

基于变精度粗集的分类方法^{*})

赵永强 潘 泉 张洪才

(西北工业大学自动控制系 西安710072)

摘 要 基于差别矩阵的约简算法简单、计算量小,但是传统的差别矩阵不能处理噪声数据。为提高差别矩阵的适用范围,提出一种能够容忍数据中包含噪声的变精度差别矩阵,并给出改进的基于条件属性偏序关系的约简算法。最后,将这一方法用于对多类图像的分类过程中,将分类结果与BP网络的分类结果和基于传统Skowron差别矩阵方法的分类结果相比较表明这种分类方法具有较好的结果。

关键词 变精度粗集,差别矩阵,图像分类

Classification Based on Variable Precision Rough Set

ZHAO Yong-Qiang PAN Quan ZHANG Hong-Cai

(Department of Automatic Control, Northwest Polytechnical University, Xi'an 710072)

Abstract The reduction algorithm based on discernibility matrix is simple and easy, but it can't deal with noisy data. To make discernibility matrix more useful, we propose a new version of discernibility matrix, so called variable precision discernibility matrix, which can tolerate the noise of information, in addition, a reduction algorithm is also presented based on the partial order relation of the conditional attribute and used to do image classification. Compared with traditional BP network and Skowron discernibility matrix based method, the result will be much better.

Keywords Variable precision rough set, Discernibility matrix, Images classification

1 引言

分类是对目标所属的类型做出判断,目前已有的分类方法多种多样,而且一些分类方法已在工程中广泛应用。神经网络是应用很广泛的一种分类方法,但是由于传统神经网络自身的缺陷^[3],如:1)初始权值的选择的不适当可以导致网络不收敛;2)网络训练时间相对较长;3)可能陷入局部极小问题而得不到最优解等,而且它不能很好地解决多类分类问题。因此许多学者提出了基于传统神经网络的改进方法如:模型神经网络、神经网络集合(ensemble)等,同时一些学者将其它新的理论应用到目标分类中并取得了很好的分类效果,粗集理论就是其中之一。

粗集理论提供了一个新的分类方法,它是建立在假设论域中的每个对象之间都有一定的相关信息^[1]。最近, Cyran^[5,6]、Slezak^[7]、Bell^[8]和 Daijin Kim^[9]将基于粗集的启发式推理应用于寻找最优特征和得到分类规则,并取得了较好的效果。但是,所有这些应用都是基于定义在精确集上的粗集来进行的并且假定决策表是完备的没有噪声的,因此这些基于传统粗集的分类方法应用在含有大量噪声数据的情况时性能就会大大降低。Ziarko^[12]提出了变精度粗集的概念,把粗集定义在概率空间上,这给研究者用粗集来处理噪声数据提供了一个好的方法。本文将在 Ziarko 变精度粗集的基础上提出概率差别矩阵的概念,并给出基于属性重要性的完备的约简算法,最后将其应用于由含噪声图像不变矩特征组成的信息表的处理中,并与 Skowron 差别矩阵的处理结果进行比较。

2 基本概念

在 Pawlak 粗集中,一个信息表系统 I 定义为 $I = \{U, A\}$, 其中 1) $U = \{u_1, u_2, \dots, u_{|U|}\}$ 是一个有限、非空对象集合,称为论域;2) $A = \{a_1, a_2, \dots, a_{|A|}\} = C \cup D$ 是一个有限、非空的属性集合(它包括条件属性集合 C 和决策属性集合 D);3) 对任一 $a \in A$ 存在一个映射 a : 从 U 到另一空间, $a: U \rightarrow a(U)$, 其中 $a(U) = \{a(u) | u \in U\}$ 称为属性 a 的值域。在粗集理论中一个重要的概念是等价关系 R :

$$R(B) = \{(x, y) | (x, y) \in U \times U,$$

$$\forall b \in B (b(x) = b(y))\} B \subseteq A$$

在粗集理论中另一个重要的概念是差别矩阵 $M^{[2]}$, 其定义如下:

$$M(C) = \{m_{i,j}\}_{|C| \times |C|}$$

$$m_{i,j} = \begin{cases} \phi & d(u_i) = d(u_j) \\ \{c \in C | c(u_i) \neq c(u_j)\} & d(u_i) \neq d(u_j) \end{cases}$$

差别矩阵 M 中的元素是可以区分对象 u_i 和 u_j 的属性集合。根据 Skowron^[2]的证明,通过差别矩阵求取信息系统的核和约简是很容易的,但是求取最小约简是一个 NPC 问题。在用粗集进行分类时,为了保证分类算法的鲁棒性,需要一定的冗余信息,因此不需要计算信息表的最小约简。而在目标分类中,比较重要的衡量算法的标准是算法的计算复杂性和算法的时间复杂性。因此,在求取信息表的约简时,需考虑约简算法的计算复杂性和算法的时间复杂性。

由差别矩阵的定义可以看出通过差别矩阵来处理带有噪

^{*}) 基金项目:国防十五重点预研项目(102010302),研究生创新基金(Z20030046),校青年基金(521020101-0900-020101)。赵永强 博士生,主要研究领域为信息融合、小波分析、智能计算等。潘 泉 博导,主要从事小波理论、图像处理、多目标跟踪、信息融合、智能控制方面的研究。张洪才 博导,主要从事多目标跟踪、信息融合、智能控制方面的研究。

声的信息是很难的,处理结果的可信度也难以保证,这是因为差别矩阵是完全定义在精确的等于不等于关系上的。通过这种差别矩阵的定义所获得的分类规则难以保证有较好的鲁棒性,因此有必要对差别矩阵的定义进行修正以得到鲁棒的分类规则。Ziarko 给出了变精度粗集(VPRS)的模型^[12],这一模型为处理噪声信息提供了一个强有力的工具,这将在第3部分详细描述。

3 VPRS 约简

Pawlark^[11]和 Skowron^[2]所提出的粗集理论是定义在精确数据集上的,但是在实际的信息表中往往包含大量的噪声,用定义在精确数据集上的粗集来处理这种信息表会带来很大的误差。处理这种噪声数据的一种方法是应用概率统计

$$m_{i,j} = \begin{cases} \phi & (P(D|[u_i]_{IND}) \geq \beta \wedge P(D|[u_j]_{IND}) \geq \beta) \vee (P(D|[u_i]_{IND}) < \beta \wedge P(D|[u_j]_{IND}) < \beta) \\ \{c \in C; |c(x_i) - c(x_j)| > NTh\} & (P(D|[u_i]_{IND}) \geq \beta \wedge P(D|[u_j]_{IND}) < \beta) \vee (P(D|[u_i]_{IND}) < \beta \wedge P(D|[u_j]_{IND}) \geq \beta) \end{cases}$$

$M(C)$ 叫做概率差别矩阵, NTh 是一个由信息表中噪声程度决定的噪声阈值可以通过自适应的方法计算获得, β 是在 $(0, 0.5]$ 取值的数^[12]。在概率空间定义差别矩阵是用“主要包含关系”代替“等于关系”,这样就可以提高系统的处理噪声数据的能力。

Pawlark^[11]提出一个约简 R 必须满足独立性条件,即: $\phi \neq R \subseteq C, \forall r \in R, POS_R(D) \neq POS_{(R-r)}(D)$,在此 POS 是条件属性 R 依据决策属性 D 所确定的正区域。一个约简如果满足独立性条件则称为 Pawlark 约简,否则为约简^[4];如果一个算法能够确保对于任何信息系统得到 Pawlark 约简,则这一约简算法为完备的。约简算法的完备性对于目标分类来说是很重要的,只有用完备的约简算法对信息表进行约简才能保证约简后的信息表中的信息没有缺失,并且分类能力不变。

WangJue^[4]证明了基于条件属性偏序关系的约简算法可以得到 Pawlark 约简,并且是一个完备的约简算法;赵凯^[14]在此基础上,提出一种“针对用户需求的约简算法”,该算法以用户对条件属性的关心度建立偏序关系,但未证明该算法是完备的。下面将以对“针对用户需求的约简算法”作适当的改进使其能完成基于概率差别矩阵的约简,运算过程进一步简化并证明其完备性。

对于一个信息系统 $I = \langle U, A \rangle, n = card(A), M$ 是信息系统的概率差别矩阵,并通过概率差别矩阵来计算信息系统的核属性集 $Core$ 。在计算 M 之前,对于所有的条件属性依据属性重要性(由先验信息得到)定义一个偏序关系 $PO: a_1 < a_2 < \dots < a_n$ 。在计算概率差别矩阵时,让概率差别矩阵中的矩阵元素 $m_{i,j}$ 中属性集合满足偏序关系 PO 。

算法起始时让约简 $Re = \phi$,约简算法描述如下:

- (1) 取 $Re = Re \cup Core, m_{i,j} = m_{i,j} - Core, C = C - Core$;
- (2) $c = FirstOd(C)$,从 M 中选择第一个出现的包含 c 的差别项 a ,并取 $a = cB$;
- (3) 如果 $a = \phi$,则回到(2);
- (4) $Re = Re \cup c, m_{i,j} = m_{i,j} - \{c\}, m_{i,j} = m_{i,j} - B$;
- (5) 如果 $m_{i,j} \neq \phi$ 则回到(2),否则得到约简 Re 并退出。

下面将证明该算法是完备的。在此之前,首先引入约简的独立性定义、Pawlark 约简的定义和算法完备性的定义。

Pawlark^[4]提出一个约简 R 必须满足独立性条件,即: $\phi \neq R \subseteq C, \forall r \in R, POS_R(D) \neq POS_{(R-r)}(D)$,在此 POS 是条件属性 R 依据决策属性 D 所确定的正区域。一个约简如果满足独立性条件则称为 Pawlark 约简,否则为约简^[3];如果一个算法能够确保对于任何信息系统得到 Pawlark 约简,则这一约简

的方法,因此将粗集定义在概率空间上就可以提高其处理噪声数据的能力。Ziarko^[12]在粗集上定义了如下概率:

$$P(D|[u_i]_{IND}) = \frac{card(D \cap [u_i]_{IND})}{card([u_i]_{IND})}$$

这一概率描述了决策为 D 的 u_i 等价类的数量占 u_i 等价类的总量的比例。公式中 $[u_i]_{IND}$ 表示的 u_i 的等价类集合, $card([u_i]_{IND})$ 表示集合 $[u_i]_{IND}$ 中所包含元素的数量。根据定义 $P(D|[u_i]_{IND})$ 和 Skowron 差别矩阵的概念,我们将 Skowron 差别矩阵推广到概率空间上定义,得到概率差别矩阵(PDM):

定义1 对一给定的信息系统 $I = \langle U, A = C \cup D \rangle$,用概率定义差别矩阵 $M(C)$:

$$M(C) = \{m_{i,j}\}_{|C| \times |C|}$$

算法为完备的,只有用完备的约简算法对信息表进行处理才能保证得到的是一个约简而不是近似约简,使得最终的分类规则能达到相对最小。根据 Pawlark 的定义,可以导出如下命题:

命题1^[4] Re 是信息系统 $I = \langle U, A, D \rangle$ 的一个 Pawlark 约简,当且仅当 Re 是信息系统 $I' = \langle U', Re, D \rangle$ 的核属性。其中 U' 是当条件属性为 Re 时,从论域 U 中构造的对象集合。

下面利用命题1来证明算法的完备性。

证明:如果上面的算法要结束, Re 必须满足 $M' = \{a; a \cap Re = \phi, a \in M\} = \phi$,则 Re 是信息系统 I 的一个约简。

因此只需证明 Re 是信息系统 I' 的核属性:

I' 的差别矩阵为 M_r ,定义集合 $M'_r = \{a; a \cap (Re - \{r\}) = \phi, a \in M_r\}$,如果 $M'_r = \phi$,则 r 在 Re 中是可分辨的,因此 $Re - \{r\}$ 是 I 的一个约简而非 Pawlark 约简^[3],因此必须证明 $M'_r \neq \phi$ 。假设 r 不是 I' 的核属性,对于 $\forall a \in M_r, a$ 取值为 $r\{b\}$ 并且 $\phi \neq \{b\} \subset A, \{b\} \cap Re = \phi$,但是在信息系统 I', U' 仅包含由 Re 确定的元素,因此有 $\{b\} = \phi$,因此 r 为信息系统 I' 的核属性。

由命题可知,由算法得到的约简 Re 是 Pawlark 约简,因此算法是完备的,证毕。

4 实验仿真

在这一部分,我们将分别用本文中提出基于概率差别矩阵的方法和基于 Skowron 差别矩阵方法来对由9类(180幅)飞机图像的不变矩特征组成的信息表^[16]进行约简,然后用最近距离法进行分类^[9]比较分类效果。在图1,列出了包含噪声的9类飞机的图像。

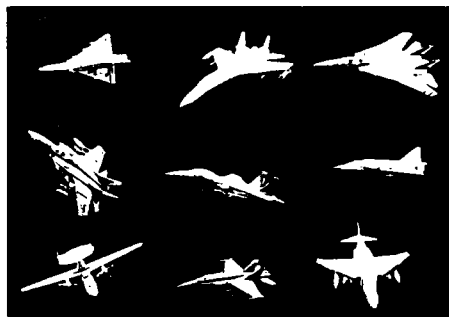


图1 9类飞机图像

4.1 数据预处理

首先对这些飞机图像进行基于矩的不变特征提取,常用的矩不变特征提取方法有:Hu 不变矩^[16]、小波不变矩^[13,15]、Zernike 不变矩等。考虑到本文主要处理噪声数据的影响和相似目标的分类,以及小波不变矩对噪声的敏感和细节信息的描述等方面的优势,在此用小波不变矩对图像进行特征提取。

提取180幅目标图像的小波不变矩得到一个69列(属性数)180行(目标数)一个信息表,决策属性为每个目标所对应的类别。取90行数据作为训量数据,全部180行数据作为待检测数据用BP 网络进行分类。首先计算待检测小波矩数的散度^[15]进行特征提取,然后用提取的小波矩特征对神经网络进行训练,然后对待检测信息表进行分类得最终率不超过60%,并且在整个训练过程中所化的训练时间很长(所用仿真

软件平台:MATLAB6.1,硬件平台:处理器 P4 1.6G, 256DDR 内存,60G 硬盘,训练所用时间约为1300s),训练过程的收敛也很困难。

4.2 粗集分析

得到信息表之后,分别计算传统的差别矩阵和本文中所提到的概率差别矩阵,然后用算法1分别进行属性约简。在进行约简时,将信息表中每类目标的数量不断地增加(每类飞机共有20个目标,待分类的目标取为全部目标),通过仿真发现约简后属性的数量、最终的分类正确率都会随着目标数量有相应的边变化,并比较通过两种差别矩阵进行约简所化的时间(所用仿真软件平台:MATLAB6.1,硬件平台:处理器 P4 1.6G,256DDR 内存,60G 硬盘),具体数据如下表1所示。

表1 变精度粗集于传统粗集在约简后的属性数量和分类正确率上的比较

总目标数		36	45	54	63	72	81	90	99	108	117	126	135	144	153	162	171	180
变精度方法	约简时间(s)	0.8	1.2	1.9	2.5	3.3	4.2	5.4	6.6	8.0	9.2	10.5	11.2	13.9	15.1	16.9	18.9	18.9
	属性数	7	8	5	7	7	7	7	7	10	10	10	10	10	14	14	14	14
	分类正确率%	33	40	44	56	62	64	70	74	80	83	86	89	95	99	99	100	100
传统方法	约简时间(s)	0.8	1.3	1.8	2.5	3.3	4.2	5.3	6.5	8.7	9.3	10.7	11.3	13.8	14.4	17.8	19.9	22.0
	属性数	5	5	5	6	6	6	7	7	7	7	7	7	7	7	7	7	7
	分类正确率%	36	45	49	54	57	60	63	69	73	76	81	84	89	95	97	99	100

从表1可以看出,随着信息表中目标数量的增加,约简后的属性数量会作相应的增加,相应的识别率也逐渐地增加直到全部正确分类。这说明,在用粗集进行分类时,信息表中信息的数量是有一定的要求而且由于粗集分类是一种经验分类方法对在信息表中没有出现的新对象的分类能力较差,使得当信息表中的总目标数较小时,分类正确率很低,而当信息表中包含足够的信息时,分类正确率相应地很高。从总体上看,用本文所提的概率差别矩阵方法所对应的分类正确率要比传统的差别矩阵方法所对应的分类正确要高,并且约简所化的时间也少,由此可以看出本文所提的方法对于噪声数据分类效果较好。

结论 本文讨论了用变精度粗集对噪声数据进行分类,并且分析了 Skowron 差别矩阵处理噪声数据的缺陷,构造了概率差别矩阵以处理噪声数据并提出相应的约简算法。通过实验仿真,可以看出这种方法可以很好地处理噪声数据,并能得到很好的分类结果。

参考文献

- 1 Pawlark Z. Rough set—Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht, Boston, London. 1991
- 2 Skowron A, Rauszer C. The Discernibility Matrix and Function in Information Systems. Intelligent Decision Support—Handbook of Application and Advances of the Rough Sets Theory. In: Słoiniski R ed. 1991. 331~362
- 3 Ahlqvist O, Keukelaar J, Oukbir K. Rough classification and accuracy assessment. Int. J. Geographical information science. 2000, 14: 475~496
- 4 Wang Jue, Wang Ju. Reduction Algorithms Based on Discernibility Matrix: The Ordered Attributes Method. Journal of Computer Science and Technology, 2001, 16: 489~504

- 5 Cyran K, Mrozek A. Rough Set in Hybrid Methods for Pattern Recognition. International Journal of Intelligent Systems, 2000, 15: 919~938
- 6 Cyran K, Mrozek A. Rough Set in Hybrid Methods for Pattern Recognition. International Journal of Intelligent Systems, 2001, 16: 149~168
- 7 Slezak D, Wroblewski J. Classification Algorithm Based on Linear Combination of Feature. In: Proc of PKDD'99 Springer-Verlog 1999. 548~553
- 8 Bell D A, Guan J W. Computational Methods for Rough Classification and Discovery. Journal of the American Society for Information Science, 1998, 5: 403~414
- 9 Kim D, Bang S-Y. A Handwritten Numeral Character Classification Using Tolerant Rough Set. IEEE Trans on Pattern Analysis and Machine Intereence, 2000, 22: 923~937
- 10 Fayyad U M, Irani K B. On the handing of continuous valued attributes in decision tree generation. machine learning, 1992, 8: 87~102
- 11 Getis A. Some thoughts on the impact of large data set on regional science. Annals of Regional Science, 1999, 33: 145~150
- 12 Ziarko W. Variable Precision Rough Set Model. Journal of Computer and System Sciences, 1993, 46: 39~59
- 13 Shen D, Horace H S Ip. Discriminative Wavelet Shape Descriptors for recognition of 2-D Patterns. Pattern Recognition, 1999, 32: 151~165
- 14 赵凯, 赵岷, 王珏. 一种针对用户需求的约简算法. 计算机科学, 2002, 29 (9. Special): 329~334
- 15 徐旭东, 周源华. 基于小波矩不变量的模式识别方法. 红外与毫米波学报, 2000, 19: 215~218
- 16 赵永强, 张洪才, 潘泉. 基于粗集理论的鲁棒分类器. 计算机科学, 2002, 29 (9. Special): 120~122