doi:10.3772/j.issn.1002-0470.2015.10-11.002

基于优选特征属性偏序结构分析的白细胞图像分类规则发现①

郝连旺②*** 洪文学* 魏 鹞③**

(*燕山大学电气工程学院 秦皇岛 066004)

(**秦皇岛市第一医院 秦皇岛 066000)

摘 要 基于形式概念分析和属性偏序结构理论,提出了一种白细胞图像分类规则发现模式,从而建立了高效的白细胞图像分类方法。用该方法,首先在大量的白细胞图像区域特征测定实验基础上对白细胞图像优选特征进行了离散化处理,针对白细胞图像数据集构建了形式背景,依据分层类坐标矩阵的属性偏序结构生成方法生成了白细胞图像数据集属性偏序结构图;然后通过对属性偏序结构图分析,发现了6 类白细胞相应的6 条分类规则;最后,依据分类规则建立了二分树分类器,并且在实际白细胞图像数据集测试实验中取得了94.04% 的平均分类精度,该精度明显高于其它3 种经典算法,证明了基于优选特征属性偏序结构分析获取的白细胞图像分类规则的可用性、简单性和有效性。

关键词 形式概念分析, 属性偏序结构, 白细胞图像, 分类规则

0 引 言

人体血液中的白细胞种类及数值变化情况表征了人体健康状况[1]。人体外周血正常白细胞可分为六类,即分叶核中性粒细胞(SEG)、杆状核中性粒细胞(BAN)、淋巴细胞(LYM)、单核细胞(MON)、嗜碱性粒细胞(BAS)及嗜酸性粒细胞(EOS)。彩色白细胞图像自动分类技术是面向目标对象的有监督模式分类问题,需要把所处理的图像数据从所在的模式空间映射到特征空间再映射到类型空间[2],涉及到计算机视觉、图像理解、模式识别、知识表示等诸多研究领域。为了通过模型来预测类标记所未知的对象类,分类问题必须找出能够描述、区分数据类的模型[3]。并且,实际的白细胞图像分类问题往往都是混合属性数据非平衡分类问题,因此,采用传统的分类方法,难以得到满意的分类效果[4]。

数据集内结构的概念分析是一个非常令人感兴趣的研究领域。概念格是形式概念分析理论的核心数据结构[5],属性偏序结构是根据属性偏序的性质和数学意义建立的一种属性层次结构,由经过概念格中选定顶点的完全子格构成,是一种基于二元关系建立的层次结构[6],适用于提取分类规则。为此,本文在文献[7]完成的对白细胞图像类间特征优选工作的基础上,提出了一种基于优选特征属性偏序结构分析的白细胞图像分类规则发现方法。

1 属性偏序结构分类方法

形式概念分析由德国数学家 Wille 于1982 年提出,它反映了概念间的泛化与例化关系的二元关系[8]。一般来说,在形式概念分析理论指导下进行分类有两种方式,一种是利用概念格寻找分类关联规则,并运用经验或启发式构造分类器;另一种是利用概念的层次关系和概念与内涵的同一性寻找概念,以构成概念分类器[9]。属性偏序结构是一种基

① 国家自然科学基金(61273019),河北省科学技术研究与发展计划(12270329)和燕山大学博士基金(B897)资助项目。

② 男,1979 年生,博士,副研究员;研究方向:模式识别,医学信息处理;E-mail: haolw@ysu.edu.cn

③ 通讯作者,E-mail: weikun@medmail.com.cn

(收稿日期:2015-04-14)

于二元关系建立的层次结构,能够构成概念分类器,更适于提取分类规则。概念格是一种以概念为层次的结构关系,可以提取待识别对象不同层次的多尺度知识描述,方便求解分类规则模式。但由于概念格缺乏层次,人工不易发现概念间关系。而属性偏序结构是一个下部封口的倒树形结构,比概念格更加简单明了,意义更加清晰,可视化效果更好,更加利于对属性特征的识别,便于剔除冗余信息,挖掘显著特征知识体系。

1.1　分层类坐标矩阵属性偏序结构生成方法

分层类坐标矩阵是形式背景矩阵向属性偏序结构转化的中间过程,是矩阵化的偏序结构,它将原来由形式背景矩阵到偏序结构图的转化简化为由形式背景矩阵到分层类坐标矩阵的转化。一个偏序结构只对应一个分层类坐标矩阵,完全包含了属性分层的坐标信息和偏序结构图中的对象、属性信息。生成分层坐标矩阵的流程见图 1。

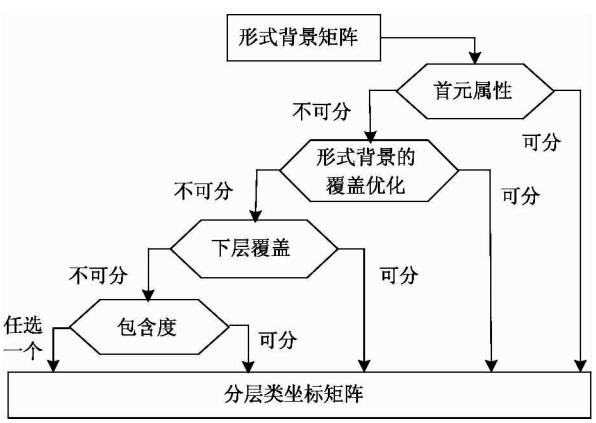

图 1　生成分层坐标矩阵的流程图

属性偏序结构图生成思路如下:首先,将原始形式背景进行预处理,得到形式背景矩阵;然后,对形式背景矩阵进行处理,使形式背景矩阵转化为分层类坐标矩阵;其次,由于分层类坐标矩阵只具有相对坐标,不具有绝对坐标,需将其转化为可存储偏序结构图结点绝对坐标的 X 坐标矩阵和 Y 坐标矩阵;最后,将 X 坐标矩阵和 Y 坐标矩阵中对应的结点连结起来,并在相应位置标注属性信息和对象信息,就构成了属性偏序结构图。

1.2　属性偏序结构分类应用流程

属性偏序结构是以研究属性间的相关性为目标的数据挖掘工具,是建立在形式概念分析基础上的,是对给定形式背景中属性层次关系的研究。本文提出了如图 2 所示的属性偏序结构分类应用流程,应用属性偏序结构关系对白细胞图像已经优选的特征做了进一步分析,目的是挖掘出白细胞图像优选特征分类规则。

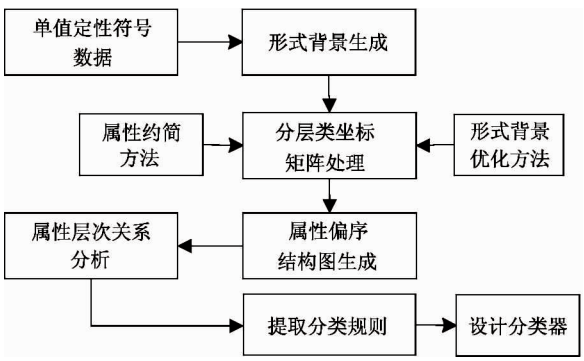

图 2　属性偏序结构分类应用流程图

2　白细胞图像优选特征属性偏序结构分析

2.1　实验数据来源

本文所用的 A、B 两组实验数据来源分别为秦皇岛市第一医院提供的健康男性成人对象甲(一人隔天二次采集)、乙丙(二人隔天三次采集)外周血血涂片。血涂片制作采用常规 Wright 染色,用 Olympus BX50 显微镜将血涂片细胞经油镜放大 l000 倍,通过彩色摄像机进入 NYD-100 型医学图像分析系统,以 BMP 文件方式采集。考虑到彩色细胞图谱亮度的微小变化会对颜色直方图分布产生较大影响,对原始样本图片进行了预处理,将图片中的白细胞从背景中分割,调整亮度[10],并进行了人工分类,实验数据情况见表 1。

表 1　实验数据情况表

数目(幅)	BAN	SEG	EOS	BAS	LYM	MON
A 组	44	33	48	40	51	65
B 组	134	148	81	53	154	101

2.2 优选特征选取

为了实现用较低的类间特征维数对正常人体外周血白细胞高效分类,文献[7]提出了基于属性层次关系的白细胞图像类间特征选取方法,本文在此基础上,进行了深入研究。本文首先选取了常见的白细胞图像70个特征,并用基于类间重叠系数矩阵的白细胞图像特征预选方法优选出了18个预选特征,然后在对六分类白细胞图像特征背景进行基于属性偏序结构的属性选取后,通过定义并计算属性度数值大小,有效划分了类间特异性较高的属性和类间普遍性较高的属性,将70种类间特征优化为7种。按照功能可以划分为如表2所示的属性特征群。

表2 按照功能划分的属性特征群

类间区分能力	代号	特征描述
类间普遍特征	e,o,g,c	细胞核圆形度,对细胞浆按照灰度级差分法提取的纹理特征,细胞浆区域色调 H 参数平均值,细胞核面积占分比
类间特异特征	i,m,h	细胞浆区域 r 参数均方差值,细胞核、浆之间 S 参数平均值对比,细胞浆区域 R 参数平均值

2.3 优选特征离散化分析

本文对2.1节中A组数据进行了分割实验,经过预处理后,并按照标准的白细胞图像特征测定方法,测定了相关特征数值范围。经观察分析,得出如下结论:

(1)类间特异特征在至少2个白细胞类别对间起到明显的划分效果,且在白细胞图像六分类行为中,可以较好划分对象类。选用图像数据标准,由标准偏差值可知每类白细胞图像的类间特异特征值分布较为集中;当 $i \geqslant 95$ 时,极可能为MON、EOS、SEG;当 $m \geqslant 58$ 时,极可能为LYM、BAN;当 $h \geqslant 250$ 时,极可能为BAN、SEG。可构建如表3的形式背景图。

表3 类间特异性特征形式背景图

	BAN	SEG	EOS	BAS	LYM	MON
$i \geqslant 95$		×	×			×
$m \geqslant 58$	×				×	
$h \geqslant 250$	×	×				

(2)类间普遍特征具有较好的分类性能,每个普遍特征都能仅依靠自身特性将白细胞图像至少分为2~4个对象类。选用图像数据标准,由标准偏差值可知每类白细胞图像的类间普遍特征值分布较为集中,且具有阶段性;细胞核面积占分比 c 和细胞浆区域色调 H 平均值 g 虽然在六分类中具有较好可分性,但在六分类中更具有较明显的二分性,即分为二个对象类。LYM、BAS的核面积占分比明显大于BAN、SEG、EOS、MON;BAS、EOS的胞浆区域色调 H 参数平均值明显大于LYM、SEG、EOS、BAN。用形式背景方式表示如表4。

表4 属性c和g的形式背景图

	BAN	SEG	EOS	BAS	LYM	MON
$c \geqslant 52$				×	×	
$g \geqslant 256$			×	×		

(3)细胞核圆形度 e 特征值分布相对集中,能够划分固定区间,便于离散化。细胞核可按照圆形度区间可划分为很圆 $e_1(0 \sim 0.6)$、较圆 $e_2(0.6 \sim 1.2)$、一般 $e_3(1.2 \sim 1.8)$、近方形 $e_4(1.8 \sim 2.4)$、不规则 $e_5(2.4 \sim 3.0)$。

以上通过分析类间特异特征、普遍特征,从而将数据离散化。这些特征涉及形状特征、色彩光密度特征、纹理特征,物理特性跨度较大,从物理意义上难以融合,因此本文暂不考虑按照灰度级差分法提取的纹理特征o。

2.4 构建优选特征形式背景

为便于分析,本文在A组数据样本集中随机抽取了27幅白细胞图像进行分析。根据形式概念分析中构建形式背景的理论,由离散化后的数据可以构造出优化后的白细胞图像数据集的全属性(包括决策属性)形式背景(表5)。形式背景中每个实例代表一个预处理后白细胞图像。该数据集是对应生

成的六个条件属性和六个决策属性的数据集合。其中条件属性：A = {e,$c \geqslant 52$,$g \geqslant 256$,$i \geqslant 95$,$m \geqslant 58$,$h \geqslant 250$}，决策属性：D = { BAS, MON, EOS, LYM, BAN, SEG }。

表 5 离散化后的白细胞图像数据集形式背景

	e_1	e_2	e_3	e_4	e_5	c_y	g_y	i_y	m_y	h_y	BAN	SEG	EOS	BAS	LYM	MON
1	×					×									×	
2	×							×	×	×						×
3	×					×	×	×	×	×			×			
4	×							×	×							×
5		×				×	×	×	×	×			×			
6		×				×	×	×	×	×			×			
7		×						×	×							×
8		×				×			×	×					×	
9		×						×	×							×
10		×					×	×		×		×				
11			×					×		×	×					
12			×					×	×							×
13			×				×	×		×		×				
14			×			×	×	×	×	×			×			
15			×					×	×							×
16				×				×		×	×					
17				×			×	×		×		×				
18				×		×	×	×	×				×			
19				×		×	×							×		
20				×		×	×		×					×		
21				×		×									×	
22					×			×		×	×					
23					×	×	×	×	×				×			
24					×		×		×					×		
25					×		×	×		×		×				
26					×	×	×							×		
27					×	×			×	×					×	

2.5 构建白细胞图像数据集属性偏序结构

按照 1.1 中基于分层类坐标矩阵的属性偏序结构图生成方法，由表 5 可以生成白细胞图像数据集属性偏序结构图。为实现只对各种综合特征分析，就能判断出实例分类，本文将决策属性去掉，建立了去掉决策属性后的偏序结构图，见图 3。

3 分类规则发现及应用分析

3.1 分类规则内容

属性偏序结构是一种基于二元关系建立的层次结构，适用于提取分类规则。模式是属性的有序组合，属性偏序结构图中的每一条路径就表示一种模式。通过对不同模式的分析，可以发现对于分类有指导意义的知识。通过对图 3 的分析，可以发现如下分类规则：

(1) 如果条件属性 a7，a10，a13，a14 同时成立，那么可以得到状况 o1 = 10，o2 = 17，o3 = 25，o4 = 13，也就是说，当白细胞图片中胞浆区域 r 参数均方差值和胞核区域 R 参数平均值均较大，并且细胞核面积占分比和核浆间 S 参数平均值对比均较小时，可

以判定此白细胞可能为分叶核中性粒细胞。同时,该细胞胞核圆形度值较大,呈现不规则状,且胞浆区域色调 H 平均值较大。如实例 10、13、17、25。

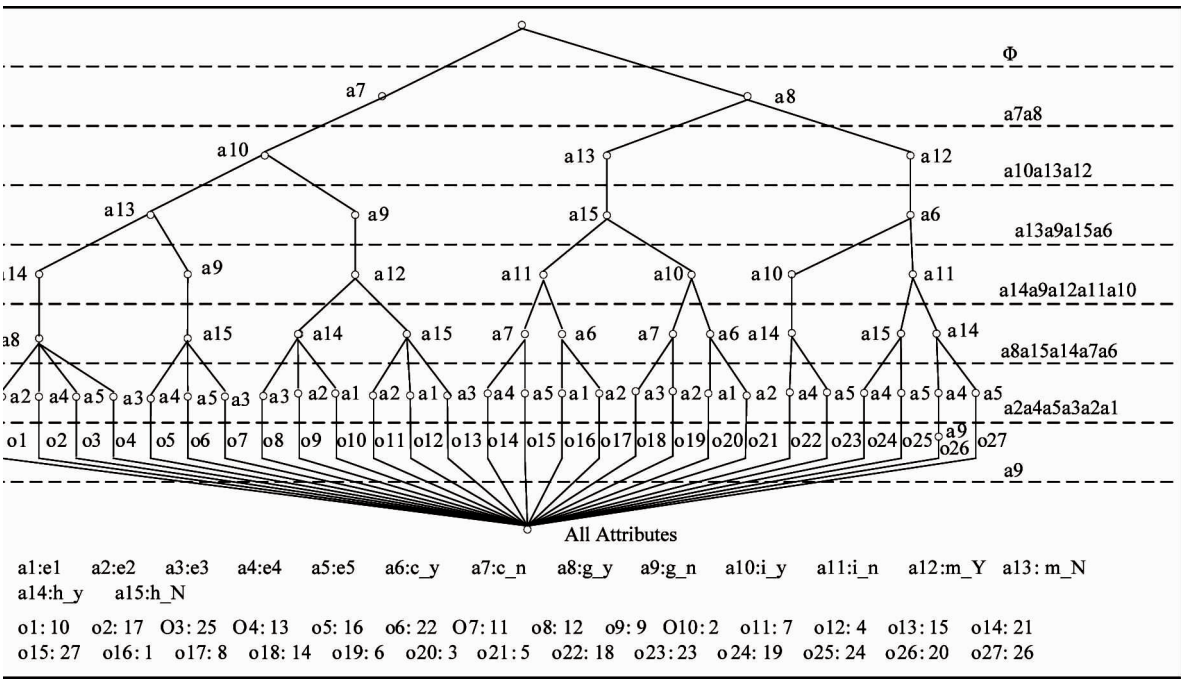

图 3　去掉决策属性后的偏序结构图

(2)如果条件属性 a7,a10,a13,a15 同时成立,那么可以得出状况 o5 = 16,o6 = 22,o7 = 11,也就是说,当白细胞图片中胞浆区域 r 参数均方差值较大,并且,胞核区域 R 参数平均值、细胞核面积占分比和核浆间 S 参数平均值对比均较小时,可以判定此白细胞可能为杆状核中性粒细胞。同时,该细胞细胞核非圆形,且胞浆区域色调 H 平均值较小。如实例 11、16、22。

(3)如果条件属性 a7,a10,a12 同时成立,那么可以得出状况 o8 = 12,o9 = 9,o10 = 2,o11 = 7,o12 = 4,o13 = 15,也就是说,当白细胞图片中胞浆区域 r 参数均方差值和核浆间 S 参数平均值对比均较大,并且,细胞核面积占分比较小时,不论胞核区域 R 参数平均值大小,可以判定此白细胞可能为单核细胞,同时,该细胞细胞核为圆形或近圆形,且胞浆区域色调 H 平均值较小。如实例 2、4、7、9、12、15。

(4) 如果条件属性 a8,a13,a15,a11 同时成立,那么可以得出状况 o14 = 21,o15 = 27,o16 = 1,o17 = 8,也就是说,当白细胞图片中胞浆区域 r 参数均方差值、核浆间 S 参数平均值对比和胞核区域 R 参数平均值均较小,不论胞浆区域色调 H 平均值大小,可以判定此白细胞可能为淋巴细胞,同时,该细胞细胞核面积占分比较大、胞核为近圆形。如实例 1、8、21、27。

(5)如果条件属性 a8,a10 同时成立的话,那么由此就可以得出状况 o18 = 14,o19 = 6,o20 = 3,o21 = 5,o22 = 18,o23 = 23,也就是说,当白细胞图片中胞浆区域色调 H 平均值和胞浆区域 r 参数均方差值均较大,不论胞核区域 R 参数平均值大小,可以判定此白细胞可能为嗜酸性粒细胞,同时,该细胞细胞核面积占分比较大、核浆间 S 参数平均值对比较大,细胞核不规则。如实例 3、5、6、14、18、23。

(6)如果条件属性 a8,a12,a11 同时成立,那么可以得出状况 o24 = 19,o25 = 24,o26 = 20,o27 = 26,也就是说,当白细胞图片中胞浆区域色调 H 平均值较大,同时胞浆区域 r 参数均方差值和胞核区域 R 参数平均值较小时,可以判定此白细胞可能为嗜碱性粒细胞,同时,该细胞细胞核面积占分比较大、核浆间 S 参数平均值对比较大。如实例 19、20、24、26。

3.2 对比实验

为了验证有效性,本文依据以上分类规则建立了二分树的白细胞图像分类器,选取了另外三种分类方法,对 2.1 节中 B 组实验数据做了对比实验。文献[11]中最小错误率的贝叶斯判别法属于参数识别法,它采用了 15 维的多维正态分布假设模型对白细胞图象进行分类;文献[12]选取并测定 22 个特征值后,根据已确定的血细胞分类的决策规则,采用统计分类的专家系统产生式规则来识别细胞;文献[13]采用了分步选择的策略,筛选出 54 个最优特征参数组合,进行了 BP 神经网络分类器设计。4 种方法在留一法验证下测试分类精度,其结果见表 6。

表 6 对比实验结果

方法来源	方法描述	特征数量	平均分类正确率(%)
文献[11]	基于最小错误率贝叶斯判别分类器	15	88.97
文献[12]	基于统计规则库匹配的统计决策表分类器	22	93.00
文献[13]	基于 BP 网络的分类器	54	90.01
本文	基于优选特征属性偏序结构的二分树分类器	6	94.04

其中,文献[11]方法在包含聚类情况的节点上性能较差,仅取得了88.97% 的平均分类正确率;文献[13]方法在 BP 网络每一个权值调整中容易偏差,取得了90.01% 的平均分类正确率;文献[12]方法中规则库中规则条件过于严格,取得了相对较高的 93.00% 的平均分类正确率;本文方法取得了 94.04% 的最好平均分类正确率,高于其它三种方法,主要因为本文基于优选特征属性偏序结构分析的白细胞图像分类规则清晰简单,便于建立一种标准二分树分类器,它把一个复杂的白细胞六分类问题转化为 4 层 5 个二分类问题来解决,根据各种属性特征值在属性多层次结构空间内将模式逐步进行由粗到细的分类,体现了人类根据各种知识进行推理的思维过程。

4 结 论

本文基于形式概念分析和属性偏序结构理论,提出了一种白细胞图像分类规则发现方法。该方法对人体外周血正常白细胞图像优选特征值进行了离散化分析,根据分层类坐标矩阵原理,针对白细胞图像数据集构建了优化后的形式背景和属性偏序结构,然后通过分析该属性偏序结构,提取了 6 类白细胞的相应 6 条分类规则,以此应用决策树原理建立了二分树分类器。6 条分类规则清晰简单,建立决策树分类器方便可行。在实际白细胞图像数据集测试实验中,该方法取得了 94.04% 的平均分类精度,高于其它经典分类方法,证明了基于优选特征属性偏序结构分析获取的白细胞图像分类规则的可用性和有效性,也证明其能够有效解决实际白细胞图像混合属性数据分类问题。为随后要建立的基于属性偏序结构关系的白细胞图像分类方法奠定了理论基础。该方法还可以推广应用于其它领域的计算机图像自动识别问题。

参考文献

[1] 丁报春. 生理学. 北京:北京大学医学出版社,2009. 32-35

[2] Khashman A. Investigation of different neural models for blood cell type identification. *Neural Computing & Applications*, 2012,(21):1-7

[3] 肖玮,陈性元,包义保. 基于多级关联信号树的高效可重构网包分类方法研究. 高技术通讯,2014,24(9):928-934

[4] Slavakis K, Giannakis G, Mateos G. Modeling and optimization for big data analytics:(statistical) learning tools for our era of data deluge. *IEEE Signal Process Mag*, 2014, 31: 18-31

[5] Ganter B,Wille R. Formal Concept Analysis:Mathematical Foundantions. New York: Springer-Verlag, 1999

[6] Hong W X,Yu J P,Cai F et al. A new method of attribute reduction for decision formal context. *ICIC Express Letters Part B:Applications*, 2012,3(5): 1061-1068

[7] 郝连旺,洪文学,李婷. 基于属性层次关系的白细胞图像类间特异特征选取方法研究. 生物医学工程学杂

志,2014,31(6):1202-1206

[8] Wille R. Restructuring lattice theory: An approach based on hierarchies of concepts. In Ordered Sets Dordeche: Reidel Publishing Company, 1982. 445-470

[9] Martin B, Eklund P. From Concepts to Concept Lattice: A Border Algorithm for Making Covers Explicit. Germany:Springer Verlag, 2008:92-105

[10] 郝连旺,洪文学. 基于多颜色空间特征融合的彩色白细胞图像识别. 生物医学工程学杂志, 2013, 30(5): 909-913

[11] 张勇. 彩色白细胞显微图像分析与识别:[博士学位论文]. 西安:西安交通大学机械工程学院,1999. 102-104

[12] 汤学民,林学闾,何林. 白细胞图像自动识别系统的研究. 生物医学工程学杂志, 2007, 24(6):1250-1255

[13] 周颖颖. 彩色白细胞图像的特征选择与分类识别:[硕士学位论文]. 南京:东南大学生物科学与医学工程学院, 2006. 53-55

Classification rule discovering for leucocyte images based on analysis of preselected features' attribute partial-ordered structure

Hao Lianwang[*,**], Hong Wenxue[*], Wei Kun[**]

([*]College of Electrical Engineering Yanshan University, Qinhuangdao 066004)

([**]The First Hospital of Qinhuangdao, Qinhuangdao 066000)

Abstract

A novel model for finding leucocyte image classification rules was proposed based on formal concept analysis and the theory of attribute partial-ordered structure, and then, an efficient method for leucocyte image classification was established. According to the method, the optimized leucocyte attributes were discretized based on the analysis of the obtained experimental measurement results, and the optimized formal concept and the attribute partial-ordered structure were established based on the hierarchical class coordinate matrix for actual leucocyte images dataset. Then, based on the optimized attribute partial-ordered structure, six classification rules for six kinds of leukocytes were extracted. Finally, a binary decision tree classifier was established according to the classification rules, and an average classification accuracy of 94.04% was achieved. The classification accuracy is significantly higher than the other 3 kinds of classical algorithms, showing the better usability, simplicity and effectiveness of the classification rules obtained based on analysis of the attribute partial-ordered structure.

Key words: formal concept analysis, attribute partial-ordered structure, leucocyte image, classification rule