基于粒度计算的覆盖算法*)

赵 姝 张燕平 张 铃
(安徽大学计算智能与信号处理教育部重点实验室 合肥 230039)

摘 要 为了更好地解决高维海量数据的分类问题,本文提出一种基于粒度计算的覆盖算法。该算法以粒度计算为理论依据,指出在分析研究某一问题时,可以适当将其属性、论域或者结构粗化,求得某个商空间,在该商空间中抓住事物的本质对其研究,对某些在同一个粗粒度世界无法识别或者彼此特征区别很弱的对象可以换一个粒度世界对其分析,从而全面了解整个问题;以构造性学习算法——覆盖算法为具体实现工具,得到多个商空间中的结果,最终由商空间理论中的函数合成法获得完整结果。实验证明这种基于粒度计算的覆盖算法在解决分类问题时是行之有效的。

关键词 粒度计算,商空间,覆盖算法,分类

The Covering Algorithm Based on Granular Computing

ZHAO Shu ZHANG Yan-Ping ZHANG Ling
(Key Laboratory of Intelligent Computing & Signal Processing of Ministry of Education, Anhui University, Hefei 230039)

Abstract In order to solve the classification problems of many dimensions and large amount of samples better, a covering algorithm based on granular computing is put forward in this paper. The algorithm, whose theory is granular computing, points out that the domain, characters or structure of a problem can be coarser properly when it is analyzed, and a corresponding quotient space is gotten; in the quotient space, the essence of problems can be studied. Then the samples, which are unidentified or whose features are unconspicuous, are classified in the different granular worlds easily. And with the function synthesis method in quotient space theory, different results by the covering algorithm of the different granular worlds are combined finally. The experiments show the rationality and feasibility of this algorithm when the classification problems are analyzed.

Keywords Granular computing, Quotient space, Covering algorithm, Classification

1 引言

在分类问题中,对高维海量数据的处理已有不少的研究,统计学中的主成份分析[1],特征向量的计算[2],知识挖掘中粗糙集方法(属性函数)的约简[3],或是定义人为的标准,对各特征属性进行选择[4]等。

从某种意义上说,这正是粒度思想的一种体现。这种思想和人类思维活动的特点类似,即人类在解决复杂问题时并不一次性地考虑问题的全部细节,而是先把问题分解或简化,忽略其中的部分细节,然后从简化的较抽象的层次开始,层层研究,实现从局部到全局的解决问题的方法。然而,这些方法的思想和人类思维活动的特点还不完全一致。它们的一个共同点就是希望找到一组特征,对论域中所有元素都合适,即利用这些特征能把讨论的所有数据进行合乎要求的分类(指分类问题)。然而要对分类问题中的海量高维数据进行分析研究,采用统一的属性特征约简方式对所有数据进行合乎要求的分类以达到最终的分类结果是很难的,正如人类看问题一样,对诸多不同类事物只从统一的某个角度分析,是很难按要求对它们进行分类的。

Zadeh 的模糊逻辑理论[5~8]、Pawlak 的粗糙集理论[9]、张铃教授和张钹院士的商空间理论[10]正是目前粒度计算最主要的支柱[11]。目前处理复杂、知识不完备、不确定系统各种方法中,粒度计算模型是较为成功的。

为了实现这种在不同粒度世界里进行分类的思想,本文首先介绍这种思想的理论基础——描述粒度世界的商空间理论,主要包括投影划分法和属性函数合成法;随后重点给出基于粒度计算理论的覆盖算法,并用实例说明这种算法的有效性和合理性;最后给出全文结论,并指明下一步研究方向。

2 商空间理论——描述不同粒度世界的数学模型

商空间法[10]就是将不同粒度世界与数学上的商集概念统一起来,表示对象模型的方法,即以商集作为不同粒度世界的数学模型方法。用一个三元组(X,f,T)来描述一个问题。X表示问题的论域;$f(\cdot)$表示论域的属性;T是论域的结构,指论域中各元素的相互关系。分析或求解问题(X,f,T),是指对论域X及其有关的结构属性进行分析、研究。

2.1 投影划分法

设元素x的属性函数是多维的,如n个属性函数分量$f_1,f_2,\cdots,f_n$。若暂不考虑其i个属性$f_1,f_2,\cdots,f_i$,将f_{i+1},$f_{i+2},\cdots,f_n$属性相同的归为一类。

2.2 属性函数合成法

设已知(X_1,f_1,T_1)和(X_2,f_2,T_2),求其合成空间$(X_3,$

*)国家自然科学基金(项目批准号:60475107)资助、973 计划(项目批准号:2004CB318108)、国家自然科学基金(项目批准号:60675031)资助、教育部博士基金(项目批准号:20040357002)资助、安徽省教育厅重点自然科学基金(项目批准号:2006KJ015A)资助、安徽省自然科学基金(项目批准号:0504200208)资助。赵 姝 博士研究生,主要研究领域为智能计算,人工神经网络。

$f_3, T_3)$，即要求属性函数满足下面条件。

1) $p_i f_3 = f_i, i=1,2$ (1)

$p_i : (X_3, f_3, T_3) \to (X_i, f_i, T_i), i=1,2$ 是投影。

2)设 $D(f, f_1, f_2)$ 是某一给定的最优判别准则，则有

$D(f_3, f_1, f_2) = \min D(f, f_1, f_2)$

或 $= \max D(f, f_1, f_2)$ (2)

其中 min(max)是对一切满足(1)式的 X_3 上的一切属性函数 f 取的。

多个商空间(X_i, f_i, T_i)，$i>2$ 的属性合成同理可得。

3 基于粒度计算的覆盖算法

设研究的问题用(X, f, T)表示，暂不考虑结构 T，仅研究(X, f)。这里 X 指的是所有参与学习的样本，f 指的是各样本的属性。算法的主要思想是对 f 颗粒化得出$[f]$，再依此求得$[X]$，由它们构成的商空间$([X], [f])$就是所求的分类器。所取商属性$[f]_i$ 不同，则据此构造的商空间$([X]_i, [f]_i)$也不同，故样本学习后得到多个不同层面的商空间，即不同粒度下的分类器；测试样本时，对这多个分类器$([X]_i, [f]_i)$进行属性合成，求得最后结果。

主算法如下：

S1：根据属性 f 用领域覆盖[12]对 X 求解，得出覆盖组 $C = \{C_1^1, C_2^1, \cdots, C_{n_1}^1, C_1^2, C_2^2, \cdots, C_{n_2}^2, \cdots, C_1^s, \cdots, C_{n_s}^s\}$，其中每一覆盖子组$\{C_1^j, C_2^j, \cdots, C_{n_j}^j\}$只覆盖第 j 类的样本点，n_j 表示第 j 类样本的覆盖领域数，则 C 就是覆盖算法得到的解；

S2：用分量差的方法取属性子集为$[f]_1$，取特征不明显的样本集为论域$[X]_1$。

S3：在$[f]_1$ 构成的粒度世界上对论域$[X]_1$ 利用覆盖算法进行求解，得覆盖领域 C_1；

S4：用分量差的方法取属性子集为$[f]_2$，取特征不明显的样本集为论域$[X]_2$。

S5：$[X]_1 \leftarrow [X]_2$，$[f]_1 \leftarrow [f]_2$，回 step1，直至论域为空或小于某个 n 值。

S6：测试样本。

结束。

在保持样本一定识别率的情况下，可以进一步将研究的粒度世界变粗，对样本进行学习，大大减少网络中神经元的个数。这里将粒度变粗则是采用文[10]中所述的对论域进行颗粒化的方法，由等价关系 R 得出的划分恰是由领域覆盖算法得到的各覆盖领域中的样本。故若上述覆盖领域数≫样本类别数，以覆盖组中每个覆盖的圆心(样本集合)构成的$[X]$作为新的论域 X，回 $s1$。

下面讨论主算法中子过程的实现。

属性子集$[f]_i$，$i=1,2$ 的求法：分量差。

1)对用领域覆盖法求出的每个覆盖取一点对，即取覆盖领域圆心与覆盖领域外最近距离的异类点，构成点对。

$D = \{(o_j^i, s_j) \mid o_j^i$ 为 C_j^i 的圆心，$\min_{s_j \notin y^i} dist(o_j^i, s_j)$，$y^i$ 为 C^i 的类别，$i=1,2,\cdots,s, j=1,2,\cdots,n_i\}$

2)统计点对的各属性分量 f_i 之差的绝对值之和，$[f]_1 = \text{SelectMax}(F, m)$，这里的 m 值一般取特征属性个数的 1/3，在有先验知识的前提下，可以根据经验来确定。

设 $n = n_1 + n_2 + \cdots + n_s$，属性 f 有 t 个分量 $f_1, f_2, \cdots, f_t$。

令 $F_i = \sum_{j=1}^{n} | f_i(d_{1j}) - f_i(d_{2j}) |$，其中$(d_{1j}, d_{2j}) \in D$，$F = \{F_1, F_2, \cdots, F_t\}$，$i=1,2,\cdots,t$。

SelectMax(F, m)，从集合 F 中选取前个 m 最大值。

算法中所述“特征不明显”的样本指的是求出的点对，这些点对有些是原本特征属性不同的异类样本因为取其属性子集导致不同类样本的特征属性$[f]_1$ 完全相同，导致产生噪声，有些是异类样本的属性$[f]_1$ 差异不大，力求换个角度以更好区分它们，故对这些样本，我们将作进一步研究，换个角度(即换个属性集或者说换个粒度空间)进行分析。求解属性子集$[f]_2$ 仍然用分量差的方法。

对样本的测试，设测试集为 T，对 $\forall x \in T$，由不同粒度下的分类器得，其决策属性

$f_{决} = \{[f]_{i决} \mid \max(num[f]_{i决})) \&\& \min(i)$

if $x \in [X]_i, num([f]_{i决}) ++\}$ (3)

其中 $f_{决}$ 表示测试样本的决策属性，即所属类别；$[f]_{i决}$ 表示测试样本 x 在第 i 个分类器$([X]_i, [f]_i)$下得到的类别，函数 $num()$ 表示计数。属性函数合成的最优准则为取最大值 max，如果有多个极大值相等，则取最先出现该值的下标。

4 基于粒度计算的覆盖算法的应用

我们可以将该算法直接应用到实际中，通过和直接使用领域覆盖算法进行比较，说明从同一粒度世界的不同侧面和不同粗粒度世界分析、研究问题的正确性。为了更好地进行文献检索，我们整理了近 3000 篇文章，对其进行分类学习。样本共 10 类，每一个样本向量 15225 维(即 15225 个属性分量)，10 类表示的是统计了 10 种不同类型的文章，如医学、文学、环境等等，每个属性分量表示的是这些文章中每个字出现的频率。

表 1 基于粒度计算的覆盖算法与领域覆盖算法的实验结果比较

	学习样本数	测试样本数	覆盖数	正确识别个数	拒识个数	正确识别率(%)	拒识率(%)
覆盖算法	278	2521	15	2268	243	89.96	9.64
基于粒度计算的覆盖算法	278	2521	12	2505	4	99.36	0.16

可以看出基于粒度计算的覆盖算法的领域覆盖数减少，反映在网络上就是神经元的个数有所减少；直接用覆盖算法的识别率比用基于粒度计算的覆盖算法识别率略低，那是因为后者对于某特征属性集下拒识状态和特征不明显的样本在另一个特征属性集下(即同一个粒度下的另一个侧面)做进一步学习，使得这些样本被划分开，从而提高识别率。实验中取的属性个数和保留覆盖个数对实验结果影响不大，可根据实际情况做具体调整。

测试时，由于不同覆盖所取的属性特征不完全一致，因此测试样本可能属于不同的覆盖领域，进而导致所属类别不完全一致，在该算法中，采取投票的方式，由式(3)得测试样本所属类别的票数最多，就认为此样本属于该类别，若有多个类别票数相同，则以学习时覆盖领域出现的先后为依据，认为该样本属于最早出现的那个领域的类别(票数相同情况下)，这样做的理由是认为第一次所取属性分量是主分量，由于在主分量的情况下某些样本识别不清才选择其他分量继续学习。

为了进一步说明覆盖领域的圆心可以作为粗粒度空间的论域，我们将学习样本数增加，得到如下结果。

表 2 基于粒度计算的覆盖算法与领域覆盖算法以及 SVM 算法的实验结果比较

	学习样本数	测试样本数	覆盖数	正确识别个数	拒识个数	正确识别率(%)	拒识率(%)
覆盖算法	2521	278	27	262	0	94.24	0
基于粒度计算的覆盖算法(较细粒度)	2521	278	24	273	0	98.20	0
基于粒度计算的覆盖算法(较粗粒度)	24	278	14	223	50	80.22%	17.99
SVM 算法	2521	278	/	257	/	92.45%	/

继续将覆盖的圆心作为粗粒度下的样本进行学习,虽然正确识别率有所下降,拒识率有所上升,但这是粒度变粗的必然结果,论域变粗的过程中会丢失一些信息,但是其覆盖数大大减少,计算复杂性大大降低。是否将粒度继续变粗,则要根据实际需要进行。

结论 本文提出粒度计算与覆盖算法想结合的思想,并将其应用于实际数据中,得到理想的实验结果。指出在分析研究某一论域时,可以适当将论域粗化,在粗粒度世界抓住事物主要特征进行讨论,并且对某些在一个粗粒度世界无法识别或识别很弱的特征可以通过不同侧面对其进行分析,甚至可以进一步将讨论的粒度世界粗化,从而全面了解原事物,最终获得论域完整的分类结果。我们目前将不同粒度世界看作同样重要,测试样本时,只是进行一般的投票。事实上,人类在考虑问题时,对问题的不同粒度的分析是有主次之分的,因此,在计算机进行粒度计算时亦可考虑给不同粒度世界赋上权值,对由于不同粒度世界而得到的不同结果可以采取加权的形式处理。这种加权的思想,也正是概率逻辑神经网络的思想。此项工作正在进一步研究中。

参 考 文 献

1 Vapnik V N. Statistical Learning Theory [M]. INC. New York: John Wiley & Sons, 1998
2 边肇祺,张学工. 模式识别[M]. 北京:清华大学出版社,2001
3 Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data[M]. Kluwer Academic Publishers, Dordrecht, 1991
4 张燕平. 机器学习中特征提取的新方法—重复覆盖算法[J]. 安徽大学学报,2002(3)
5 Zadeh L A. A Theory of Approximate Reasoning [M]. New York: Machine Intelligence, 1979. 149～194
6 Zadeh L A. Fuzzy Logic = Computing With Words [J]. IEEE Transactions on Fuzzy Systems, 1996(4):103～111
7 Zadeh L A. Towards Theory of Tuzzy Information Granulation and Its Centrality in Human Reasoning and Fuzzy Logic [J]. Fuzzy Sets and Syatema, 1997(19):111～127
8 Zadeh L A. Announcement of GrC, 1997. http://www.cs.uregina.ca/～yyao/GrC/
9 Pawlak Z. Rough Sets Theoretical Aspects of Reasoning about Data [J]. Dordrecht, Boston, London: Kluwer Academic Publishers, 1991
10 张铃,张钹. 问题求解的理论及应用[M]. 北京:清华大学出版社,1990
11 Zhang Ling. Granular Computing based on Rough Sets, Quotient Space Theory, and Belief Functions [J]. Lecture Notes in Computer Science, 2003, 2871:152～159
12 张铃,张钹. 多层前向网络的交叉覆盖设计算法[J]. 软件学报,1999,10(7):737～742

(上接第 218 页)

由对比结果可以看出,基于优势关系的信息系统正域约简在保持数据原有信息和基础上,可以有效地进行规则获取,对于处理具有序关系的数据集,两种算法的样本识别率大致相同,总体上 PDRIS 算法对未知样本的误识率比算法 A1 低,正确识别率基本比算法 A1 高,这说明用优势关系处理序信息系统比先将信息系统离散化后用等价关系来处理能得到更好的结果。

另外,基于优势关系对序信息系统获取的规则更符合人类对现实世界的认识。如 Iris(鸢尾属植物)标准数据集,每个元组包含 Iris 花的 4 个属性(萼片的长、宽,花瓣的长、宽),共有 3 种不同的花(Iris Setosa, Iris Versicolour, Iris Virginica)。该数据集获取的其中三条规则为:

$b \geqslant 3.5 \wedge c \geqslant 1.3 \wedge d \geqslant 0.2 \rightarrow d=0(\alpha=1,\beta=10/75)$,

$b \geqslant 2.7 \wedge c \geqslant 4.9 \wedge d \geqslant 1.8 \rightarrow e=2(\alpha=1,\beta=22/75)$,

$a \geqslant 5.0 \wedge b \geqslant 2.3 \wedge c \geqslant 3.3 \rightarrow d=1(\alpha=2,\beta=21/75)$。

即当萼片宽大于等于 3.5cm,花瓣长大于等于 1.3cm,花瓣宽大于等于 0.2cm 时该花是 Iris Setosa,且在该测试集中满足这种特性的元组有 10 个,另两条规则的含义类同。

结论 传统的粗糙集规则获取方法在处理连续属性时都先对连续属性进行离散化等处理,导致数据信息部分丢失。要想从基于优势关系的不协调信息系统中获取简洁的知识就必须对系统进行知识约简。本文提出的基于优势关系的不协调信息系统正域约简并获取规则的方法有效解决了连续属性和偏序关系问题,有一定的实用价值。然而基于属性重要性的正域约简算法在执行效率上有待改善,寻找更高效的属性约简方法是我们接下来的研究内容。

参 考 文 献

1 Pawlak Z. Rough Sets. International Journal of Computer and Information Sciences, 1982, 11:314～356
2 Beynon M. Reducts within the Variable Precision Rough Set Model: a further Investigation. European Journal of Operational research, 2001, 134:592～605
3 Pawlak Z. Rough Sets—Theoretical Aspects of Reasoning about Data. Boston: Kluwer Academic Publishers
4 Wu W Z, Zhang M, Li H Z, Mi J S. Knowledge Reduction in Random Information Systems via Dempster-Shafer Theory of Evidence. Information Sciences, 2005, 174(3-4): 143～164
5 王国胤. Rough 集理论与知识获取. 西安:西安交通大学出版社, 2001
6 Stefanowski J. Handling Continuous Attributes in Discovery of Strong Decision Rules. RSCTC'98, 1998. 394～401
7 Sai Y, Yao Y Y, Zhang N. Data Analysis and Mining in Order Information Table. 2001 IEEE Int. Conf. on Data Mining. IEEE Computer Society Press, 2001. 497～504
8 Shao M W, Zhang W X. Dominance Relation and Rules in an Incomplete Ordered Information System. International Journal of Intelligent Systems, 2005, 20:13～27
9 张文修, 姚一豫, 梁怡. 粗糙集与概念格. 西安:西安交通大学出版社, 2006
10 张文修, 米据生, 吴伟志. 不协调目标信息系统的知识约简. 计算机学报, 2003, 26(1):12～18
11 徐伟华, 张文修. 基于优势关系下的协调近似空间. 计算机科学, 2005, 32(9):164～166
12 邵明文, 张红英. 序信息系统上的优势关系与规则获取. 工程数学学报,2005, 22(4):697～702
13 谢宏, 程浩忠, 牛东晓. 基于信息熵的粗糙集连续属性离散化算法. 计算机学报, 2005, 28(9):1570～1574