基于投影寻踪的高光谱图像异常检测并行算法

王 维 赵慧洁 董 超

(北京航空航天大学 仪器科学与光电工程学院, 北京 100191)

摘 要: 投影寻踪方法能有效提取数据中的非高斯结构凸显异常信息,但在求解最优投影方向时存在计算量大、运行时间长的问题,为提高处理效率,提出一种机群环境下的并行算法. 选用偏度和峰度组合作为投影指标,将所有像素光谱作为特定投影方向集依次搜索,求解最优投影方向. 在并行计算各候选方向投影指标时,分割图像数据分布存储于各机群结点,数据子块朝候选方向并行投影后,将指标计算式变形分解,使各结点在指标计算过程中所需数据均为本地数据,解决数据局部性问题,并采用一种"轮流作主"的机制提高算法负载均衡程度. 利用实用型模块化高光谱仪数据在机群系统上进行测试,达到了较好的加速效果,表明该并行算法具有良好的并行性能.

关 键 词: 遥感; 图像异常检测; 并行算法; 投影寻踪; 特定投影方向集

中图分类号: TP 751. 1

文献标识码: A **文 章 编 号**: 1001-5965(2009)03-0342-05

Parallel algorithm of anomalies detection in hyperspectral image with projection pursuit

Wang Wei Zhao Huijie Dong Chao

(School of Instrument Science and Opto-electronics Engineering, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: Projection pursiut can extract non-Gaussian structure in hyperspectral data to reveal the anomalies information, but searching the best projection directions is a computational intensive task. To improve the process efficiency, a parallel process algorithm under cluster system was presented. The combination of skewness and kurtosis was selected as projection index (PI). Using all the pixels' spectral as a special projection dierection set and searching the best projection directions in it. While parallel computing the PI value for each candidate dierection, the hyperspectral data was distributed to each computing node after partitioned evenly. After projecting each data subblock to a candidate direction in parallel, the index computation was transformed and decomposed. This makes all the data needed during index computation be in local memory for each node and decreases the communication. Furthermore, a "be host in turn" method was put forward to improve the degree of load balance. Using an operative modular imaging spectrometer data to test the efficiency on cluster, the results show that the parallel algorithm achieves good parallel performance.

Key words: remote sensing; image anomalies detection; parallel algorithm; projection pursuit; special projection direction set

高光谱遥感是一种全新的遥感技术,能为图像像元提供数百个波段的光谱信息. 随着其空间及光谱分辨率的不断提高,传感器采集的每景数据量可达 GB 级,丰富的光谱信息能反映光谱间细微差异,提高了遥感探测能力,使其广泛应用于军事目标探测、森林火灾监控等领域. 但这些应用均对数据处理有实时或者近实时的需求以做出快速决策,这就需要针对高光谱海量数据处理设计具有高性能的并行处理算法. 随着高光谱图像数据挖掘的深入,并行处理已成为一种提高处理速

收稿日期: 2008-03-17

作者简介: 王 维 (1984 -),男,湖南常德人,硕士生, wangweibh@126. com.

度的有效途径[1].

在高光谱图像异常检测应用中,利用投影寻踪(PP, Project Pursuit)方法可将高光谱图像的高维数据投影至低维空间,凸显出图像中的异常成分后通过阈值分割实现异常目标提取. 对于最优投影方向的求解,文献[2]提出了一种特定候选投影方向集搜索的方法. 这种方法简单易行,一般情况下求解精度均能满足异常目标检测需求,但需要较长处理时间. 为提高算法处理速度,本文在机群环境下,设计了一种并行处理算法,有效地缩短了算法执行时间. 实验结果表明该并行算法具备了较强的适用性和有效性.

1　算法思想

高光谱图像中异常目标的光谱特征与背景存在较大差异. 一般情况下,自然背景符合高斯分布,而异常目标则为远离高斯分布的数据结构,利用此特性可在没有先验信息的条件下进行异常检测.

假设一 k个波段 n个像素的高光谱数据,所有像素可以按顺序构成一 $k\times n$的数据矩阵 X, $X=[x_1,\cdots,x_{n-1},x_n]_{k\times n}$. 对 X使用主成分变换预处理[3],保留前 d个特征值较大的主成分,实现数据的降维降噪处理,得到 d维预处理结果 X_d, $X_d=[\hat{x}_1,\cdots,\hat{x}_{n-1},\hat{x}_n]_{d\times n}$. 投影寻踪异常检测就是寻找最大化投影指标函数 $F(\cdot)$的最优方向 V_{best},使线性投影 $V_{best}^{T}X_d$ 最大程度地凸显数据中远离高斯分布的异常信息. 为检测多类异常成分,在获得一类异常信息后,数据朝方向 V_{best}的正交子空间投影以去除数据集内该方向所代表的异常成分,然后在剩余空间中进行新的检测. 方向 V 的正交子空间定义为式(1), X_d 朝正交子空间投影定义为式(2). 基于最优投影方向投影可以获得一系列凸显各类异常特征的分量图像,对这些分量图像使用直方图阈值分割即可提取出异常目标.

$$P_v^{\perp}=I-V\cdot(V^{T}\cdot V)^{-1}\cdot V^{T} \tag{1}$$

$$X_d'=P_v^{\perp}\cdot X_d \tag{2}$$

文中采用偏度与峰度(K_3, K_4)的组合作为投影指标函数 $F(\cdot)$描述投影图像远离高斯分布的程度[4]. 投影向量的非高斯性越大,指标值(PI, Projection Index)越大.

$$F(V^{T}X_d)=PI(V)=K_3^2+K_4^2/12 \tag{3}$$

$$K_3=PI_{skewness}(V)=\sum_{i=1}^{n}(Y_i-\bar{Y})^3/(n-1)\cdot S^3 \tag{4}$$

$$K_4=PI_{kurtosis}(V)=\sum_{i=1}^{n}(Y_i-\bar{Y})^4/(n-1)\cdot S^4-3 \tag{5}$$

式(4)、式(5)中, Y为投影向量 $V^{T}X_d$; $\bar{Y}$为 Y的均值; S为 Y标准差; n为 Y的维数. 求解最优投影方向时, X_d 中所有像素光谱曲线构成一特定投影方向集,依次计算每一光谱作为投影方向时的投影指标 PI,最大化投影指标的光谱曲线即为最优投影方向.

2　并行算法

串行算法通过计算特定投影方向集中各像素光谱曲线的投影指标得到最优投影方向. 该计算过程需要多次扫描全局数据,存在大量矩阵乘法,是串行算法的计算瓶颈. 为实现并行处理,一种直接的方案即在各计算结点冗余复制全图数据,划分特定方向集,并行求解各子集的最优方向,最后通过交互比较得到全局最优方向. 这种方法简单易行,但该检测算法要求图像背景不能过小,当某些实时并行处理系统结点内存稍小时,往往不能满足该方案冗余复制全图数据的需求. 本文设计的并行方法划分图像数据,通过分解计算并行求解每一候选方向投影指标,获得了与前一种方案相当的加速效果,且受图像数据大小及硬件内存容量限制小,具备更广的适用能力.

2.1　投影指标的并行求解

为尽量减少各结点间通信交互,根据参与计算的结点个数 P将数据 X按列划分为 P个大小一致的数据子块 X_r, $r\in[1,P]$. 为进行各数据子块的并行预处理,根据主成分变换原理,先将协方差计算分解至各计算结点并行进行,主结点获得完整的协方差矩阵后计算特征矩阵 U 与特征值对角阵 D,并将其广播至所有结点. 各结点根据 U 与 D 对本地数据子块进行预处理得 X_{dr}[5].

由于预处理结果 X_d 分解为 P个子块 X_{dr},在计算第 i个候选方向 V_i(X_d 第 i列)的投影指标时,首先由该方向所在结点将其广播至其它 $P-1$ 个结点,然后数据 X_d 朝方向 V_i 的并行投影在分布式系统中的变形如式(6). 式中, $Y(n)$为 X_d 的投影向量, n表示向量维度, $Y(n/P)_r$ 为 X_{dr}的投影分量.

$$V_i^{T}X_d=Y(n)=V_i^{T}[X_{d1},\cdots,X_{dr},\cdots,X_{dP}]=[Y(n/P)_1,\cdots,Y(n/P)_r,\cdots,Y(n/P)_P] \tag{6}$$

并行投影后,为计算投影向量 $Y(n)$的投影指标,首先需要计算 $Y(n)$的均值 $\bar{Y}$和标准差 S. 为最小化数据间相关性,对 $Y(n)$标准差 S的计算

进行如下变形:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}Y_i^2 + n\cdot\bar{Y}^2 - 2\cdot\bar{Y}\cdot\sum_{i=1}^{n}Y_i\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n}Y_i^2 - n\cdot\bar{Y}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n}Y_i^2 - \frac{1}{n}\cdot\sum_{i=1}^{n}Y_i\cdot\sum_{i=1}^{n}Y_i\right) = \frac{1}{n-1}\left(\sum_{1}Y_i^2 + \cdots + \sum_{P}Y_i^2\right) - \frac{1}{n(n-1)}\cdot\left(\sum_{1}Y_i + \cdots + \sum_{P}Y_i\right)^2 \quad (7)$$

其中 $$\sum_{1} = \sum_{i=1}^{n/P}$$

$$\sum_{P} = \sum_{i=1+n(P-1)/P}^{n}$$

变形后,各结点根据本地投影分量 $Y(n/P)_r$ 计算分量 $\sum Y_i^2$ 和 $\sum Y_i$ 的值并传递给主结点,主进程根据各结点结果计算出向量 $Y(n)$ 的 $\bar{Y}$ 和 S,并返回所有计算结点. 各结点则根据 $\bar{Y}$ 和 S 并行计算投影向量 $Y(n)$ 的投影指标 PI 分析指标函数中的偏度和峰度计算式可知,两者可进行如下变形:

$$K_3 = \sum_{i=1}^{n}(Y_i - \bar{Y})^3/(n-1)\cdot S^3 = \sum_{i=1}^{n/P}(Y(n/P)_{1i} - \bar{Y})^3/(n-1)\cdot S^3 + \cdots + \sum_{i=1+n(P-1)/P}^{n}(Y(n/P)_{Pi} - \bar{Y})^3/(n-1)\cdot S^3 = (K_3')_1 + \cdots + (K_3')_r + \cdots + (K_3')_P \quad (8)$$

$$K_4 = \sum_{i=1}^{n}(Y_i - \bar{Y})^4/(n-1)\cdot S^4 - 3 = \sum_{i=1}^{n/P}(Y(n/P)_{1i} - \bar{Y})^4/(n-1)\cdot S^4 - \frac{3}{P} + \cdots + \sum_{i=1+n(P-1)/P}^{n}(Y(n/P)_{Pi} - \bar{Y})^4/(n-1)\cdot S^4 - \frac{3}{P} = (K_4')_1 + \cdots + (K_4')_r + \cdots + (K_4')_P \quad (9)$$

由变形可知所有结点均可局部计算其本地投影分量 $Y(n/P)_i$ 的偏度分量 $(K_3')_r$ 和峰度分量 $(K_4')_r$,然后由主结点对所有偏度、峰度分量求和获得投影向量 $Y(n)$ 的偏度与峰度,并根据式(3)计算出方向 V_i 的投影指标 $PI(V_i)$.

2.2 并行性能的影响因素

并行粒度与负载均衡是影响并行效率的2个重要因素[6]. 为了减少通信次数,可增大并行粒度 G,即各结点同时计算 G 个候选方向的指标后再进行通信交互. 对不同粒度进行测试,实验表明 $G=100$ 时通信时间最少,且粒度较小,易负载均衡. 此外,若固定由主结点收集指标分量并计算指标 PI(V),此时其它结点将处于等待状态,未能充分利用系统资源. 为提高负载均衡程度,采用一种"轮流做主"的平衡机制,各结点将计算出的指标分量传送至一伪"主"结点后,该伪"主"结点并不立即进行 PI(V)的计算,而是同其它结点再次计算下一轮候选方向的指标分量,直到所有计算结点均轮流作为伪"主"结点获得指标计算任务后,各结点并行计算这 $G\cdot P$ 个候选方向的投影指标. 实验测试了该方法在不同结点个数下的负载比,表1结果表明算法达到了较好的负载均衡.

表1 不同结点个数下的负载比

结点数	2	4	6	8	12	16
负载比	1.008	1.013	1.025	1.114	1.057	1.08

并行求解完所有候选方向的投影指标后,比较各结点上的最优方向得全局最优投影方向 V_{best}. 同时各结点根据 V_{best} 进行去正交操作,去除 V_{best} 所代表的异常成分后重新进行上述并行处理,检测新异常成分. 最后,所有结点将各最优投影方向下的本地数据投影结果传回主结点,由主结点分割出各类异常目标. 两结点下的并行处理流程如图1所示.

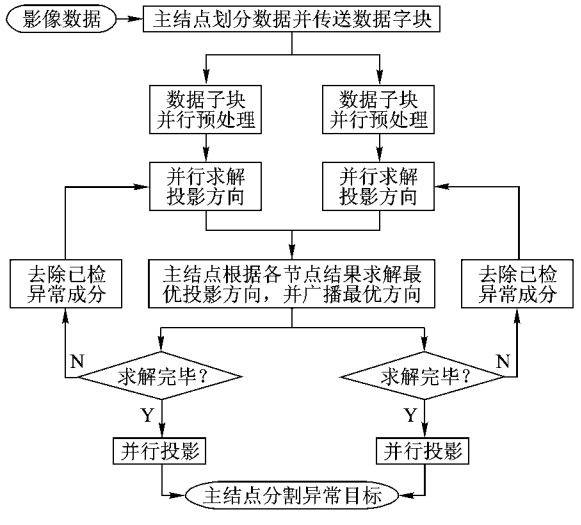


图1 并行算法流程图

3 实验结果

实验采用图2中实用模块化高光谱仪数据,共150行,103列,94波段. 图中49个仿真目标点为背景与同航带目标光谱及白噪声线性混合而成. 前6行目标自上而下混合程度依次下降,最后一行为纯像元. 可认为该图像中有不同含量的仿

真目标点及农田中小道 2类异常目标,背景与目标光谱如图 3所示,混合程度高的仿真目标、小道与背景间相似度高,需要凸显异常特征后才能有效检测.

并行处理算法通过分解计算实现分布式处理,并未改变该算法思想,因此处理结果与串行算法一致,获得的各类异常成分分量图像如图 4所示.利用灰度直方图的尾部波澜计算阈值对凸显特征的分量图像进行目标分割,较好地分割出了图 4a中异常成分含量 20%以上的点目标和图 4b、图 4c中垂直及水平方向小道的具体位置,如图 5所示.

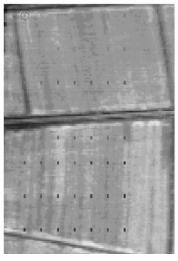


图 2　数据某波段图

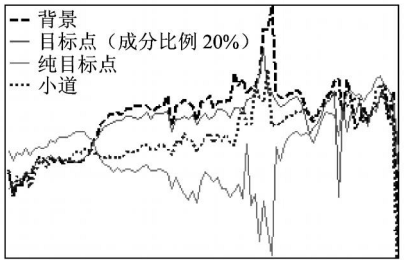
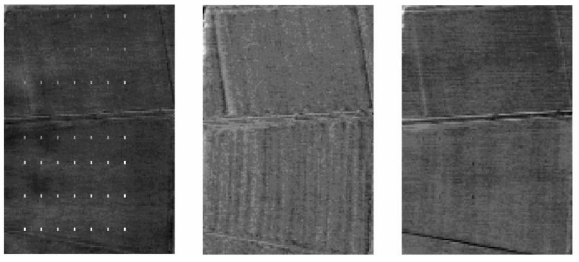


图 3　异常成分与背景光谱



a仿真点分量　　b垂直小道分量　　c水平小道分量

图 4　前 3个异常类型分量图像



a仿真点目标　　b垂直小道目标　　c水平小道目标

图 5　各类型异常分量图像的目标分割结果

并行实验在 16结点机群系统下进行.实验比较了文中所提 2种并行方案性能,如表 2所示.并行方法 1采用的是分割投影方向集,各结点冗余复制全图数据的方案;并行方法 2采用的是本文所述的分割图像数据,并行计算每一候选方向投影指标的方案.图 6中加速比曲线表明,2种方法均能达到较好的并行加速效果.并行方法 1的加速比曲线基本呈线性增长,但由于数据预处理只能串行处理,一定程度地限制了并行效率;并行方法 2的加速比曲线在计算结点个数增长至 6时,加速效果减缓,这是由于结点个数增多后算法通信开销增大,导致并行效率下降,而当结点个数进一步增多时,由于数据划分,各结点数据量更小,数据的计算寻址时间更短,使加速性能又有所回升,加速比最终达到 15. 1.该结果表明,本文所述并行方法在 16结点机群并行处理下能达到并行方法 1的加速效果,具备近实时的处理能力,且适应性更强.

表 2　并行实验结果

结点个数	并行方法 1			并行方法 2		
	时长/s	加速比	效率	时长/s	加速比	效率
1	182. 1	1	1	182. 1	1	1
4	47. 1	3. 86	0. 96	47. 59	3. 83	0. 96
8	24. 4	7. 46	0. 93	29. 56	6. 16	0. 77
12	16. 9	10. 8	0. 90	18. 48	9. 85	0. 82
16	13. 1	13. 9	0. 87	12. 06	15. 1	0. 94

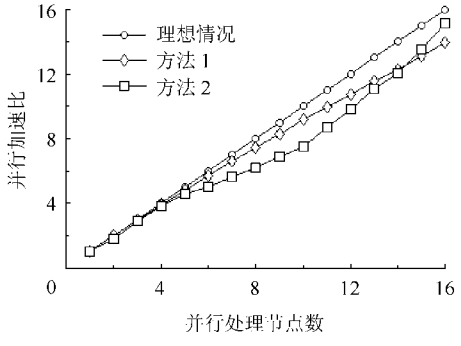


图 6　并行算法的加速比曲线

4　结束语

投影寻踪方法能较好地自动检测高光谱图像中分布异常的目标.本文利用机群并行环境,针对特定投影方向集的最优投影方向求解方法设计了一种较细粒度的并行处理算法,通过分解投影指标很好地解决了数据并行计算的局部性问题.并与另一种并行实现方案进行比较,结果表明该并行算法适应性更强的同时也具备了良好的加速性能.

参考文献 (References)

[1] Plaza A, Valencia D, Plaza J, et al. Commodity cluster-based parallel processing of hyperspectral imagery[J]. Journal of Parallel and Distributed Computing, 2006, 66(3): 345 - 358

[2] Ifarraguerri A, Chang Chein-I. Unsupervised hyperspectral image analysis with projection pursuit[J]. IEEE Transactions on Geoscience and Remote Sensing, 2000, 38(6): 2529 - 2538

[3] Chiang S S, Chang C I, Ginsbern IW. Unsupervised target detection in hyperspectral images using projection pursuit[J]. IEEE Transactions on Geoscience and Remote Sensing, 2001, 39(7): 1380 - 1391

[4] 崔恒建, 陈广雷. 带有误差变量的偏度和峰度正态性检验[J]. 北京师范大学学报:自然科学版, 2000, 36(1): 9 - 13
Cui Hengjian, Chen Guanglei. Skewness and kurtosis statistics for testing normality in ev model[J]. Journal of Beijing Normal University (Natural Science), 2000, 36(1): 9 - 13 (in Chinese)

[5] 胡冰, 周海芳, 王攀峰. 遥感图像 PCA 融合并行算法研究与实现[J]. 微电子学与计算机, 2006, 23(10): 153 - 161
Hu Bing, Zhou Haifang, Wang Panfeng. A parallel algorithm of PCA image fusion in remote sensing and its implementation[J]. Microelectronics and Computers, 2006, 23(10): 153 - 161 (in Chinese)

[6] 陈国良, 安虹, 陈崚, 等. 并行算法实践[M]. 北京: 高等教育出版社, 2004: 294 - 317
Chen Guoliang, An Hong, Chen Ling, et al. Practice of parallel algorithm[M]. Beijing: Higher Education Press, 2004: 294 - 317 (in Chinese)

(责任编辑:文丽芳)

(上接第 287页)

实时性的重要因素. 在目前误码率为 1.2×10^{-4} 的情况下,所有消息的最差延时仍小于其最大允许延时,说明在发生一定的误码率的情况下,网络仍可以保证实时性. 网络负载率结果表明(图 9下):随着误码率的增加,网络负载率增加. 这是由于误码导致增加了网络恢复通信的链路占用开销和消息的重传开销,这些额外开销均导致负载增加.

4 结束语

本文实现了在 OPNET仿真环境下 CAN仿真系统的构建,完成了节点模型、链路模型和数据帧模型的设计与实现. 并通过文献[9]中的网络模型和数据验证本文所建模型的真实性. 在该仿真环境下,对典型的 CAN网络拓扑模型进行仿真,通过分析和对比仿真结果,讨论了 CAN网络的性能特点. 本文所构建模型真实地反映了 CAN协议的特点,具有一定的通用性,并测试了 CAN网络的性能特点,从而为 CAN网络更广泛的应用提供了宽阔的新舞台.

参考文献 (References)

[1] ISO 11898, Road vehicles—interchange of digital information—controller area Network for high-speed communication ISO[S]

[2] Kotzian J, Srovnal V. CAN based distributed control system modeling using UML[C]//2003 IEEE International Conference on Industrial Technology. Maribor, Slovenia: IEEE, 2003, 2: 1012 - 1017

[3] Acle J P, Reorda M S, Violante M. Early, accurate dependability analysis of CAN-Based networked systems[J]. Design & Test of Computers, 2006, 23(1): 38 - 45

[4] Donlin A. Transaction level modeling: flows and use models[C]//Proceedings of the 2nd IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis. Stockholm, Sweden: CODES + ISSS, 2004: 75 - 80

[5] Schirner G, Dorner R. Result-oriented modeling-a novel technique for fast and accurate TLM[J]. IEEE Trans on Computer-Aided Design of Integrated Circuits and Systems, 2007, 26(9): 1688 - 1699

[6] 孙松, 孙川. 测控网络中 CAN总线及其实现[J]. 计量与测试技术, 2004, 31(07): 23 - 25.
Sun Song, Sun Chuan. CAN bus in measuring and controlling system and its implementation[J]. Metrology and Measurement Technique, 2004, 31(07): 23 - 25 (in Chinese)

[7] Pinho L M, Vasques F. Reliable real-time communication in CAN networks[J]. IEEE Trans on Computers, 2003, 52(12): 1594 - 1607

[8] Punnekkat S, Hansson H, Norstrom C. Response time analysis under errors for CAN [C]//2000 Proceedings of Real-Time Technology and Applications Symposium. Washington: IEEE, 2000: 258 - 265

[9] Pinho L M, Vasques F, Tovar E. Integrating inaccessibility in response time analysis of CAN networks[C]//2000 Proceedings of Factory Communication Systems. Porto, Portugal: IEEE International Workshop, 2000: 77 - 84

(责任编辑:娄 嘉)