遥感影像均值漂移分割算法的并行化实现

沈占锋，骆剑承，吴 炜，胡晓东

（中国科学院 遥感应用研究所，北京 100101，shenzf@irsa.ac.cn）

摘 要：本文采用遥感影像数据的均值漂移算法并行化方法来解决均值漂移不能处理过大影像、处理速度过慢等问题，通过分析均值漂移算法的原理，提出了一种新的数据“缓冲区”式分块方法，并进而分别对不同的数据块进行并行分割，从而消除了该算法对数据量的限制，有效避免计算机在处理过大影像时而产生的内存不足问题，并从效率角度对算法进行了改进.

关键词：多尺度分割；均值漂移；并行化；数据划分

中图分类号：TP393109 **文献标志码**：A **文章编号**：0367－6234(2010)05－0811－05

Implementation of parallelization of mean-shift algorithm for multi-scale segmentation of remote sensing images

SHEN Zhan-feng, LUO Jian-cheng, WU Wei, HU Xiao-dong

(Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing 100101, China, shenzf@irsa.ac.cn)

Abstract: Aimed at the problem that the mean shift algorithm sometimes can not compute large volumes of image data, or the data-computing speed may be very slow, by analyzing the principle of the mean-shift image segmentation, this paper presents an image-segmentation idea by processing the data in parallel computing environment, i.e. Data Partition Method with Buffer (DPMB), then we can compute different parts of data separately. By this method, we can avoid the limit of data amounts and memory errors problem of computer, which improves the efficiency of data segmentation.

Key words: multi-scale segmentation; mean shift; parallel computing; data partition

在面向对象的高分辨率遥感影像信息提取与分析过程中，影像分割是实现影像的对象化划分的一个重要步骤，其实现的精度、速度将直接影响着影像处理与分析的精度与效率[1]. 对于多光谱的遥感影像分割任务来说，目前应用较多的分割算法包括 Definiens 推出的多分辨率影像分割算法、分水岭分割算法以及均值漂移分割算法等. 其中多分辨率影像分割算法属 Definiens 专利，其实现方法并未完全公开，均值漂移是一种有效的统计迭代算法，并且证明具有较好的算法收敛性，已广泛应用于聚类分析、跟踪、图像分割、图像平滑、滤波、图像边缘提取和信息融合等方面[2-3]. 但是，这些算法都存在着一个统一的问题，那就是随着影像数据量的加大，算法都需要将相应的数据一次性调入内存进行分析与处理，这就使得当影像数据量过大时，算法有可能出现异常，而且分割的速度会骤然降低，给海量数据（例如需要处理整景的融合数据）的快速处理带来了困难，因此需要对算法进行改进以适应这种情况.

针对大数据量的均值漂移分割算法改进可以体现在两个方面，一种方法是从算法本身进行改进，如选择或构造不同的核函数[4]、构建不同的目标模型[5]、进行算法的优化[6]等，另一种方法则是从算法的计算机实现角度来进行改进，如进行算法的并行化改进等，本文正是从这一角度对

收稿日期：2008－12－12.

基金项目：国家高技术研究发展计划资助项目(2009AA12Z123，2009AA12Z121)；国家自然科学基金项目(40971228，40871203)；国家科技支撑计划重大项目(2006BAJ02A01，2006BAJ14B08).

作者简介：沈占锋(1977—)，男，博士，副研究员；
骆剑承(1970—)，男，研究员，博士生导师.

高分辨率遥感影像的均值漂移算法进行研究与实现的. 通过对算法进行并行化改进,其优点一方面可以提高影像数据的处理速度(其实现方法可以在一台计算机上进行多线程影像计算,也可以采用 Cluster 等方法将多台处理器在局域网内部连接起来进行并行协同计算),另一方面还可以提高海量数据计算的能力,例如在面向农业土地覆盖类型调查及林业林相图生成过程中,很多时候需要将一景的 SPOT5 融合影像进行整体图像分割和边界提取,而如此大的数据量则是一般计算机的内存无法容忍的,采用这种数据分块方法直接克服了算法对数据量及内存开辟的限制,具有更广的应用范围.

本文通过分析研究均值漂移算法的原理及实现过程,对算法密集段进行并行化算法实现,在数据划分过程中,通过采用数据"缓冲区"式划分方法消除了并行数据块之间明显的"分隔线",使得并行化算法处理结果更趋于合理,并可有效提高算法的处理效率.

1 均值漂移遥感影像分割算法

1.1 均值漂移影像分割方法

1.1.1. 均值漂移算法原理

均值漂移是一种特征空间中的自动聚类算法,是一种非参数估计密度函数的方法,对先验知识要求少,完全依靠训练数据进行估计,可以用于任意形状密度函数的估计,对于不同结构的数据具有很好的适应性和稳健性,不需要事先确定类别数,能使特征空间中每一个点通过有效的统计迭代"漂移"到密度函数的局部极大值点[7].

设核函数 $\boldsymbol{K}_H(-)$ 如果满足一定的统计矩约束就是概率密度函数,可以用于非参数概率密度估计. 若样本集 $\{x_i\}_{i=1}^n$ 是依密度函数 $f(x)$ 经过 n 次独立抽样得到,则给出的密度函数估计为

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{K}_H(x - x_i). \tag{1}$$

其中核函数满足

$$\boldsymbol{K}_H(x) = |\boldsymbol{H}|^{-1/2}\boldsymbol{K}(\boldsymbol{H}^{-1/2}x).$$

在实际应用过程中,矩阵 $\boldsymbol{H}$ 的选择对结果有着直接影响. 为了减少计算的复杂性,往往选择对角阵

$$\boldsymbol{H} = \mathrm{diag}[h_1^2, h_2^2, \cdots, h_d^2],$$

或单位矩阵的比例阵

$$\boldsymbol{H} = h^2\boldsymbol{I}.$$

其中后者的优点是只需要指定一个大于零的带宽 h. 在这种情况下,确定核函数带宽后,(1) 式中的密度估计算子就可以转化成一种如下的更为常见的形式:

$$\hat{f}(x) = \frac{1}{nh_d}\sum_{i=1}^{n}\boldsymbol{K}\left(\frac{x - x_i}{h}\right).$$

核密度估计的质量由密度函数及其估计值来决定,即

$$\hat{f}(x) = \sum_i w_i\boldsymbol{H}(x - x_i).$$

其中,权重系数 w_i 满足约束条件 $\sum_i w_i = 1$. 若核 $\boldsymbol{K}_H(-)$ 是核 $\boldsymbol{K}(-)$ 的影子核,则均值漂移向量的定义为

$$m(x) - x = \frac{\sum_i w_i\boldsymbol{K}(x - x_i)x_i}{\sum_i w_i\boldsymbol{K}(x - x_i)} - x. \tag{2}$$

其中,$m(x)$ 为 x 处的样本均值. 数据点向样本均值移动的迭代过程 $x \leftarrow m(x)$,称为均值漂移算法. 迭代过程中 x 所经过的位置,即序列 $\{x, m(x), m(m(x)), \cdots\}$ 称为 x 的轨迹.

式(2) 定义的均值漂移向量正比于概率密度函数 $f(x)$ 在 x 处的梯度. 均值漂移的方向总是指向具有最大局部密度的地方,在密度函数极大值处,漂移量趋于零,$\nabla f(x) = 0$,所以均值漂移算法是一种自适应快速上升算法,它可以通过计算找到最大的局部密度在什么地方[8],并向其位置"漂移",这就是均值漂移算法的原理.

1.1.2 均值漂移分割算法实现流程

根据以上均值漂移算法的原理,对基于均值漂移的遥感影像分割算法进行了实现,实现流程如图 1 所示.

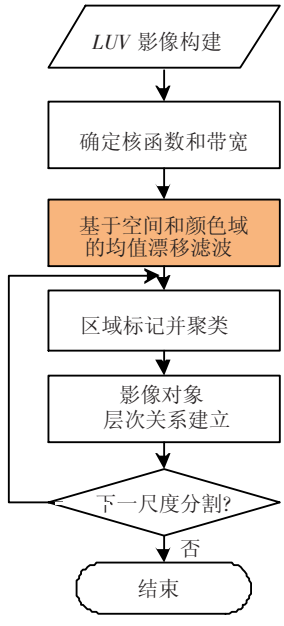


图 1 均值漂移算法的实现流程

均值漂移分割算法需要首先将色度空间转

换到 *LUV* 特征空间，进行核函数及相应参数的确定并进行均值漂移滤波，再进行影像的聚类过程. 假设彩色图像的特征空间为 L，则图像中不同颜色的物体对应特征空间上不同的聚类，彩色图像映射到特征空间 L 后，再结合像素在图像中的位置，即空间信息 (X,Y)，就能得到每个像素在 5 维特征空间中的值，即 (X,Y,L^*,U^*,V^*)，其中 L^* 表示图像的亮度，U^* 和 V^* 分别表示色差. 在此基础上采用聚类算法，就可以把空间和颜色欧氏距离相近的点归为一类，从而实现彩色图像的分割.

均值漂移滤波器在较好抑制图像噪声的同时，最大程度上保留了边缘或者其它结构特征，基于均值漂移的影像分割算法正是基于均值漂移滤波的结果进行的. 但是在均值漂移滤波运算过程中，因为包括了图像到高维空间的映射以及核中心位置的迭代运算，内存要求非常大，而且运算时间要比邻域均值滤波等其他滤波方法要长的多，因此，均值漂移滤波过程就构成了均值漂移分割算法的任务密集段，是算法并行化过程中需要首先考虑并实现的部分.

1.2　均值漂移算法的并行化分割思路

对于均值漂移影像分割算法来说，由于算法的任务密集段为影像的均值漂移滤波过程，因此在进行算法的并行化过程中，应该首先考虑此部分的并行算法的实现. 在均值滤波过程中，假设 x_i 与 $z_i, i = 1,2,\cdots,n$ 为 d – 维输入待滤波影像像元，对于每一个影像像元均有

1）初始化 $j = 1$ 以及 $y_{i,1} = x_i$；

2）根据公式(2) 计算 $y_{i,j+1}$，直到 $y = y_{i,c}$；

3）$z_i = (x_i^{\mathrm{s}}, y_{i,c}^{\mathrm{r}})$.

其中 s 与 r 分别代表特征向量的空间与范围，滤波数据 x_i^{s} 通过滤波过程将收敛于 $y_{i,c}^{\mathrm{r}}$.

根据上述均值漂移的实现过程可知，实际上，均值漂移滤波过程是对每个像元逐步计算并逼近其“漂移点”的过程，而这个过程将随着影像像元个数的增加而呈倍数的增加，也是直接影响算法速度的“瓶颈段”. 由于图 1 所示的基于空间和颜色域的均值漂移滤波部分是影像并行化过程需要首要考虑的步骤，其并行化实现过程是在主节点数据划分后将不同数据块分别在不同位置（可为不同线程、不同进程或其他方式处理）进行滤波，并在此基础上进行相应的对象特征计算与表达，并由主节点进行处理结果合并进行后续过程，图 1 的滤波过程经并行化改进的实现流程如图 2 所示.

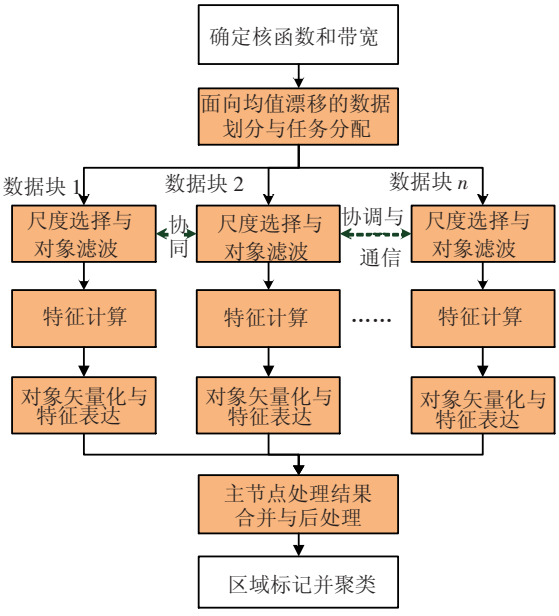


图 2　均值漂移算法的并行化实现思路

在图 2 所示的并行化过程中，数据分块划分策略将直接决定着数据在并行算法中的处理效果及效率[9-10]. 由于均值漂移算法需要对全局性的参数进行统计计算，因此不存在如参考文献[9]或文献[2]中所述的“全局统计—局部处理”的情况，即整个算法的数据处理过程均为全部影像像元计算，而不是如同参考文献[9]或文献[2]中所列的仅对像元或基于窗口处理的情况，这就给数据的划分与分块处理带来困难. 因为如果分块不当的话，就将直接影响着均值漂移滤波的结果，并进而影响均值漂移分割的效果.

在进行影像分块后，就可以对不同的影像数据块分别进行均值漂移滤波过程，而且也可以应用并行计算环境对不同的数据块进行同时滤波，以达到数据快速处理的效果，其实现可以为基于 MPI 的局域网集群计算环境，以及同一台计算机的多线程编程实现等. 一般来说，进行完滤波并进行影像分割后，还需要对构建不同的数据对象层次关系及对象合并规则，并根据用户定义的尺度进行不同级别的对象合并，最后形成不同的多尺度分割效果. 采用均值漂移算法可以进行一次滤波、多次合并，由于合并过程实现速度较快，因此不需要对此过程进行并行化实现，而且也可以多次进行，以实现用户端的多尺度分割效果（见图 1 多尺度分割的箭头指向）. 以下将就均值漂移滤波过程的并行化进行深入的探讨.

2　均值漂移算法的并行化实现

根据上面的分析，在进行均值漂移分割算法的并行化实现过程中，将重点对均值漂移滤波过

程的数据划分方法进行研究,如图3所示. 如果采用最简单的数据划分方法,如图3(a)~3(c),3(a)为原始数据块,3(b)为分为两块的示意图,3(c)为两块处理完毕后的合并效果图. 在3(b)所示的两块数据划分与分别滤波过程中,彼此间并没有考虑对方数据对本块数据的影响,而且不同的图像数据块具有不同的全局统计特征,因此在将结果进行合并再进行矢量化后,则形成了两块数据块的分割区域不能完好对应的问题,因此导致最终的3(c)处有一条明显的"分块界线",同时,3(c)的上部分与下部分本该属于同一区域的部分也被分为了不能很好对应的两块,而且两块的边界线在"分块界线"处也不一致,不符合实际情况.

针对这种对应不一致的情况,对数据划分方法进行了改进,提出了一种新的"增加缓冲区"式数据划分方法对此问题进行解决,如图3(d)~3(g)所示. 图中3(d)为原始数据块,3(e)为带"缓冲区"的数据划分方法,即每部分的数据行数都比简单分为两份多一些,这样在滤波处理每一块数据时,在接近边界时都会对边界有一定的缓冲过程,从而在3(f)的合并过程中,可以根据用户的设定方法进行缓冲区的数据进行有效取舍,拼接线内的部分实际上是参照两个部分的滤波结果的综合结果,并根据规则完成其合并过程,形成3(g)的合并结果.

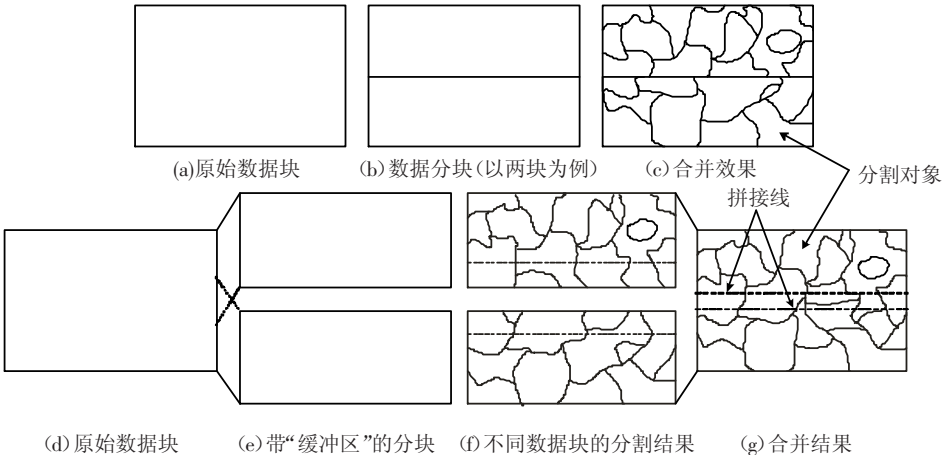


图3　面向均值漂移算法的影像并行化处理数据划分

其中,合并的规则与步骤如下:

1) 分析两个冗余数据块在分割线两侧对应的所有数据块,统计分块后各块平均RGB值及面积;

2)根据各块面积、对应的RGB值以及相对分割尺度计算合并阈值(阈值对应到尺度及RGB值上);

3)对分割线两侧各分割块分别进行判定并进行分割块合并,更新RGB值及对应分割线.

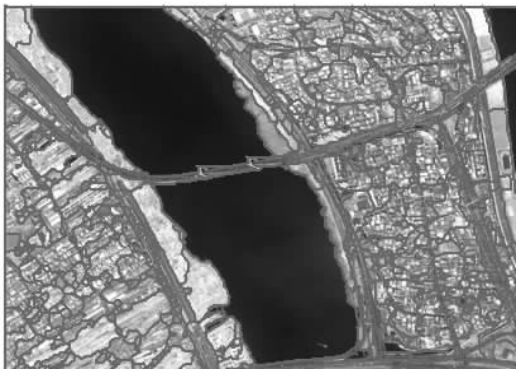
3　影像数据分割及效果分析

根据以上的研究思路,对均值漂移影像分割算法的并行化进行了实现,用户可以一次性根据需要指定多个分割尺度,并可选择是否采用并行化实现方法(可包括面向多线程并行实现与MPI的集群环境多进程并行实现等). 实验过程中,分别采用了平均数据划分方法及"缓冲区"式数据划分方法进行了实验,并对采用不同的分块方法效果进行了对比.

图4所示的实验影像为一幅IKONOS影像,实验数据大小为1 600×1 100,图4(a)为直接将图像划分为两个1 600×550大小的数据块并进行均值滤波后的分割结果,图4(b)为将图像采用图3(d)~3(g)所示的方法进行的数据块划分,数据块大小为1 600×650,即每个数据块在纵向上延伸了100行,以达到分割效果缓冲的目的. 从图4(a)中可以明显看出中间具有一条分隔线,而图4(b)中这条分隔线则很不明显,其效果类似于数据整体均值漂移分割处理后的效果.

进一步的实验中,还采用了同样方法对SPOT5整景融合影像的分割测试,实验数据大小为22 000×22 000,数据量接近1.4 GB,实验过程中对数据进行了分16块分别进行处理(处理过程采用基于串并混连的并行处理方式实现),每块在纵向上延伸了100行,达到了同样的效果,而如果不采用分块方法,则由于算法需要开辟过大内存的原故,无法直接对此数据文件进行处理.

(a)普通分块方法



(b) 带“缓冲区”分块方法

图 4　不同数据划分方法的效果示意图

4　结　论

为提高大数据量影像数据的分割速度,以及克服部分大数据量数据无法直接处理的问题,本文针对均值漂移算法的实现特点与过程,提出采用算法的并行化改进的方式对算法进行改造. 由于均值漂移算法的滤波过程需要进行全局性数据分析,常规的简单的数据划分方法会导致不同数据块的分割结果合并后出现较明显的分隔线问题,本文在这种算法并行化实现方法分析的基础上,提出了一种“缓冲区”式数据划分方法,这种方法通过缓冲区间的数据过渡,能够有效地解决“分隔线”问题,使得影像数据的分割结果同不分块的效果类似. 相应地,本文的算法亦可推广到其他类似的算法并行化改进过程中.

本文仅从算法角度提出并论证了这种数据划分方法的可行性,但并没给出缓冲区选取大小的依据;同样,在这种方法应用到多线程、基于 MPI 的多进程并行计算过程的效率问题,笔者也没进行较为详细的论证,这些问题都有待于进一步研究、分析与论证.

参考文献:

[1] 周成虎,骆剑承. 高分辨率卫星遥感影像地学计算[M]. 北京: 科学出版社. 2009.

[2] ZALIK K R, ZALIK B. A sweep-line algorithm for spatial clustering [J]. Advances in Engineering Software, 2009 (40): 445 – 451.

[3] MUKHERJEE D P, LEVNER Y P, ZHANG H. Ore image segmentation by learning image and shape features [J]. Pattern Recognition Letters, 2009 (30): 615 – 622.

[4] 王永忠, 梁彦, 赵春晖. 基于多特征自适应融合的核跟踪方法[J]. 自动化学报,2008,34(4): 393 – 399.

[5] SHEN Z, LUO J, HUANG G, *et al.* Distributed computing model for processing remotely sensed images based on grid computing [J]. Information Sciences, 2007 (177): 504 – 518.

[6] INNOCENTI E, SILVANI X, MUZYA A, *et al.* A software framework for fine grain parallelization of cellular models with OpenMP: Application to fire spread [J]. Environmental Modelling & Software, 2009(24): 819 – 831.

[7] 李乡儒, 吴福朝, 胡占义. 均值漂移算法的收敛性[J]. 软件学报,2005,16(3). 365 – 374.

[8] COMANICIU D, MEER P. Mean Shift: A robust approach toward feature space analysis[J]. IEEE Transactions On Pattern Analysis and Machine Intelligence, 2002,24(5). 603 – 619.

[9] 沈占锋, 骆剑承, 陈秋晓. 高分辨率遥感影像并行处理数据分配策略研究[J]. 哈尔滨工业大学学报, 2006,38(11). 1968 – 1972.

[10] CHEN H F, MEER P. Robust regression with projection based M-estimators [C]//Proceedings of the Ninth IEEE international Conference on Computer Version. Washington: IEEE Computer Society, 2003: 878 – 885.

(**编辑**　张　宏)