关注。具体地说,就是将n维向量作变换$T:D\to S^*,x\in D$,即:

$$T(x)=(x,\sqrt{(d^2-|x|^2)}) \quad (4)$$

*)本研究得到国家自然科学基金(60175018)、(60135010)和国家973项目(G1998030509)的资助。王伦文 博士研究生,主要从事信号处理、人工智能等领域研究。张 铃 教授,博士生导师,主要从事人工智能等领域研究。

其中，$d \geqslant \max\{|x|, x \in D\}$，这里的 D 为 n 维空间中的有界集合，S^n 是 $n+1$ 维空间中的超球面。这一变换将 D 一一映射到一个半径为 d 的超球面上。

那么 $(W * x - \theta) > 0$，则表示球面上落在由超平面 p（其方程为式(3)）所分割的正半空间的部分，这部分恰好是球面上的某个球形领域。这样，就将神经元与球面的球形领域联系起来，利用神经元的这种几何意义能非常直观地进行神经网络的各种研究。

由上面给出的神经元的几何意义得知，构造一个网络，对给定的样本集能进行符合要求的分类，等价于求出一组领域，对给定样本集 K 中的点，能按分类的要求用领域覆盖将它们分隔开来。这样，就将神经网络的最优设计问题转化成某种求最优覆盖的问题[2,3]。

这种算法的优点是几何意义明显，它从给定的数据出发，逐步构造出所需的覆盖领域来，对训练样本100%识别，能实现对多类别、大规模的模式分类，它成功地实现了平面双螺旋分类[3]。这种算法的另一个突出的优点是将样本原原本本地投影到超球面上，分类精度达到最高，区别于其他分类方法随分析粒度变化而改变精度。但是也存在两个缺陷：一是只能解决特征为数值型和能转换成数值型的非数值型问题。因为式(4)中的变量 x 是数值型的 n 维向量。可是，大多数问题既有数值型特征，又有不可转换为数值型的非数值型特征，限制了算法的应用范围。二是处理速度有待提高，特别是大规模的模式分类。算法的复杂度为：

$$O((K * S * W)^2) \qquad (5)$$

其中，K 是类别数目，S 是每类中的训练样本数目，W 是输入向量的维数。大规模的模式分类中，K、S 和 W 通常比较大，因此运算量较大[4]。

模式分类的一种特殊形式是分层分类，它也包含于式(1)和(2)中。即可以将式(1)和(2)视为第一层分类，它组成子集 $U_k(i=1,2,\cdots,n)$。在第二层分类，每个子集 U_k 本身继续按式(1)和(2)分类为子集 $U_{kj}(j=1,2,\cdots,m)$。这一过程可以根据需要延续下去。另外，对于集合 U 来说，存在许多分解均满足式(1)和(2)，但对于使用者来说，仅有少数，而在许多情况下甚至仅有一种是实用上感兴趣的。在此，只需考虑将符合使用者的愿望分解出来，不考虑其他的。

目前，粗糙集理论应用比较广泛，该理论不仅适合处理非数值型数据，而且能根据非数值型属性约简属性，消除冗余信息。如果我们将粗糙集理论与覆盖算法结合起来，采用粗糙集理论针对非数值型特征进行第一层分类，利用覆盖算法针对数值型特征进行第二层分类，能较好地解决上述问题。因为第一层分类不仅降低输入数据维数 W，而且只将我们感兴趣的子集分离出来，不考虑无关的子集，从而减少后续分类的样本数目 S，根据式(5)将减少后续算法的复杂度；第二层分类利用覆盖算法能提高分类精度。

3. 基于粗糙集理论的知识简化和粗分类

粗糙集理论[5]是由 Pawlak 于80年代初提出，是一种处理模糊和不确定性知识的数学工具。其主要思想是在保持分类能力不变的前提下，通过知识约简，导出问题的决策或分类规则。

知识约简[6,7]是粗糙集理论的核心内容之一。我们在对一个事物做出判断和决策时，并不是依据被判事物的全部特征，而是依据最主要的一个或几个特征。知识约简就是根据这一原理，剔除知识库中的冗余知识，简化规则。

定义1 信息系统 $S=(U,A,\{V_a\},a)$ 是一个四元组，其中 U 是非空有限集合，称为论域；A 是非空有限集合，称为属性集合；V_a 是属性 $a \in A$ 的值域；$a: \rightarrow V_a$ 为一单射，使论域 U 中的任一元素取属性 a 在 V_a 中的某一唯一值。若 A 由条件属性集合 C 和结论属性集合 D 组成，C 和 D 满足 $C \cup D = A$，$C \cap D = \phi$，则称 S 为决策系统，常用 $(U, C \cup D)$ 表示；当结论属性集合只有一个元素时，也常用 $(U, C \cup \{d\})$ 表示。

定义2 若 P,Q 是 U 上的两个等价关系的集合，设 P 和 Q 是全域 U 上的等价关系的族集，称 $POS_P(Q) = \bigcup_{X \in U|Q} P_*(X)$ 为 Q 的 P-正区域，记作 $POS_P(Q)$，它是全域 U 上所有使用分类 U|P 所表达的知识能够正确地分类于 U|Q 的等价类的个体的集合。若 $r \in P$，且 $POS_{IND(P-\{r\})}(IND(Q)) = POS_{IND(P)}(IND(Q))$，则称关系 r 在族集 P 中是 Q-可约去的，否则称为 Q-不可约去的，如果在族集 P 中的每个关系 r 都是 Q-不可缺的，则称 P 关于 Q 是独立的，否则就称为 P 关于 Q 是相关的。若 $S \subset P$，且 $POS_S(Q) = POS_P(Q)$，称 S 为 P 的 Q-约简，族集 P 中的所有 Q-约简的交，称为族集 P 的 Q-核，记作 $CORE_Q(P)$，它是 P 中所有 Q-不可约去的关系的集合。

根据定义2，可以对属性集进行约简，得到最小属性集 CORE(C)。

粗糙集分类是依据信息系统的等价关系或不可分辨关系对研究的论域 U 进行分类。事实上信息系统的一个属性就是 U 上的一个等价关系，这里的分类就是根据属性进行分类。

定义3 对决策系统 $(U, C \cup \{d\})$，$B \subseteq C$ 是条件属性集合的子集，称二元关系 $IND(B,\{d\}) = \{(x,y) \in U \times U | d(x) = d(y)$ 或者 $a \in B, a(x) = a(y)\}$ 为不可分辨关系，其中，x,y 为 U 元素。称 x,y 为属性 B 的等价关系，也称 x,y 关于属性 B 为同一类。

结合定义3，我们可以在最小属性子集基础上，将论域 U 用等价关系进行分类。

由上可知，利用粗糙集理论可以对信息系统进行属性约减和粗分类。属性约减在不丢失信息的前提下，消除冗余的属性；粗分类是针对非数值型的属性，从论域 U 中挑选出各属性值满足我们要求的子集，供覆盖算法细分。粗分类后输入数据不仅样本数减少，而且维数下降，后续的覆盖算法运算复杂性将大大降低。

4. 粗糙集与覆盖算法相结合的分类方法

粗糙集理论与覆盖算法在模式分类中各有优势。粗糙集方法能简化规则，消除冗余信息；覆盖算法能提高分类精度。在解决实际问题中，根据决策属性选择解决方法。设决策属性为 D，条件属性 $C = C_1 \cup C_2$，其中 C_1 是非数值型条件属性，C_2 是数值型的。有三种方法供选。

方法1：若 $D = f(C_1)$，即：D 只由 C_1 决定，可用粗糙集方法[7]；

方法2：若 $D = f(C_2)$，即：D 由 C_2 唯一决定，可用覆盖算法[8]；

方法3：若 $D = f(C_1 \cup C_2)$，即：D 由 C_1 和 C_2 共同决定，则采用我们研究的二者相结合的方法。

例如，一信息表是一个班学生的基本情况表，若 D＝性别，则用方法1；若 D＝全体学生平均成绩，则用方法2；若 D＝男女学生的平均成绩，则用方法3。

方法3具体实现步骤如下：

(1)非数值型条件属性的约简

①求出决策属性 D 的非数值型条件属性 C 的正区域 $POS_c(D)$；

②对每一个非数值型条件属性 c_i，计算 $POS_{C-(c_i)}(D)$，若 $POS_{C-(c_i)}(D)=POS_C(D)$，则 c_i 是可约去的，否则，c_i 不可约去，从而得非数值型条件属性的核 CORE(C)；

③求含 CORE(C)的最小集合 P，使得 $=POS_P(D)=POS_C(D)$，得 C 的最小约简 P。

(2)删除样本中不包含于 P 中的属性所在的列，得简化的学习样本。

(3)对数据粗分类，根据需求筛选出满足非数值型条件属性的子集。令 $P_i\in P,(i=1,2,\cdots,m)$，$P_i$ 的值域为 $\{V_{Pi}\}$，其中 $g_{Pi}\in\{V_{Pi}\}$ 满足要求，可按照 $\bigcap(P_i=g_{Pi}),(i=1,2,\cdots,m)$ 求出子集。具体方法为：

① $Class\leftarrow\varnothing$，$Temp\leftarrow U$，$i=1$；

② 取 $Temp$ 中的第 i 个元素记为 $x(i)$，其中 $i=1,2,\cdots,n$，n 为论域对象的个数；

③if$((P_1=g_{P1})\&(P_2=g_{P2})\&\cdots\&(P_m=g_{Pm}))$，$Temp\leftarrow Temp-x(i)$；$Class\leftarrow Class\cup x(i)$；

④ else $Temp\leftarrow Temp-x(i)$；

⑤ 若 $Temp=\varnothing$转⑥，否则，$i=i+1$，转②；

⑥ 输出 U$\leftarrow Class$。即：满足特定非数值型条件属性子集仍放回 U 中。

(4)覆盖算法分类的具体实现，请参考文[8]。

由上可知，方法3的步骤(1)和(2)降低了输入数据的维数，步骤(3)减少后续覆盖算法的样本个数，步骤(4)具有较高的分辨能力。因此方法3既快速又准确。

5. 实验及结果

我们用某型号接收机接收并采集通信信号，提取其特征向量，如表1所示，信号1、信号2和信号3的特征是从采集广播电台的数据中提取的，信号4和信号5的是自己用某型号电台发射信号，经接收和采集而提取的。表1、2中，U 为论域，即信号个体，X_1为调制样式，X_2为信号属性，X_3为频率(KHz)，X_4为强度，X_5为带宽(KHz)，X_6为调制参数，X_7为语种，Y 为决策属性，即结果。

表1 通信信号的部分特征表[9]

U	X_1	X_2	X_3	X_4	X_5	X_6	X_7	…	Y
1	AM	民	864	61	3.31	0.17	汉	…	a
2	AM	民	936	84	3.37	0.16	汉	…	b
3	AM	未	15500	96	3.51	0.19	汉	…	c
4	FSK	军	10500	43	1.42	1.2	汉	…	d
5	FSK	军	12300	51	1.16	1.0	汉	…	e

$C=\{X_1,X_2,\cdots,X_n\}$为条件属性集，$\bigcap(C-X_7)=\bigcap C$，所以 X_7为 C 中可省略的，同理约其他冗余属性，得到约简后的属性集 $S_{core}=\{X_1,X_2,\cdots,X_6\}$，如表2所示。

表2 通信信号的部分特征的核心属性表

U	X_1	X_2	X_3	X_4	X_5	X_6	Y
1	AM	民用	864	61	3.31	0.17	a
2	AM	民用	936	84	3.37	0.16	b
3	AM	未知	15500	96	3.51	0.19	c
4	FSK	军用	10500	43	1.42	1.2	d
5	FSK	军用	12300	51	1.16	1.0	e

表2是完备属性集，因为$\bigcap(C-\{X_i\})\neq\bigcap C$，即 X_i 是不可省略的$(i=1,2,\cdots,6)$。表2与表1相比，信号的属性个数减少了，即：输入数据维数降低了。至此，我们可以根据需求，利用非条件属性进行粗分类。若需求为：$X_1=$“AM”，$X_2=$“民用”，经过粗分类，只有信号1和信号2满足要求，接着对这两个信号用覆盖算法细分，其他的三个信号就不再作细分的考虑。因此减少了进行覆盖计算的样本数。这对大规模模式分类具有重要的意义。

我们在不同时间分别大量接收并采集表1中的所有信号，提取其特征向量，用粗糙集方法(方法1)、覆盖算法(方法2)和粗糙集与覆盖算法相结合的方法(方法3)实验比较，结果如表3所示。

表3 通信信号识别结果表

识别方法	训练样本数	测试样本数	识别时间(秒)	识别正确率
方法1	200	500	1.8	85.2%
方法2	200	500	2.1	96.4%
方法3	200	500	1.4	99.2%

通过表3，可以发现粗糙集和覆盖算法相结合的方法比单独使用粗糙集或覆盖算法，在识别时间上和识别正确率上都有大幅度的提高，说明这种方法是有效的。

结束语 将粗糙集方法和覆盖算法结合起来，能较好地利用粗糙集理论约减非数值型属性的冗余信息，提取非数值型属性的最小核心属性集，有效降低属性维数，并根据非数值型属性进行粗分类，将符合使用者意图的部分子集先分离出来，不考虑剩余子集，从而减少后续覆盖算法的样本个数。同时又能发挥覆盖算法的优势，对数值型特征构造领域覆盖的几何模型，精确地对模式进行分类，很好地实现模式分类的快速和准确的统一。该方法在通信信号识别中已经得到证实，对于其他领域模式识别和机器学习，也有很好的借鉴作用。

参考文献

1 [联邦德国]H. 尼曼著，周冠雄，李梅译. 模式分类. 科学出版社，1988

2 Zhang Ling, Zhang Bo. A Geometrical Representation of McCulloch-Pitts Neural Model and Its Applications. IEEE Transactions on Neural Networks, 1999, 10(4): 925～929

3 张铃，张钹. M-P 神经元模型的几何意义及其应用. 软件学报，1998, 9(5): 334～338

4 吴鸣锐，张钹. 一种用于大规模模式识别的神经网络算法. 软件学报，2001, 12(6): 851～855

5 Pawlak Z. Rough Sets Theoretical Aspects of Reasoning about Data. Warsaw: Nowowiejska, 1990

6 Stepaniuk J. Attribute Discovery and Rough Sets. In: Komorowski J, Zythow J, eds. Principles of Data Mining and Knowledge Discovery, 1997. 145～1551

7 王国胤. Rough 集理论与知识获取. 西安交通大学出版社，2001. 5

8 王伦文，等. 一种改进的覆盖算法及其应用. 模式识别与人工智能，2003(1)

9 张贤达，保铮. 通信信号处理. 国防工业出版社，2000