卫星云图感兴趣区域自动提取方法研究

来 旭[1],李国辉[1],赵福华[2]
LAI Xu[1],LI Guo-hui[1],ZHAO Fu-hua[2]

1.国防科技大学 信息系统与管理学院 系统工程系,长沙 410073
2.湖南省气候中心,长沙 410007
1.College of Information System and Management,National University of Defense Technology,Changsha 410073,China
2.Climate Center of Hunan Province,Changsha 410007,China

LAI Xu,LI Guo-hui,ZHAO Fu-hua.Automatic ROI extraction from satellite cloud image based on gray level histogram. Computer Engineering and Applications,2009,45(30):230-233.

Abstract: All kinds of clouds are the interests for people in the satellite cloud images.Considering the complexity of cloud image,this paper proposes a weighted FCM method based on the gray level histogram for cloud image segmentation.The ROI is finally acquired after finishing post-process.A general clustering method needs the class number specified by people.This paper implements automatically confirming the optimized class number based on the clustering validity index.Cloud image segmentation is the important step during the process of ROI extraction.This paper combines the weighted idea with FCM to make the clustering more scientific.On the other hand,the clustering object is transformed from pixel to gray level histogram.The modified algorithm executes more efficiently.The experiment result demonstrates the ROI extraction method can classify the image content into six regions of interest:Land,water,stratus,middle cloud,cirrus and cumulonimbus.The results are consistent with the objective facts.
Key words: Region Of Interest(ROI);clustering validity index;threshold partition;Fuzzy Clustering Method(FCM)

摘 要:卫星云图中人们感兴趣的区域(ROI)往往是各类云团,针对卫星云图内容的复杂性,利用直方图模糊加权 C 均值聚类方法实现云图的图像分割,对分割结果进行后处理,最终获取云图内的感兴趣区域。常规聚类方法需要人工指定类个数,影响了 ROI 提取过程的自动化程度。引入修正聚类评价指标,基于该指标实现最佳类别个数的自动确定。云图分割是感兴趣区域提取过程的关键,采用的直方图模糊加权 C 均值聚类方法在原有算法基础上,引入样本权重概念,使得聚类过程更为合理;同时将聚类对象由原始像素转换为灰度直方图,提高了聚类过程执行效率。实验结果表明设计的感兴趣区域提取方法能较为准确地分辨出陆地、水体、低云、中云、卷云、对流云六类区域,提取结果与客观实际一致。
关键词:感兴趣区域(ROI);聚类有效性指标;阈值分割;C 均值聚类
DOI:10.3778/j.issn.1002-8331.2009.30.068 **文章编号:**1002-8331(2009)30-0230-04 **文献标识码:**A **中图分类号:**TP391

1 引言

卫星云图是一种反映大气云团分布情况的遥感图像,目前主要使用的是红外云图。云团在云图中一般呈大范围弥散状分布,而且不同的云团存在叠加的情况,这些特点为云图的判读和理解带来了困难。引入感兴趣区域提取思想,从云图中划分出典型区域, 这些区域一方面可为云团训练样本库的建立服务,另一方面有助于减轻人工判读时的工作负担。图像分割是 ROI 提取的关键技术,常用的图像分割采用阈值化思想,基于类别可分性准则的 Otsu 方法[1]是这类思想的代表。随着聚类研究的深入,提出了许多基于聚类准则的图像阈值化方法[2]。对于内容相对单一的图像,常规算法就可以确定阈值实现目标与背景区的分割,但卫星云图内容复杂,且具有模糊特性,基于这些特点, 采用基于模糊 C 均值聚类思想的多阈值图像分割算法实现云图区域的划分。基于卫星云图的灰度直方图确定阈值可以较大地提高图像分割的效率,结合云图内容的特点,提出两阶段云图感兴趣区域提取方法:第一阶段首先对云图进行分块化处理,基于每个图像快进行感兴趣区域提取;第二阶段对全部图像块中提取的感兴趣区域进行聚类,完成感兴趣区域间的融合,最终获得面向整幅云图的感兴趣区域。

2 聚类有效性分析指标

聚类方法中的分区聚类法理论基础扎实, 实现非常简便,获得了广泛的使用, 但是分区聚类法需要预先设定类别个数,目前一般采用人工方式确定。这种方式需要很强的领域知识,同时还要掌握数据的全局分布情况才能较合理的确定类别个数。人工方式限制了聚类方法的自动化实现。因此,实现感兴趣

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60473116)。
作者简介:来旭(1979-),男,博士研究生,主研方向:图像智能处理、多媒体数据挖掘;李国辉,教授,博士生导师;赵福华,高级工程师。
收稿日期:2008-06-10 **修回日期:**2008-09-04

区域的自动化提取需首先解决类别个数确定的自动化问题。

聚类有效性指标是一种衡量聚类结果合理性的指标,假设聚类类别数 C 的范围为[1,$C_{\max}$],最佳类别数 C_{opt} 是令聚类指标达到某个极值的点,基于聚类有效性指标,最佳聚类类别个数确定问题转化为搜索有效性指标极值问题。

常用的聚类有效性指标有 3 类:(1)基于数据集模糊划分的有效性指标,如分离度、划分熵[3]等。(2)基于数据集几何结构的方法,如划分系数[4]、分离系数[5]、Xie-Beni 指标[6]和基于图论的有效性函数等。(3)基于数据集统计信息的方法。三种方法分别从不同角度评估聚类的效果,但都存在一定的不足,例如第 1 类方法与数据集的结构特征缺乏直接的联系;第 2 类方法表述及运算复杂;第 3 类方法需要根据数据集做出合理的统计假设,缺乏实际应用中的可操作性。

根据卫星云图内容的特点,在感兴趣区域提取过程中采用 C 均值模糊聚类方法,在模糊划分有效性指标的基础上,结合考虑数据集的几何结构信息,引入划分模糊度(partition fuzzy degree)[7]这一概念。将原有的划分熵与划分模糊度相结合,定义一种修正的划分模糊度作为聚类有效性指标,以此确定最佳聚类个数。

定义 1　对于给定的类别数 c 和模糊划分矩阵 $\boldsymbol{U}$,数据集 $\boldsymbol{X}$ 的模糊划分熵定义为:

$$H(\boldsymbol{U},c)=-\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}\cdot\ln(u_{ij}) \tag{1}$$

约定当 $u_{ij}=0$ 时,有 $u_{ij}\cdot\ln(u_{ij})=0$。划分熵是一种衡量聚类结果模糊程度的标准,根据定义,当分类越分明时,$H(\boldsymbol{U},c)$的值越小;当分类越模糊时,$H(\boldsymbol{U},c)$的值就越接近于 ln$c$。

定义 2　对于给定的类别数 c 和模糊划分矩阵 $\boldsymbol{U}$,数据集 $\boldsymbol{X}$ 的划分模糊度定义为:

$$PF(\boldsymbol{U},c)=\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\left|u_{ij}-(u_{ij})_H\right| \tag{2}$$

式中$(u_{ij})_H=\begin{cases}1,u_{ij}=\max\limits_{1\leqslant k\leqslant c}\{u_{ij}\}\\0\end{cases}$即$(u_{ij})_H$对应于数据集的硬划分指标函数。划分模糊度也是一种判定分类模糊性程度的标准,当划分结果越分明时,$PF(\boldsymbol{U},c)$的值越小;而分类越模糊时,$PF(\boldsymbol{U},c)$的值就越接近 2-2/c。

由于 $H(\boldsymbol{U},c)$和 $PF(\boldsymbol{U},c)$的值都随类别数 c 的增加而呈递增趋势,这种递增趋势将影响对 $H(\boldsymbol{U},c)$和 $PF(\boldsymbol{U},c)$的全局或局部极值点的检测,而这些极值点可能对应到最优类别数,根据两种指标增长速率的差异,引入了修正划分模糊度,基于该指标的曲线将更易于确定最佳类别数。

定义 3　对于给定的类别数 c、模糊划分矩阵 $\boldsymbol{U}$,数据集 $\boldsymbol{X}$ 的修正划分模糊度定义为:

$$MPF(\boldsymbol{U},c)=\frac{PF(\boldsymbol{U},c)}{H(\boldsymbol{U},c)} \tag{3}$$

约定当 $\boldsymbol{U}$ 是硬划分时,$MPF(\boldsymbol{U},c)=0$。基于式(3)定义的修正划分模糊度,最优聚类类别个数判定准则:

$$MPF(\boldsymbol{U}^*,c^*)=\min_{c}\left(\min_{\Omega_c}MPF(\boldsymbol{U},c)\right) \tag{4}$$

式中:Ω_c 为不同的类别数 c 对应的"最优"分类矩阵有限集,($\boldsymbol{U}^*$,c^*)为最优的聚类结果,c^* 为最佳聚类类别数。

3　基于直方图的加权模糊 C 均值聚类算法

图像分割是指把图像分解成各具特性区域的过程,它是感兴趣提取流程中关键的环节。随着模糊理论的引入,聚类技术获得了新的发展,其中最有代表性的算法是模糊 C 均值聚类算法。研究对象为卫星云图,由于云图内容复杂多变,具有很强的模糊与不确定性,硬性的聚类并不适合云图的分割,通过隶属度计算对象属于各类的不确定程度,能更好地符合云图内容的特点。

令 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$表示一组具有 n 个样本的数据集,其中 $\boldsymbol{x}_i=[x_{i1},x_{i2},\cdots,x_{ip}]^{\mathrm{T}}$,表示第 i 个样本的 p 个特征值。数据集 $\boldsymbol{X}$ 的模糊 C 均值聚类问题可以表述为下面的数学规划问题:

$$\min J_m(\boldsymbol{U},\boldsymbol{V})=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}d_{ij}^{2} \tag{5}$$

式中:c 为聚类类别数(1<c<n),m 为模糊加权指数(m>1),通常取 m=2,$d_{ij}=\|x_j-v_i\|$为样本 x_j 到聚类中心 $v_i\subset R^p$ (1≤i≤c)的欧式距离,u_{ij} 为第 j 个样本属于第 i 个聚类中心的隶属度,此外,$\boldsymbol{U}=\lfloor u_{ij}\rfloor$是一个 $c\times n$ 阶矩阵,$\boldsymbol{V}=[v_1,v_2,\cdots,v_c]$是一个 $p\times c$ 阶聚类中心矩阵。Bezedk[8]给出了求解上述数学规划问题的交替优化算法,即模糊 C 均值聚类算法(FCM),算法核心是利用拉格朗日乘子法推导出针对公式(5)的优化迭代公式:

$$u_{ik}=\frac{1}{\sum\limits_{j=1}^{c}\left(d_{ik}/d_{jk}\right)^{2(m-1)}} \tag{6}$$

$$V_i=\frac{\sum\limits_{k=1}^{n}u_{ik}^{m}X_k}{\sum\limits_{k=1}^{n}u_{ik}^{m}} \tag{7}$$

根据式(6)、(7)不断调整类中心和隶属度矩阵,最终使目标函数达到最优。

不同的样本数据对 FCM 聚类过程的影响效果是不同的,常规 FCM 算法中并没有强调这种差异,针对不同的样本赋予不同的权值,引入样本加权思想,使聚类过程更加合理。引入权重后公式(5)的形式变为:

$$\min J_m(\boldsymbol{U},\boldsymbol{V})=\sum_{i=1}^{c}\sum_{j=1}^{n}\omega_i u_{ij}^{m}d_{ij}^{2} \tag{8}$$

其中,ω_i 为每个样本的权重系数,且应满足条件 $\sum\limits_{i=1}^{n}\omega_i=1$。再次利用拉式乘子法,推导出引入权信息的迭代优化公式,隶属度公式不变,类中心公式如式(9)所示。

$$V_i=\frac{\sum\limits_{k=1}^{n}\omega_i u_{ik}^{m}X_k}{\sum\limits_{k=1}^{n}\omega_i u_{ik}^{m}} \tag{9}$$

权系数主要作用在聚类中心的调整,当 $\omega_i=1/n$ 时,即各样本对分类影响一致时,加权 FCM 算法退化为常规 FCM 算法。

加权 FCM 算法单次迭代的计算集中在 V_i 和 U_{ik} 的乘、除法运算。针对像素聚类每次迭代计算 V_i 至少需要 $c\times n$ 次乘法和 c 次除法,计算 U_{ik} 至少需要 $n\times c\times(n-1)$次除法。精简数据集,减少参与聚类的对象个数可以有效提高聚类算法效率,对

于 1 024×768 的图像,如果基于像素聚类,n=786 432;如果基于直方图聚类,n=256,且聚类样本数 n 不会伴随图像的增大而改变。基于这一思想,选取云图灰度直方图作为聚类对象。直方图的峰谷反映图像灰度的分布情况,峰是像素灰度分布集中区域,谷是像素分布的稀疏地带。每一个峰都是紧密集中在某个中心值周围的一类像素;类(峰)与类(峰)的分界值存在于谷中[9]。该文利用灰度及其出现的频数作为待聚类的对象,与加权 FCM 聚类算法结合成为基于直方图的加权 FCM 聚类算法。直方图中某级灰度频数的计算如公式(10):

$$h_i=\frac{n(i)}{M\times N},i=0,1,\cdots,L-1 \tag{10}$$

其中 $n(i)$表示灰度为 i 的像素个数,L 为总的灰阶数,h_i 满足条件 $\sum_{i=0}^{L-1}h_i=1$,可直接将 h_i 作为权重。

综上所述,直方图加权 FCM 算法作为云图分割算法有三方面优势:(1)模糊机制适于云图这类内容复杂,对象间区别无明显界线的情况。(2)样本加权使聚类过程考虑的因素更加全面,聚类过程更为合理。(3)划分对象由图像像素转变为直方图减少了运算量,提高了聚类效率。

4　云图感兴趣区域自动提取

云图感兴趣区域自动提取主要以图像分割技术为基础,前文阐述了与云图分割相关的两个关键问题的解决方法,一是针对基于图像内容的了最佳聚类类别数的确定方法;另一个是基于聚类技术的图像分割算法的实现。结合卫星云图内容的实际特点,该文设计的云图感兴趣提取算法包括下述步骤:

步骤 1　云图区块化处理。

卫星云图由于覆盖面积大,图像包含的信息非常复杂。直接针对原图进行操作,不但计算量较大,而且增加了算法执行的难度,因此,先对原始云图进行分块处理,算法分析的对象先集中在内容相对较为简单的区域内,获得了分块处理的结果后,再进行全局处理。

步骤 2　基于分块区域的最佳类别数的确定。

根据修正划分模糊度这一指标,在[1,$C_{\max}$]区间内找出分块区域对应的最佳类别数。$C_{\max}$ 表示最大类别数,文献[10]指出 $C_{\max}\leqslant 2\ln(n)$,其中 n 为样本个数。采用基于直方图的聚类方法,n 的值为云图灰度等级数。确定了类别数的区间后,绘制出“类别数—修正模糊度”曲线,由曲线的极值点分布确定最佳类别数。

步骤 3　针对分块区域的图像分割的实现。

根据步骤 2 确定的分块区域对应的最佳聚类数,提取分块区域的直方图,利用加权 FCM 算法可获得不重叠的区域,这些划分出的对象可称为潜在的原子感兴趣区域。

步骤 4　对聚类得到的区域进行腐蚀、膨胀处理,获得图像块内的原子感兴趣区域。

分块利用聚类得到的分割区域往往存在着明显的孔洞效性,主要是因为在某些较大的区域中间零散分布着一些面积很小存在灰度差异的区域,直方图加权 FCM 算法会把这些小区域同包含它们的大区域划分为不同的类别,但是专家在人工判读云图时,往往会把它们划归为同一类云团,这些小的差异可以看成是纹理信息。采用腐蚀和膨胀等图像处理技术对分割结果进行处理,有效解决了孔洞问题。

步骤 5　对所有原子感兴趣区域进行再次聚类,获得全局感兴趣区域。

云团在云图中表现为大范围的连续分布,经过步骤 1 的分块处理,往往会影响云团的完整性;同时步骤 3 的聚类只是考虑本区块的情况,获得的原子感兴趣区域只是局部云团分布的反应。因此为了获得反应云团全局信息的感性趣区域,对所有的原子感兴趣区域在特征空间内,利用 FCM 聚类技术进行二次聚类,将相似的原子区域划分到同一类别中。进行二次聚类前,应完成的准备工作包括:(1)确定最佳类别数,可利用前面的修正模糊度指标实现。(2)确定从原子感兴趣区域内提取的特征,作为 FCM 算法处理的对象。红外云图最重要的信息是灰度,人工判读云图就是根据灰度的明暗程度区分云团的种类。该文的目标是得到云图内的感兴趣区域,为后续训练样本的获取服务,并不是完成精确的云团种类的区分,时效性是感兴趣区域提取算法更加关注的指标。同多维的特征向量相比,单个特征反映的信息有限,但是多维特征向量必然会加大算法处理的难度,延长处理时间,无法满足感兴趣区域提取的时效性。因此确定将原子感兴趣区域内的灰度均值作为二次聚类处理的特征值。

5　实验结果与分析

实验采用风云 2 号卫星(FY2C)2007 年 1 月 14 日 14 时08 分拍摄的红外 1 卫星云图作为研究对象,验证该文感兴趣区域提取算法的有效性。图 1 为该云图的聚类评价指标曲线,横轴为类别数,纵轴为聚类评价指标。其中绿线对应的是划分熵,蓝线表示划分模糊度,红线表示修正模糊度。从红线的走势可以发现从类别数为 6 开始,修正模糊度的下降趋势明显趋缓,结合气象学领域知识,可确定该云图的最佳类别数 C_{opt}=6。

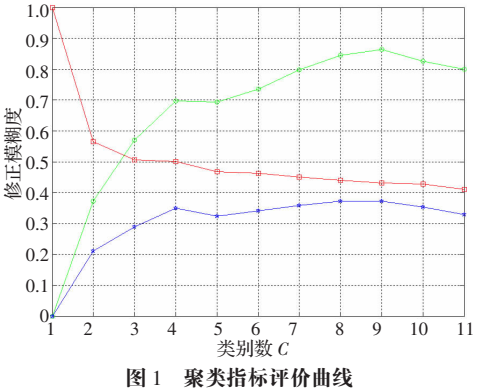
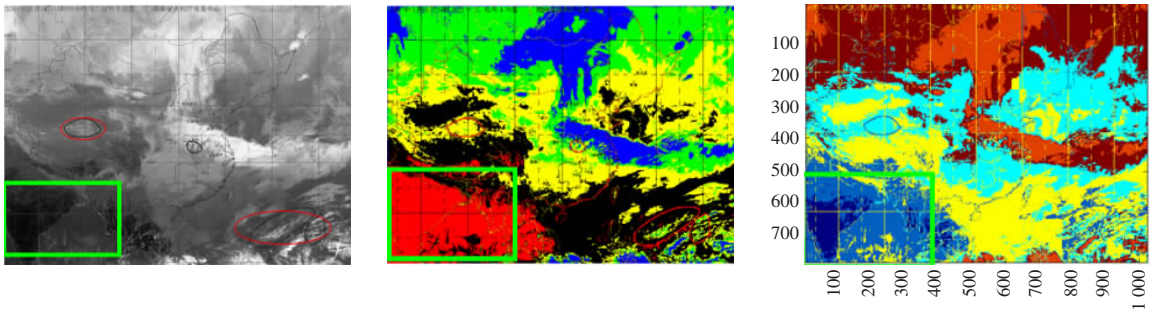


图 1　聚类指标评价曲线

上述计算是基于整幅图像进行处理的,在确定聚类个数后,利用直方图加权 FCM 聚类算法得到云图的分割结果,如图 2(b)所示。该文的第 4 章内容中强调了分区处理的思想,先对原图进行分块处理。然后再按照感兴区域提取算法流程处理,获得的分割结果如图 2(c)所示。通过比较,基于区块处理的结果优于基于整幅云图的处理结果,例如,在对陆地和水体的区分时,前者就可以有效地进行区分,后者就将它们划规为一类(图 2 中绿色方框标记的)。最终获得 6 个全局感兴趣区域如图 3 所示。图 2 中红色椭圆形区域是由专家手工标注的具有典型



(a)原图　　(b)整幅图应用直方图分割法　　(c)分块用灰度直方图分割结果

图 2　直方图加权 FCM 算法聚类效果

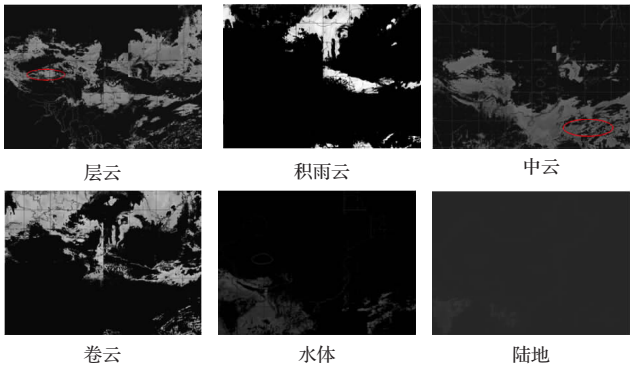


图 3　二次聚类后提取的全局感兴趣区域

云类特征的感兴趣区域,从图 3 可以发现,该文算法自动提取的全局感兴趣区域与专家标定的感兴趣区域保持一致,证明了算法的有效性。

参考文献:

[1] Ostu N.A threshold selection method from gray level histogram[J]. IEEE Tran SMC,1979,9(1):62-66.

[2] 韩思奇,王蕾.图像分割的阈值法综述[J].系统工程与电子技术,2002,24(2):92-102.

[3] Bezdek J C.Pattern recognition with fuzzy objective function algorithms[M].New York:Plenum Press,1981.

[4] Dunn J C.Wellseparated clusters and the optimal fuzzy partions[J].J Cybernet,1974,4:95-104.

[5] Gunderson R.Application fuzzy Isodata algorithm startracker printing system[C]//Proc 7th Triennial World IFAC Congr,Helsinki,Finland,1978:1319-1323.

[6] Xie X L,Beni G.A validity measure for fuzzy clustering[J].IEEE PAMI,1991,13:841-847.

[7] 李双虎,张风海.一个新的聚类有效性分析指标[J].计算机工程与设计,2007,28(8):1773-1774.

[8] Bezdek J C.Pattern recognition with fuzzy objective function algorithm[M].New York:[s.n.],1981:309-321.

[9] 王璐,蔡自兴.改进的快速 FCM 算法[J].小型微型计算机系统,2006,26(10):1775-1777.

[10] 范九伦,裴继红,谢维信.基于可能性分布的聚类有效性[J].电子学报,1998,26(1):127-130.

(上接 226 页)

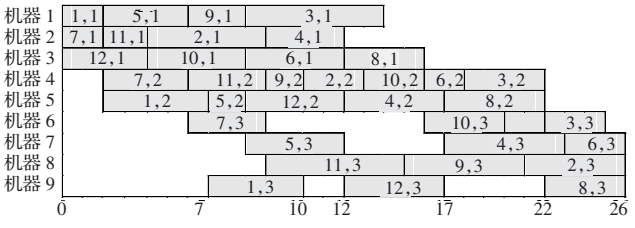


图 1　最优调度的甘特图

中所求得的结果,说明该文中的算法在收敛速度和稳定性方面的效果是比较好的。

4　结束语

从实例仿真可以看出,提出的遗传算法基因编码和解码方法简洁,能清楚地反映出调度方案关系;收敛速度快,能有效地解决柔性 Flow-shop 调度问题。通过比较,解的质量是比较满意的,有一定广泛实际应用的良好前景。

参考文献:

[1] Johnson S M.Optimal two-and three-stage production schedules with set-up times included[J].Naval Research Logistics Quarterly,1954,1(1):61-68.

[2] Hejazi S R,Saghafian S.Flowshop-scheduling problems with makespan criterion:A review[J].International Journal of Production Research,2005,43(14):2895-2929.

[3] Murata T,Ishibuchi H,Tanaka H.Multi-objective genetic algorithm and its applications to flow shop scheduling[J].Computers and Industrial Engineering,1996,30(4):957-968.

[4] Reeves C.A genetic algorithm for flow shop sequencing[J].Computers and Operations Research,1995,22(1):5-13.

[5] Wang Hong.Flexible flow shop scheduling:Optimum,heuristics and artificial intelligence solutions[J].Expert Systems,2005,22(2):78-85.

[6] Pezzella F,Morganti G,Ciaschetti G.A genetic algorithm for the flexible job-shop scheduling problem[J].Computers and Operations Research,2008,35(10):3203-3212.

[7] 王万良,姚明海,吴云高,等.基于遗传算法的混合 Flow-shop 调度方法[J].系统仿真学报,2002,14(7):863-865.

[8] 王万良,吴启迪.生产调度智能算法及其应用[M].北京:科学出版社,2007:77-83.

[9] 王凌.车间调度及其遗传算法[M].北京:清华大学出版社,2003.

[10] 王小平,曹立明.遗传算法—理论、应用与软件实现[M].西安:西安交通大学出版社,2002:28-38.

[11] 周明,孙树栋.遗传算法原理及应用[M].北京:国防工业出版社,1999:146-254.

[12] 雷英杰,张善文,李续武,等.Matlab 遗传算法工具箱及应用[M].西安:西安电子科技大学出版社,2005.