第30卷，第3期
2010年3月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 30, No. 3, pp743-747
March, 2010

基于线性最小二乘支持向量机的光谱端元选择算法

王立国，邓禄群，张 晶

哈尔滨工程大学信息与通信工程学院，黑龙江 哈尔滨 150001

摘 要 光谱端元选择是高光谱数据解混分析的重要前提。在各种端元选择算法中，N-FINDR算法因其自动性和高效性受到广泛欢迎。然而，该算法需要进行数据降维预处理，且包含大量的体积计算导致该算法的运算速度较慢，限制了该算法的应用。为此提出基于线性最小二乘支持向量机的N-FINDR改进算法，该算法无需降维预处理，且采用低复杂度的距离尺度代替复杂的体积尺度来加速算法。此外还提出对野值点施加有效控制以赋予算法鲁棒性，以及利用像素预排序方法来降低算法的迭代次数。实验结果表明，基于线性最小二乘支持向量机的改进N-FINDR算法在保证选择效果的前提下复杂度大大降低，鲁棒性方法和像素预排序方法则进一步提高了算法的选择效果和选择速度。

关键词 高光谱图像；光谱端元选择；线性最小二乘支持向量机；N-FINDR算法

中图分类号：TP75 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2010)03-0743-05

引 言

随着遥感技术的发展，高光谱图像(HSI)得到了越来越广泛的应用。高光谱图像的空间分辨率一般较低，这种情况导致了混合像素的广泛存在，即一个像素可能是几种类别的混合。对于这类像素，将其按照一般分类方法[1]归属为任何一类都是不准确的。分析各类别成分在混合像素内所占的比例的技术称为光谱解混[2]，是高光谱数据分析的最基本、最重要内容之一，从实质上讲它是一种更为精确的分类技术。光谱解混实施的必要前提是要知道高光谱数据中包含哪些地物类别，在此背景下提取各类别代表性纯光谱的技术称为光谱端元选择，简称端元选择。在近十多年里，多种高光谱图像光谱端元选择方法相继发展起来[3]。典型的光谱端元选择方法包括：像素纯度索引(pixel purity index，PPI)[4]、N-FINDR算法[5]、迭代误差分析(iterative error analysis，IEA)[6]、光学实时自适应光谱辨识系统(optical real-time adaptive spectral identification system，ORASIS)[7]、自动形态光谱端元选择(automated morphological endmember extraction，AMEE)[8]，等等。IEA是基于迭代的方法，在这个过程中，能降低约束光谱解混误差的像素被当作光谱端元。该方法的缺点是越是先选出的光谱端元根据性越差，而像素一经被选作光谱端元便无法更新。ORASIS通过学习和矢量量化(LVQ)来进行光谱端元的选择，但该方法对于阈值参数极其敏感。AMEE利用形态学，选择光谱端元的过程中同时利用了空间和光谱信息，其不足之处在于运算量较大。PPI和N-FINDR是基于N维空间谱凸多面体的搜索光谱端元的经典例子。N-FINDR是全自动的方法，PPI是半自动化的方法。相比之下，N-FINDR因其具有全自动、无参数、选择效果较好等优点而受到广泛欢迎。但该算法需要进行数据降维预处理，且包含大量的体积计算，这也是它最为耗时的部分。并且，体积计算(即主要为行列式的计算)的复杂度将随着所选择的光谱端元数目增大而呈现立方增长，从而导致算法运算速度大大降低。

目前已有一些典型文献提出了对N-FINDR算法的改进方案。文献[9]引入虚拟维(virtual dimensionality，VD)的概念来确定待选择光谱端元的数目，对算法的实施具有一定意义，但这并不能改变该算法的如上两点不足。文献[10]采用像素预选的方式来降低后续搜索的复杂性，也是从侧面来降低算法计算量。文献[10,11]利用顺次选择的方式来代替联合选择的方式。顺次选择方法的主要缺陷在于算法的初始化端元缺乏合理性，常利用距离原点最近或(和)最远的像素，或平均光谱向量作为初始端元，而这一取法在一定程度上缺乏合理性。事实上，这种顺次选取的方式远离了N-FINDR算法的基本特征而走向IEA端元选择模式，像素一经选定便无法更新，光谱端元之间的相互依赖关系也无法得到最大满

收稿日期：2009-01-20，**修订日期**：2009-04-25

基金项目：国家自然科学基金项目(60802059)，教育部博士点新教师基金项目(200802171003)和水下智能机器人技术国防科技重点实验室项目资助

作者简介：王立国，1974年生，哈尔滨工程大学信息与通信工程学院副教授 e-mail：wangliguo@hrbeu.edu.cn

足。文献[12]提出的方法可以直接在原始数据空间上进行而免于降维预处理，因此选出的光谱端元更具合理性，在理论上突破了 N-FINDR 算法需要降维处理的传统模式，但该方法也属于顺次选取。

另一方面，基于凸几何分析的端元选择算法容易受到野值点的影响，而野值点在高光谱图像中大量存在，现有文献对此并未提出相应的解决方案。

在此背景下，本文在深入分析线性最小二乘支持向量机(Linear least square support vector machines，简记为 LLSSVM)[13, 14]模型的基础上，提出相应的改进措施，建立可在原始空间实施、免于体积计算、并对野值点干扰具有鲁棒性的 N-FINDR 改进算法。

1 N-FINDR 端元选择算法

高光谱图像全部像素在高光谱数据空间中形成一个凸多面体，每个光谱端元则对应于凸多面体的一个顶点。在这种情况下，光谱端元选择的任务就变为选择数据空间所形成的凸多面体的顶点。由于全部光谱端元作为顶点的凸多面体具有最大的体积，因此该任务又转为寻求指定数目的像素，使得由它们作为顶点的凸多面体具有最大的体积。需要强调的是，算法的实施并没有在原始数据空间中进行，而是在经过 MNF 后的变换空间中进行的，这种降维处理的目的之一是为了凸多面体的体积计算能够得以实施。篇幅所限，具体迭代过程参见文献[5]。这里特别说明一点，对于$(d+1)$个像素 p_1，p_2，…，p_{d+1}所张成的凸多面体的体积计算公式为

$$V(E)=\frac{1}{(d+1)!}\text{abs}(|E|)$$
$$\boldsymbol{E}=\begin{bmatrix}1 & 1 & \cdots & 1\\ p_1 & p_2 & \cdots & p_{d+1}\end{bmatrix} \tag{1}$$

abs(·)和|·|分别为绝对值和行列式算子。

2 提出的端元选择算法和预处理方法

2.1 基于 LLSSVM 的 N-FINDR 改进算法

有关支持向量机的理论读者可参阅文献[13-16]，此处从略。LLSSVM 的判别函数式为

$$f(x)=\sum_{i=1}^{n}\alpha_i K(x_i,x)+b=\langle w^*,x\rangle+b^* \tag{2}$$

$$\begin{cases}w^*=\sum_{i=1}^{n}\alpha_i x_i y_i\\ b^*=-\frac{1}{2}(\max_{y_i=0}(\langle w^*,x_i\rangle)+\min_{y_i=+1}(\langle w^*,x_i\rangle))\end{cases}$$

下面我们利用 1-a-r 型多类 LLSSVM 来建立 N-FINDR 算法。为了便于说明和取得可视化效果，首先考虑二维空间的情形。在图 2 中，以 A，B，C 作为顶点形成了一个三角形(二维凸多面体)，记它的体积为 V_{old}。令 A_0 为不同于 A，B，C 的点，则由 A_0，B，C 形成了一个新的三角形，记它的体积为 V_{new}。再记线段 AD，A_0D_0 分别为 A 和 A_0 到线段 BC 的距离。那么，将 A 替换为 A_0 是否为有效替换只需比较面积(二维体积)V_{old} 和 V_{new} 的大小即可。为此，原始的 N-FINDR 算法需要根据公式(1)具体地计算 V_{old} 和 V_{new}。然而，我们看到 V_{old} 和 V_{new} 的大小关系与 AD 和 A_0D_0 的大小关系是一致的，因此，比较 V_{old} 和 V_{new} 可以通过比较 AD 和 A_0D_0 来完成。图 1 给出了这种直观的说明。其中，l 是由点 B 和 C 形成的直线，l_1 为过点 A 且与 l 平行的直线，l_2 为直线 l_1 关于直线 l 的对称直线。

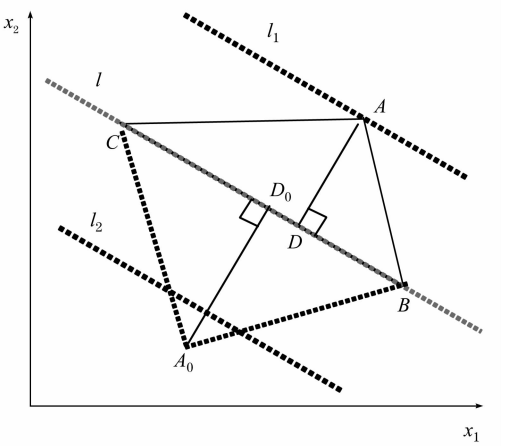

Fig. 1 Measure substitution in N-FINDR algorithm

在多类(此时为 3 类)LLSSVM 的子分类器之中，规定光谱端元 A 为一类，B 和 C 为另一类

$$f(A)=1,\ f(B)=f(C)=0 \tag{3}$$

则判别函数 $f(\cdot)$定义了一种由点·到直线 BC 的有向距离(注意，此距离正比于欧式距离)，这种距离绝对值的相对大小直接反映了多面体体积(此时为三角形面积)的相对大小关系。SVM 对于小样本学习问题具有良好的性能，而小样本条件下 LLSSVM 可以利用显性解析式方便地求解，因此，我们可以利用 LLSSVM 为端元组合 A，B，C 建立单样本分类模型，从而可以利用 SVM 判别函数进行距离计算和比较。这一情形可以容易地推广到多个端元组合或高维空间中去。

下面我们对算法的相对复杂性进行简要地分析。在 N-FINDR 算法实施中，判别函数的计算次数等于光谱端元更新次数与光谱端元个数的乘积，这将远远少于距离计算的次数。因此，在改进的 N-FINDR 算法中，计算的复杂度可以忽略判别函数的计算而近似等于距离的计算。由(2)式可知，判别函数 $f(\cdot)$的复杂度与端元数目无关，仅为光谱维度的线性复杂度。相比之下，体积计算则需计算一个$(d+1)\times(d+1)$矩阵的行列式，其复杂度随光谱端元数目的增长而立方增长。除了计算量的不同，算法在改进前后对于最终光谱端元选择结果来说是等效的。这样，原始的体积计算转化为当前低复杂度的距离判别计算。

2.2 两种预处理方法

2.2.1 野值点的检测与去除

无论是基于体积计算还是基于距离测算的光谱端元迭代搜索过程，都极易受到野值点的干扰。由 N-FINDR 算法所得到的光谱端元是对应最大体积的像素点组合，因此，野值点以其特殊的空间位置比一般像素点具有更高的概率被选作

光谱端元，从而造成较大影响甚至导致算法的彻底失败。这样的点哪怕只有一个，其潜在的影响也是难以估量的。而在高光谱图像中，野值点又是广泛存在的。如果能够确认野值点，并在光谱端元迭代搜索过程中排除考虑这些点，就能够达到我们的目的。野值点通常以更加孤立的方式存在。这样，可以以每个像素点为中心建立邻域窗，通过计算邻域窗内所包含的像素点数作为中心点的孤立程度度量指标。孤立程度度量指标越大，说明该点作为野值点的可能性就越大。采用方邻域(高维盒子)替换圆邻域，这样可以在基本不影响野值点去除效果的前提下尽可能降低计算量。

2.2.2　像素预排序

为了能够获得快速收敛的重要端元迭代搜索，每个像素点应该根据其作为端元的潜在可能性进行预先评价和排序。根据类别可分性强的像素一般分布在相应的高维几何空间角端的特性[2]，可利用投影统计的方法进行像素预排序。具体地说，当我们将每个数据光谱投影于众多的具有随机方向的测试向量时，重要像素将以较大的概率落在测试向量投影终端。通过这种方式我们就可以进行重要像素的选择。折衷考虑计算的复杂性和选择的准确性，我们只将光谱空间的各维坐标选作测试向量，这样所有的投影结果可以免除任何计算而直接由像素的光谱特征值直接排序得出。具体步骤如下：(1) 由原始数据空间的第一维到最后一维依次选出和排列对应于极大值和极小值坐标的点对；(2) 从余下的数据空间中进行第一步操作；(3) 继续进行这样的过程，直至所有的数据点都被选出、排列。待全部数据点排序完毕之后，排在最前面的若干像素点因为潜在可能性最大而有理由被选作初始重要特征，其迭代更新过程也将按照排序结果依次进行。

3　仿真实验

为方便引用，将基于 LLSSVM 的端元选择算法简记为 LLSSVM-N-FINDR，附加像素预排序时记为 PLLSSVM-N-FINDR。

关于 N-FINDR 算法的端元选择效果文献已不乏论证，而 LLSSVM-N-FINDR 算法无论是在理论上还是在下面的实验中都说明二者具有相同的选择结果。故本文只重点对改进算法的执行效率进行对比论证，同时说明两种预处理方法的有效性。

在第一组实验中，如图 2 所示，$A(-15, 0)$，$B(15, 0)$ 和 $C(0, 20)$作为已知光谱端元被随机产生的归一化系数混合成 1 000 个点(附加方差为 1 的高斯噪声)。实验中，N-FINDR 和 LLSSVM-N-FINDR 算法经过 9 504 次的迭代搜索后均得到了相同的光谱端元(图中三角形顶点)，接近于真实端元位置，而它们的搜索时间分别为 0.817 0 和 0.130 0。进一步，利用像素预排序方法后 LLSSVM-N-FINDR 的迭代次数为 3 725，搜索时间缩减为 0.056 0，较之 N-FINDR 算法速度提高了十几倍。以上结果见表 1。

进一步，将上面的数据通过后补 0 的方式扩展至 10 维空间中。此时，N-FINDR 算法无法为其进行端元选择，而 LLSSVM-N-FINDR算法仍然可以实施，选择结果不变，运行时间也与空间扩展前大致持平。

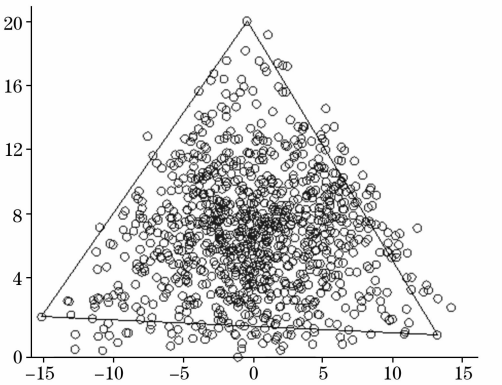

Fig. 2　Synthetic data used in experiment 1

Table 1　Comparison of running time/iterative times for experimental group 1

EM 选择方法	运行时间/s	EM 更新次数	体积/距离计算次数
N-FINDR	0.817 0	21	9 504
LLSSVM-N-FINDR	0.130 5	21	9 504
PLLSSVM-N-FINDR	0.056 2	2	3 725

在第二组实验中，我们将光谱端元的数目增加到 10，即在 9 维空间中，9 个标准的单位向量加上原点被选为光谱端元，并由它们合成 10 000 个数据点。两种方法经过 689 843 次的迭代搜索后均无误差地得到了理想的结果，而它们的搜索时间依次为 377.68 和 10.660。可以看出，随着所选择光谱端元数目的增加，本文方法的效率优势更加明显。进一步，利用像素预排序方法后 LLSSVM-N-FINDR 的迭代次数为 105 562，搜索时间仅为 1.608 0，较之 N-FINDR 算法速度提高了 200 倍以上。以上结果见表 2。

Table 2　Comparison of running time/iterative times for experimental group 2

EM 选择方法	运行时间/s	EM 更新次数	体积/距离计算次数
N-FINDR	377.68	95	689 843
LLSSVM-N-FINDR	10.663	95	689 843
PLLSSVM-N-FINDR	1.608 0	28	105 562

第三组的两次实验旨在对鲁棒性方法的有效性进行论证。实验 1 对真实光谱添加强噪声后进行端元选择；实验 2 对另一组真实高光谱数据直接进行端元选择。图 3 所给出的图像分别为原始光谱端元、利用 N-FINDR 算法直接选择所得到的光谱端元和应用鲁棒 LLSSVM-N-FINDR(RLLSSVM-N-FINDR)所选择到的光谱端元。可见，鲁棒性方法较好地克服了噪声干扰，所获得的结果接近于真实光谱端元。不难分析，每个光谱端元作为所对应类别的类中心，并不应落在数据形体的顶点而是近似顶点的位置，选择顶点只是一种理想状态下的操作。因此，无论高光谱数据有无野值点或

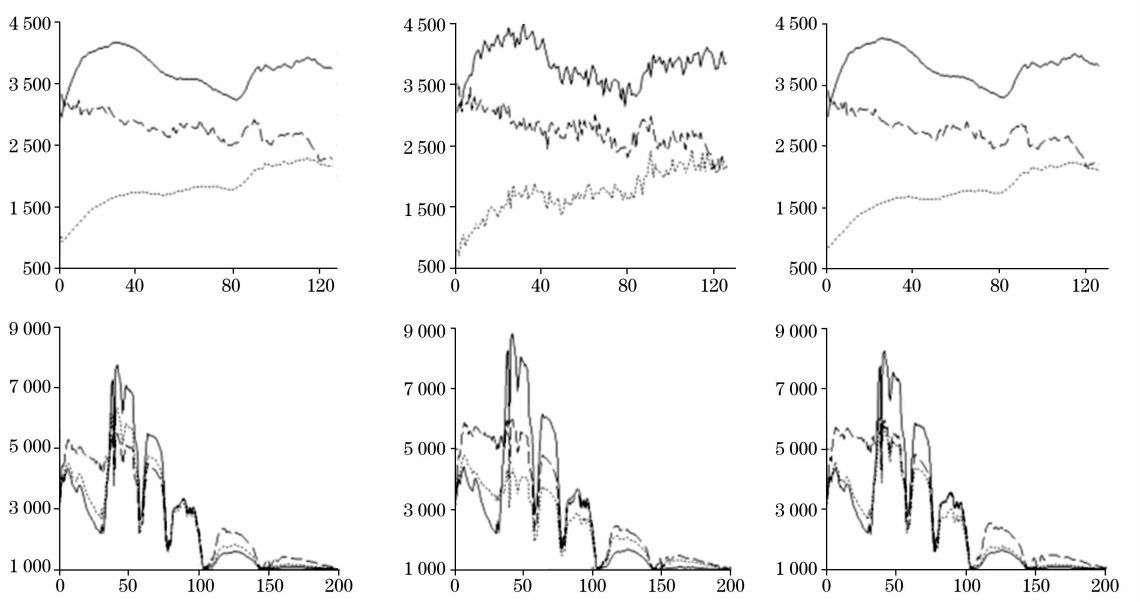

Fig. 3 Comparison of selected EMs of different methods for experimental group 3

Left: Consulting EMs; Middle: EMs selected by N-FINDR; Right: EMs selected by RLLSSVM-N-FINDR

受噪声严重干扰的数据点存在，所提出的鲁棒性方法都将有助于得到更加合理的结果。

4 结 论

本文根据线性 LLSSVM 构建了免于降维预处理的 N-FINDR 算法，该算法中利用距离比较取代原始的体积比较，使得算法的复杂度不再随着光谱端元数目的增加而增加。这种方式在选择结果上与原始算法等价，而在执行效率上明显优于原始方法，所选择的光谱端元数目越大，新算法的优势越明显。对于 SVM 理论，人们更多地了解的是它的分类和回归功能，而本文巧妙地利用它进行距离测算，解决了 N-FINDR 算法中的两大难题。需要注意的是，非线性 SVM 和 1－a－r 以外类型的分类器结构无法实现本文目的；而最小二乘类型的 SVM 虽然理论上可以换作普通类型的 SVM，但对于处理本文这样的超小样本问题效率会大受影响。此外，本文还通过控制野值点来提高算法的鲁棒特性、通过像素预排序来减少算法的迭代搜索次数，可供其他端元选择方法借用。

致谢：感谢澳大利亚新南威尔士大学遥感专家 Jia Xiuping 博士对研究内容提出的宝贵建议。

参 考 文 献

[1] TAN Kun, DU Pei-jun(谭 琨, 杜培军). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2008, 28(9): 2009.

[2] Keshava N, Mustard J F. IEEE Signal Processing Magazine, 2002, 19(1): 44.

[3] Cipar J J, Eduardo M, Edward B. Proceedings of SPIE-The International Society for Optical Engineering, 2002, 4725: 1.

[4] Boardman J W, Kruse F A, Green R O. In Summaries of the V JPL Airborne Earth Science Workshop, Pasadena, CA. 1995.

[5] Winter M E. Proc. SPIE, Imaging Spectrometry, 1999, 3753: 266.

[6] Winter M E. Aerospace Proceedings, IEEE. Big Sky MT, United States, 18-25, March, 2000. 305.

[7] Tsang K Y, Grossmann J M. Proceedings of SPIE-The International Society for Optical Engineering, 1998, 3372: 43.

[8] Plaza A, Martínez P, Pérez R, et al. IEEE Transactions on Geoscience and Remote Sensing, 2002, 40: 2025.

[9] Plaza Antonio, Chang Chein-I. Proceedings of SPIE-The International Society for Optical Engineering, n PART I, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XI, 2005, 5806: 298.

[10] Wu Chao-Cheng, Chu Shihyu, Chang Chein-I. Proceedings of SPIE-The International Society for Optical Engineering, v 7086, Imaging Spectrometry XIII, 2008, 7086: 70860C.

[11] Chowdhury A, Alam M S. Proceedings of SPIE-The International Society for Optical Engineering, v 6565, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIII, 2007.

[12] Tao Xuetao, Wang Bin, Zhang Liming. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v 4681 LNCS, Advanced Intelligent Computing Theories and Applications: With Aspects of Theoretical and Methodological Issues-Third International Conference on Intelligent Computing, ICIC 2007, Proceedings, 2007, 4681: 1029.

[13] Wu Hsu-Kun, Chen Pao-Jung, et al. IEEE International Conference on Systems, Man and Cybernetics, Conference Proceedings, 2006,

Vol. 1-6: 5106.
[14] Suykens J A K, Brabanter J D, Lukas L and Vandewalle J. Neurocomputing, 2002, 48(1-4): 85.
[15] Vapnik V N. The Nature of Statistical Learning Theory. New York: Springer Press, NY, 1995.
[16] Evqeniou T, Pontil M. Machine Learning and Its Application. Advanced Lectures, Springer Publisher, 2005. 249.

Endmember Selection Algorithm Based on Linear Least Square Support Vector Machines

WANG Li-guo, DENG Lu-qun, ZHANG Jing
College of Information and Communications Engineering, Harbin Engineering University, Harbin 150001, China

Abstract Endmember (EM) selection is an important prerequisite task for mixed spectral analysis of hyperspectral imagery. In all kinds of EM selection methods, N-FINDR has been a popular one for its full automation and efficient performance. Unfortunately, the implementation of the algorithm needs dimensional reduction in original data, and the algorithm includes innumerable volume calculation. This leads to a low speed of the algorithm and so becomes a limitation to its applications. In the present paper, an improved N-FINDR algorithm was proposed based on linear least square support vector machines (LLSSVM), which is free of dimensional reduction and makes use of distance measure instead of volume evaluation to speed up the algorithm. Additionally, it was also proposed to endow the algorithm with robustness by controlling outliers. Experiments show that the computational load for EM selection using the improved N-FINDR algorithm based on LLSSVM was decreased greatly, and the selection effectiveness and the speed of the proposed algorithm were further improved by outlier removal and the pixel pre-sorting method respectively.

Keywords Hyperspectral imagery (HSI); Endmember selection; Linear least square support vector machines (LLSSVM); N-FINDR algorithm

(Received Jan. 20, 2009; accepted Apr. 25, 2009)