文章编号：1007-4619 (2003) 05-0400-07

# 超谱遥感图像快速聚类无损压缩算法

王朝晖，周佩玲
(中国科学技术大学 电子科学与技术系，合肥 230026)

**摘 要**： K-means 聚类要求每个像素要和所有聚类中心求欧氏距离，当聚类数很多时，这是一个相当耗时的工作。改进的 K-means 聚类算法根据历史聚类结果进行初始类分割，即节约初始聚类时间，又能使历史聚类过程中形成的类间稳定关系得以保持；类内像素只和相邻的聚类中心计算距离进行聚类，随着算法的迭代进行，大量类的状态基本固定，使得聚类速度不断加快。基于改进 K-means 聚类的无损压缩算法具有充分利用历史聚类成果和收敛速度快的特点，通过提高类内像素冗余度，最大限度消除谱间冗余和空间冗余。采用多次聚类压缩的结果预测最佳聚类数的方法，可实现最小熵无损压缩。通过和 DPCM 算法概率模型的熵值比较及实验数据的分析，验证了基于聚类无损压缩效率比不聚类无损压缩效果更优。

**关键词**： 超谱图像；无损压缩；熵；K-means 聚类
**中图分类号**： P751.1 **文件标识码**： A

## 1 引 言

超光谱遥感图像的问世是遥感技术的一大飞跃，由于它具有高的谱分辨力，使原先用多光谱信息不能解决的问题，在超谱下可以得到解决。然而这种具有较高谱分辨力的优越性是以较大的数据量及较高的数据维为代价的，如超谱 AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) 图像有 224 波段。因此，如何实现快速有效的压缩始终是超谱遥感信息处理的一个重要课题。

多光谱遥感图像无损压缩主要采用预测树[1]、聚类矢量量化[2]及 DPCM[2]等方法，通过设计预测器实现去冗余。近无损压缩通过 KL 变换[1]减少波段数，同时保留主要信息。文献[4]中提出的超谱图像压缩算法主要根据相邻波段的相关性最强的特点，通过对谱向残差数据的编码实现数据压缩。

本文根据遥感像素空间几何关系，对 K-means 聚类算法[3]进行了有效改进，提出基于改进的 K-means 聚类的无损压缩算法(Clustering Lossless Compression Algorithm 即 CLCA)。通过多次聚类预测最佳聚类数的方法，可实现最小熵无损压缩(Least Entropy Lossless Compression Algorithm 即 LELCA)。并将 CLCA 算法与 DPCM 无损压缩算法[4]按概率分布模型进行参数分析和零阶熵比较，最后结合 AVIRIS 超谱数据进行算法分析，从而验证了基于聚类的 CLCA 算法的优越性。

## 2 基于聚类的超谱遥感图像的无损压缩算法(CLCA)

K-means 聚类要求每个像素要和所有聚类中心求欧氏距离，并据此比较以确定像素的归属。当聚类数很多时，这是一个相当耗时的工作，其实有很多和像素相距较远的类，根本无须判断，而且很多类的状态稳定之后，就不用重新聚类；当调整聚类数并重新聚类时，初始聚类只需对类空间分布最不规则的类，按空间伸展方向进行分割。这使聚类收敛速度得到很大提高。

CLCA (Clustering Lossless Compression Algorithm) 根据改进的 K-means 聚类(Improved K-means algorithm 即 IKA)，实现较高聚类数时的均衡聚类；由于 IKA 聚类具有充分利用历史聚类成果和收敛速度快的特点，可多次采用 CLCA 算法确定最优聚类数和最高压缩比。

### 2.1 改进的 K-means 聚类算法(IKA)

#### 2.1.1 IKA 具体步骤

设 $N_s$ 是参加聚类的像素数；$b$ 表波段数；$n_c$ 为



**收稿日期**：2002-04-17；**修订日期**：2002-07-23
**基金项目**：安徽省重点项目(No. [2000]095)。
**作者简介**：王朝晖(1970— )，男，硕士，研究图像压缩与编码、人工智能、多媒体通讯。已发表论文 5 篇。

当前类分割数;$C_i$ 为类 $i$ 中的像素集合;$N(C_i)$表示类 $i$ 的像素个数;$e_j^i$ 为类 $i$ 中第 $j$ 个像素(即 $e_j^i \in C_i$);$e_{j1}^i$为像素 $e_j^i$ 第 1 波段的值;$c_l^i$ 为类 $i$ 聚类中心 $c_i$ 的第 $l$ 波段的值;$C_l^{ilast}$为上次迭代类 $i$ 聚类中心 $c_i$ 的第 $l$ 波段的值;$C_{ij}$表示象素 $e_j^i$ 的需重新聚类判断的类集合;$d_j^i$ 表示类内像素 $e_j^i \in C_i$ 到聚类中心 $c_i$ 的距离;$P[\mathrm{i}]$和 $\mathrm{E}(\mathrm{d}^i)$分别是$\{ d_j^i \mid e_j^i \in C_i \}$ 的最大值和均值(即数组 $P[\mathrm{i}](\mathrm{i} \in [0, n_c]$是类 i 内像素偏离中心的最大值,并记该像素为 $\mathrm{P}_i$);$d_{ij}$表示类 $i$ 和类 $j$ 的聚类中心间的距离;$d_{jk}^i$ 表示像素 $e_j^i$ 到类 $k$ 的聚类中心的距离;*lastsign*[i]是类 i 的上次迭代修改标志,用于判断类间稳定关系;sign[i]是类 i 本次迭代修改标志,用于判定聚类收敛性;ε 是聚类中心最大迭代误差。聚 $\mathrm{N}_c$ 个类的 IKA 算法步骤如下:

1) 类空间初始分割 $N_c$ 个类,产生初始聚类中心 $c_i(i \in [0, N_c-1]$。

1.1) 若存在历史记录(聚类数 $N_c'$小于 $N_c$),则读出并重新构造上次聚类结果,并取 $n_c = N_c'$;否则,所有象素归为一类,使 $n_c=1$,则第一个类空间大小为 $N_s$ 个像素,并求聚类中心 $c_0$,计算类内像素 $e_j^0$ 到聚类中心的距离 $d_j^0$,取 $d_j^0$ 最大值为 $P[0]$,相对应的点为 $\mathrm{P}_0$,使 *lastsign*[0]=1。

1.2) 由图 1a 和式(1),对数组 $P[\mathrm{J}](\mathrm{J} \in [0, n_c-1]$中最大值对相应的类 i 和该类中偏离中心最远的像素 $\mathrm{e}_m^i$(记为 $P_i$),按方向 $P_iC_i = e_m{}^i - c_i$,并以通过该类中心 $c_i$ 和最远点 $e_m^i$ 中点的垂直平面分割该类,产生两子类 $C_i'$($C'$取代 $C_i$ 的位置,使 $C_i = C_i'$)和 $C_{nc}$,置 *lastsign*[i]=*lastsign*[$\mathrm{n}_c$]=1,每个子类分配空间大小为 $N(C_i)$和 $N(C_{nc})$,使 $n_c = n_c+1$,对新生的两子类分别进行如下计算(以新类 $C_i$ 为例):求类 $C_i$ 的聚类中心 $c_i$(式(2)),计算类内像素 $e_j^i \in C_i$ 到聚类中心的距离 $d_j^i$,$P[\mathrm{i}]$和 $\mathrm{E}(\mathrm{d}^i)$分别是 $\{ d_j^i \mid e_j^i \in C_i \}$ 的最大值和均值。

$$I = \frac{\overline{P_i}\,\overline{C_i}}{\| \overline{P_i}\,\overline{C_i} \|} \tag{1}$$

$$\forall\, \mathrm{e}_j^i \in C_i, j \neq m, \text{if } d_j^i < d_m^i - \frac{d_m^i + E(d^i)}{2} \Rightarrow C_i' = C'_i U \{ e_j^i \}$$

$$\text{else if } |(e_j^i - e_m^i) \cdot I| < \frac{d_m^i + E(d^i)}{2} \Rightarrow C_{n_c} = C_{n_c} U \{ e_j^i \}$$

$$\text{else } C'_i = C'_i U \{ e_j^i \}$$

$$c_l^i = \frac{1}{N(C_i)} \sum_{j=1}^{N(C_i)} e_{jl}^i, 1 \leqslant \mathrm{l} \leqslant \mathrm{b}, \mathrm{e}_j^i \in \mathrm{C}_i \tag{2}$$

$$d_j^i = \Big[ \sum_{l=1}^{b} \Big( e_{jl}^i - c_{jl}^i \Big)^2 \Big]^{1/2}$$

1.3) 若获得指定的类数 $n_c = N_c$ 时,则停止分割,转 2);否则转 1.2),继续分割操作。

2) 求各类中心之间的距离,产生聚类中心间距离矩阵 $D = \{ d_{ij}, i, j \in [0, N_c-1] \}$,记录上次聚类中心值 $c_{ilast} = c_i$,并置当前类修改标志 *sign*[i]=0($i \in [0, N_c-1]$)。

$$d_{ij} = \Big[ \sum_{k=1}^{b} (c_{ik} - c_{jk})^2 \Big]^{1/2}, i \neq \mathrm{j}, \mathrm{d}_{ij} \in \mathrm{D}$$

3) 各类中的像素和相邻聚类中心进行重新聚类(参图 1b):

3.1) 每一类 $i \in [0, N_c-1]$进行如下处理:将 $\mathrm{d}_{ij}(i \neq j, j \in [0, N_c-1]$按由小到大排序,确定其他类 j 和当前类 i 的远近程度。

3.2) 对类 $i$ 中每一像素 $e_j^i \in C_i$ 进行如下处理:

3.2.1) 按由近到远的顺序,确定搜索范围 $C_{ij}$(对 $k \in C_{ij}$,有 $k \neq i, k \in [0, N_c-1]$, $\mathrm{d}_j^i > 0.5 d_{ik}$)。

3.2.2) 若上次修改标志 *lastsign*[i]=0 且 *lastsign*[k]=0($k \in C_{ij}$),则类 $j$ 和类 $i$ 的相对关系已经确定,不用进行聚类判断,将 $k$ 从 $C_{ij}$中删除。

3.2.3) 求 $e_j^i$ 与所有邻近类 $k \in C_{ij}$的聚类中心的距离 $d_{jk}^i$,若$\{ \{ d_{jk}^i \mid k \in C_{ij} \}, d_{jj}^i \}$ 中的最小值对应的类 $l \neq i$,则置类修改标志 *sign*[i]=1, *sign*[l]=1,并修改类集合:$C_l = C_l - \{ e_{jj}^i \}$, $C_1 = C_1 + \{ e_{jj}^i \}$。

$$d_{jk}^i = \Big[ \sum_{l=1}^{b} \Big( e_{jl}^i - c_l^k \Big)^2 \Big]^{1/2}, k \in \mathrm{C}_{ij}$$

4) 更新上次类修改标志 *lastsign*[i]=*sign*[i]($i \in [0, N_c-1]$)。

5) 对所有修改标志 *sign*[i]为 1 的类 $i \in [0, N_c-1]$,重新计算 $\mathrm{c}_i$ 及 $d_j^i(e_j^i \in C_i)$,若$|c_l^{ilast} - c_l^i| < \varepsilon(l \in [1, b])$,则该类状态稳定,置 sign[i]=0。

6) 若所有 *sign*[i]=0($i \in [0, N_c-1]$),或迭代足够的次数,则聚类结束,转 7);否则转 2)。

7) 保存本次聚类结果。

### 2.1.2 IKA 聚类特点

(1) 根据历史聚类结果进行初始聚类,对与类内最大偏差值 $P[\,]$中的最大值相对应的类进行分割,即节约初始聚类时间,又能使历史聚类过程中形成的类间稳定关系得以保持。

(2) 由于只对类之间相对关系有变化的类进行重新聚类,并且类内像素只和相邻的聚类中心计算距离进行聚类,随着算法的迭代进行,大量类的状态基本固定,使得聚类速度不断加快。

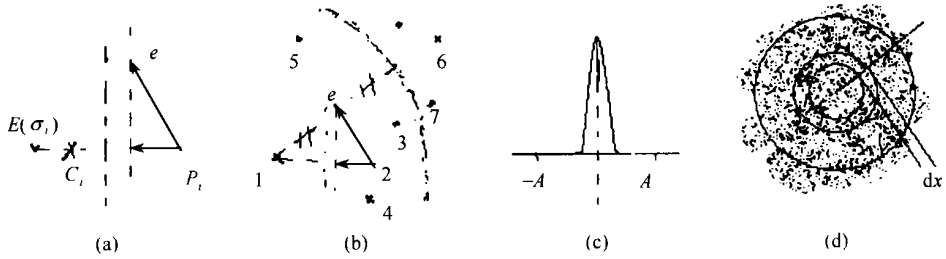


图 1 IKA 聚类示意图:(a) $C_i$ 类分割(b)类 1 内像素 $e$ 的重新聚类(c)谱间残差数据概率分布示意图(d)析限情况下类内数据呈空间均匀分布

Fig 1 IKA clustering sketch map: (a)Division of class $C_i$, (b)Re-clustering of the pixel 'e' of class 1, (c)The sketch map of the intra-spectral residue data. (d)Space uniformity distribution of intra-class data under extreme division condition.

## 2.2 CLCA 算法步骤

基于聚类的 CLCA 算法在适于高聚类数的 IKA 快速聚类的基础上,可最大限度提高类内像素冗余度,通过残差数据不定长编码实现无损压缩。由于相邻波段的相关性最强,聚类中心可以根据相邻波段的残差消除谱间冗余;类内像素的冗余通过和聚类中心的残差来消除;然后对残差数据和分类地图分别进行 huffman 统计编码。

设 $m$、$n$、$b$ 为遥感图像的长、宽和波段数,$X^i_{j,k}$ 为第 $i$ 类中第 $j$ 个样本的第 $k$ 波段的数据,$I^i_k$ 为第 $i$ 聚类中心第 $k$ 波段数据,$N(i)$ 为第 $i$ 类的样本数,$s_{ij}$ 为像素 $(i,j)$(像素空间位置坐标)的类序号,CLCA 具体步骤如下:

1) 对超谱遥感图像进行 IKA 聚类,产生 $N_c$ 个聚类中心 $I_i$,$i\in[0,N_c-1]$(取和原聚类中心最近的类内像素作为该类的聚类中心,这比直接取整或四舍五入的误差更小;使聚类中心序号根据第一波段幅值大小进行适当排序,以增强冗余度)和分类地图 $S=\{s_{ij}\in[0,N_c-1],\mathrm{i}\in[0,m-1],\mathrm{j}\in[0,n-1]\}$,并对 $S$ 进行 huffman 统计编码。

2) 聚类中心的编码:

2a. 求每个聚类中心的谱向残差:$\mathrm{d}I^i_k=I^i_k-I^i_{k-1}$,$i\in[0,N_c-1]$,$\mathrm{k}\in[1,b-1]$。

2b. 求各类第一维数据的残差:$\mathrm{d}I^i_0=I^i_0-I^{i-1}_0$,$i\in[1,N_c-1]$。

2c. 对残差数据 $\mathrm{d}I$ 进行 huffman 统计编码,$I^0_0$ 单独保存。

3) 类内像素的编码:

3a. 求各类内样本和聚类中心的残差:$\mathrm{d}X^i_{j,k}=X^i_{j,k}-I^i_k$,$i\in[0,N_c-1]$ $\mathrm{j}\in[0,N(i)-1]$,$\mathrm{k}\in[0,b-1]$。

3b. 对残差数据 $\mathrm{d}X$ 进行 huffman 统计编码。

上述算法中 1)和 2)两步可实现矢量量化的有损压缩,压缩效率将达到数百倍;另外,如果将波段进行适当重排,增加波段间的冗余,可使聚类中心得到进一步的压缩。利用历史聚类成果,快速实现不同聚类数时的压缩率,从而确定最优聚类数。

## 2.3 CLCA 算法概率模型理论分析

设像素数为 $N_s$,则对 CLCA 算法的残差数据概率分布模型进行如下假设和分析。

**2.3.1** 当聚类数达到一定数量时,类内残差数据为 $X=\{0,\pm1,\pm2,\cdots\}$,变量 $x\in X$ 近似呈 $N(0,\sigma_1^2)$ 的 Gauss 分布,均方值受限的最大熵[5]为 $H_1=\log_2(\sigma_1\sqrt{2\pi e})$,则类内像素残差总比特数为 $B_1=N_sbH_1=N_sb\log_2(\sigma_1\sqrt{2\pi e})$。

**2.3.2** 聚类中心残差数据编码:

设 $A=(f_{\max}-f_{\min})$($f_{\max}$、$f_{\min}$)分别表示所有象素幅值的最大和最小值),则中间波段数愈多,中心向量残差 $x\in[-A,A]$,呈 $\mathrm{N}(0,\sigma_2^2)$ 的 Gauss 分布(见图 1(c)),最大熵为 $H_2=\log_2(\sigma_2\sqrt{2\pi e})$,总比特数为 $B_2=N_cbH_2=N_cb\log_2(\sigma_2\sqrt{2\pi e})$。

**2.3.3** 分类地图编码:

分类地图数据 $X=\{0,1,2,\cdots,N_c\}$ 呈均匀分布,概率密度为 $p_3(x)=\frac{1}{N_c}$,熵为 $H_3=\log_2N_c$,总比特数为 $B_3=N_sH_3=N_s\log_2N_c$。



**2.3.4** CLCA 算法的总比特数:

$$B=B_1+B_2+B_3=N_sb\log_2\sigma_1+N_cb\log_2(\sigma_2\sqrt{2\pi e})$$

$$+N_s\log_2 N_c+N_s b\log_2\sqrt{2\pi e} \tag{3}$$

$B_3$ 与聚类数 $N_c$ 成对数关系，$N_c$ 的变化对 $B_3$ 影响不大。当类数较少时，$\sigma_1$ 随聚类数 $N_c$ 的增加而减少，使 $B_1$ 的 Huffman 编码效率提高，且 $B_1$ 减少的幅度远大于 $B_2+B_3$ 增加的幅度，从而提高无损压缩比。

### 2.4 最小熵无损压缩算法(LELCA)

LELCA (Least Entropy Lossless Compression Algorithm)即求 CLCA 无损压缩最少比特数时的聚类数。设若 $N_c$ 增加 $m$ 倍，$(N_c,\sigma_1)$变为$(N_c',\sigma'_1)$则聚类数 $N_c$ 和类内像素标准差 $\sigma_1$ 存在式(4)关系，并建立微分方程(式(5))，由三个初始条件可确定待定系数 $k$ 和 $x$。

$$\begin{cases}\sigma'_1=\dfrac{1}{km^x}\sigma_1\\ N'_c=mN_c\end{cases} \tag{4}$$

$$\Rightarrow x(\ln N'_c-\ln N_c)=\ln\sigma_1-\ln\sigma'_1-\ln k$$

$$\Rightarrow\begin{cases}x d(\ln N_c)=-d(\ln\sigma_1)-\ln k\\ N_c\mid_{\sigma_1=\sigma_{1n_1}}=n_1\\ N_c\mid_{\sigma_1=\sigma_{1n_2}}=n_2\\ N_c\mid_{\sigma_1=\sigma_{1n_3}}=n_3\end{cases} \tag{5}$$

$$\Rightarrow N_c=\left(\frac{k^2}{\sigma_1 k^{\sigma_1}}\right)^{1/x} \tag{6}$$

由初始条件有：

$$\ln N_c-\ln n_1=(\ln n_2-\ln n_1)\frac{\ln\sigma_{1n_1}-\ln\sigma_1+(\sigma_{1n_1}-\sigma_1)\ln k}{\ln\sigma_{1n_1}-\ln\sigma_{1n_2}+(\sigma_{1n_1}-\sigma_{1n_2})\ln k} \tag{7}$$

其中，

$$\ln k=\frac{(\ln n_3-\ln n_1)\left(\frac{\ln\sigma_{1n_1}-\ln\sigma_{1n_2}}{\sigma_{1n_1}-\sigma_{1n_2}}\right)-(\ln n_2-\ln n_1)\left(\frac{\ln\sigma_{1n_1}-\ln\sigma_{1n_3}}{\sigma_{1n_1}-\sigma_{1n_3}}\right)}{-(\ln n_3-\ln n_1)+(\ln n_2-\ln n_1)}$$

则：

$$\frac{d\sigma_1}{dN_c}=-\frac{\ln\sigma_{1n_1}-\ln\sigma_{1n_2}+(\sigma_{1n_1}-\sigma_{1n_2})\ln k}{N_c(\ln n_2-\ln n_1)}\times\frac{\sigma_1}{1+\sigma_1\ln k}$$

令 $dB/dN_c=0$，则：$\dfrac{dB}{dN_c}=\dfrac{N_s b}{\sigma_1\ln 2}\dfrac{d\sigma_1}{dN_c}+b\log_2(\sigma_2\sqrt{2\pi e})+\dfrac{N_s}{N_c\ln 2}=0$

$$\Rightarrow N_{c\text{best}}=\frac{N_s b[\ln\sigma_{1n_1}-\ln\sigma_{1n_2}+(\sigma_{1n_1}-\sigma_{1n_2})\times\ln k]-N_s(\ln n_2-\ln n_1)\times(1+\sigma_1\times\ln k)}{b\ln(\sigma_2\sqrt{2\pi e})\times(\ln n_2-\ln n_1)\times(1+\sigma_1\times\ln k)} \tag{8}$$

LELCA 算法通过多次聚类的方法预测 $N_{c\text{best}}$：

1) 取初始三个类数：$n_1$、$n_2$、$n_3$，通过 CLCA 求出对应的 $\sigma_{\ln 1}$、$\sigma_{\ln 2}$、$\sigma_{\ln 3}$。

2) 由式(8)预测方程求出 $N_{c\text{best}}$，并通过 CLCA 求出对应的 $\sigma_{1\text{best}}$ 及相应 Huffman 编码字节长度 $L_{\text{best}}$。

### 2.5 类分割极限情况

当对呈 Gauss 分布的类内数据继续分割，则更多的类内像素呈空间均匀分布(图 1(d))，注意类内残差数据 $x\in[0,R]$并不呈均匀分布，对 b 维 $V_b$ 空间，其分布概率密度为 $p_1(x)$，熵为 $H_1$。

$$p_1(x)=\frac{bx^{b-1}}{R^b}$$

$$H_1=-\int_0^R\frac{bx^{b-1}}{R^b}\log\left(\frac{bx^{b-1}}{R^b}\right)dx=\frac{1}{\ln 2}\left(\ln R+\frac{b-1}{b}\ln b\right)$$

若 $R\approx\sigma_1$，和 2.3.1 中 $H_1$ 的形式相差不大，设当聚类数达到 $N_{c0}$(此时 $\sigma_1=\sigma_{10}$)，若 $\sigma_1$ 减少至 $1/m$ 倍，则 $\sigma'_1=\sigma_1/m$，$N_{c0}'=m^bN_{c0}$，聚类数 $N_c$ 和类内像素标准差 $\sigma_1$ 存在式(9)关系。

$$\left(\frac{N'_{c_0}}{N_{c_0}}\right)^{1/b}=m=\frac{\sigma_{10}}{\sigma'_{10}} \tag{9}$$

$$\begin{cases}\dfrac{1}{b}d\ln N_c=-d\ln\sigma_1\\ N_c\mid_{\sigma_1=\sigma_{10}}=N_{c_0}\end{cases}\Rightarrow N_c=\frac{k}{\sigma_1^b},$$

由初始条件有 $k=N_{c_0}\sigma_{1_0}^b$，则：$N_c=\dfrac{N_{c_0}\sigma_{1_0}^b}{\sigma_1^b}$

$B_1+B_3=[N_s\log_2\sigma_1^b+N_s(b-1-b\ln b)/\ln 2]+N_s\log_2 N_c=N_s\log_2(N_{c_0}\sigma_{1_0}^b)+N_s(b-1-b\ln b)/\ln 2$ 为常数，则：

$$B=[N_s\log_2(N_{c_0}\sigma_{1_0}^b)+N_s(b-1-b\ln b)/\ln 2]+N_c b\log_2(\sigma_2\sqrt{2\pi e})$$

此时总熵 $B$ 由 $B_2$ 决定，每增加一个类，$B$ 增加 $b\log_2(\sigma_2\sqrt{2\pi e})$ (bit)

## 3 DPCM/D²PCM 无损压缩算法

根据文献[4],超谱图像 DPCM 无损压缩算法的具体步骤为:

1) 求每个像素的谱向残差:$\mathrm{d}I_k = I_k - I_{k-1}$, $k \in [1, b-1]$。

2) 每一波段数据 $S+P$ 整数小波变换[6]消除空间冗余。

3) 对残差数据 $\mathrm{d}I_k(k \in [1, b-1])$进行 Huffman 统计编码,$I(0,0)$值单独保存。

$D^2$PCM 对谱向数据求二次残差,对上述步骤修正为:

1)中加上 $\mathrm{d}^2 I_k = \mathrm{d}I_k - \mathrm{d}I_{k-1}$, $k \in [2, b-1]$,

3)中的残差数据为 $\mathrm{d}I_1, \mathrm{d}^2 I_k(k \in [2, b-1])$。

### 3.1 概率分布模型

**3.1.1** 每个像素谱向残差 $x \in [-A, A]$呈 $N(0, \sigma_1'^2)$的 Gauss 分布[5](见图 1(c)),熵为

$$H'_1 = \log_2\left(\sigma'_1\sqrt{2\pi e}\right),$$

总比特数为

$$B'_1 = N_s(b-1)H'_1 = N_s(b-1)\log_2\left(\sigma'_1\sqrt{2\pi e}\right)$$

**3.1.2** 第一波段残差数据为 $X = \{0, \pm1, \pm2, \cdots\}$变量 $x \in X$ 呈 $N(0, \sigma_2'^2)$的 Gauss 分布,则类内像素残差总比特数为 $B'_2 = N_s H'_2 = N_s\log_2\left(\sigma'_2\sqrt{2\pi e}\right)$。

**3.1.3** DPCM/$D^2$PCM 算法的总比特数:

$$B' = B_1' + B_2' = N_s(b-1)\log_2\left(\sigma'_1\sqrt{2\pi e}\right) + N_s\log_2\left(\sigma'_2\sqrt{2\pi e}\right) \quad (10)$$

## 4 算法比较

CLCA 和 DPCM 两种无损压缩算法的编码数差异为 $\Delta B$(式(11)),一般情况下 $\sigma_1 < \sigma'_1$ 即 $\Delta B > 0$;只有在波段数 $b$ 极高的情况下,才有可能 $\sigma_1 > \sigma'_1$,但目前的遥感器达不到这样的要求。因此 CLCA 算法比 DPCM 算法高效。

$$\begin{aligned}\Delta B = B' - B &\approx N_s(b-1)\log_2\left(\sigma'_1\sqrt{2\pi e}\right) - N_s b\log_2\left(\sigma_1\sqrt{2\pi e}\right)\\ &\approx N_s b\log_2\left(\frac{\sigma'_1}{\sigma_1}\right) > 0\end{aligned} \quad (11)$$

## 5 实验仿真

选取 224 波段的 Sook Lake (256 * 256, 16bit) AVIRIS 超谱图像(网上获取,图 2(a)),删去 1,3,89,136 等 4 个零波段,由余下的 220 波段进行算法研究,则有 $N_s = 65536$, $b = 220$, $A = (f_{\max} - f_{\min}) = 37810 - 30991 = 6819$,原始数据量为 256 * 256 * 220 * 2=28835840 字节,实验数据分析中,按式(12)定义压缩率。

$$\text{压缩率} = \frac{\text{原数据量}}{\text{压缩后数据量}} \quad (12)$$

三种压缩算法的实验数据如下:

(1) CLCA 算法

采用 IKA 聚类算法,取 $\varepsilon = 0.01$,则当 $N_c = 100$ 时(见表 1),迭代 214 次收敛,无损压缩率为 2.3075,平均每次迭代用时 8s,图 2(b)是由聚 100 类后的聚类中心,对类内像素统一表示,并经灰度均衡后的结果。

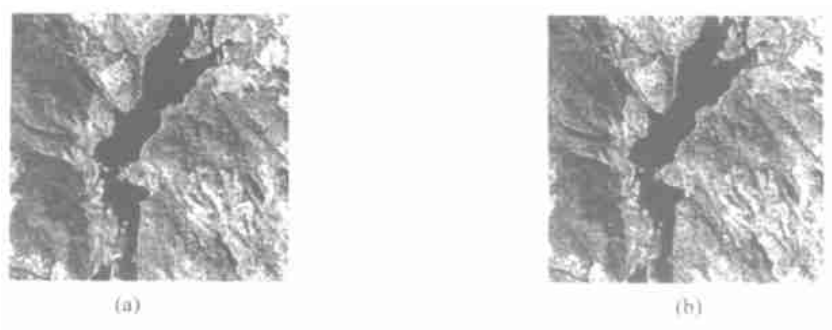


图 2 Sook Lake AVIRIS 超谱图像有损重建效果比较(a)原第 60 波段经灰度均衡后结果(b)聚 100 类后,由聚类中心量化并经灰度均衡后的第 60 波段图像

Fig.2 Sook Lake AVIRIS hyperspectral images rebuilding results comparison: (a) The 60th gray balanced band image (b) The 60th gray balanced band image rebuilt with centers clustering with 100 classes

**表 1　CLCA 和 LELCA 无损压缩结果比较**

**Table 1　Compresion results comparison of CLCA and LELCA**

| 聚类数 $N_c$ | 无损压缩字节数 $L$ | 类内残差标准差 $\sigma_1$ | 谱间残差标准差 $\sigma_2$ | $N_{cbest}$ | 无损压缩率 |
|---|---|---|---|---|---|
| 100 | 12496332 | 40.892593496802 | | | 2.3075 |
| 200 | 12342768 | 37.611525938468 | | | 2.3363 |
| 300 | 12256192 | 35.878427593448 | 183.76822461419 | 1175 | 2.3528 |
| 1175 | 12115636 | 31.250165028639 | 182.70978456112 | 1176 | 2.3801 |
| 1176 | 12115667 | 31.236940162119 | 183.01692705276 | | 2.3800 |
| 1170 | 12111931 | 31.258761693421 | 183.19476179036 | | 2.3808 |
| 1100 | 12112331 | 31.368733371511 | 183.57650159557 | | 2.3807 |
| 1130 | 12110633 | 31.274657976381 | 183.16136094010 | | 2.3810 |

取 $N_c=200$,则初始聚类在 $N_c=100$ 聚类成果的基础上,进行进一步的分割,初始类分割仅 3s,迭代 210 次收敛,平均每次迭代用时 9s,无损压缩率达到 2.3363。可见聚类数的增加对聚类迭代的速度影响不大,而原 K-means 迭代一次作时 6m(100 类),且聚类时间随聚类数的增多急剧上升。

图 3 是取每次聚类的迭代次数为 20,充分利用历史聚类成果,快速实现各聚类数时的无损压缩:图 3(a)中分类地图编码 $B_3$ 与 $N_c$ 近似为对数关系;聚类中心残差 $B_2$ 和 $N_c$ 近似为线性关系;图 3(b)中类内数据残差 $B_1$ 和 $N_c$ 近似为对数关系,符合概率模型的假设。

(2) LELCA 算法

由表 1 ,取 $n_1=100$, $n_2=200$, $n_3=300$ 作为输入参数,则由式(8)可得:$N_{cbest}=1175$,CLCA 循环 136 次收敛,此时无损压缩率为 2.3801。

取 $n_1=200$, $n_2=300$, $n_3=1175$ 作为输入参数,再次由式(8)预测方程求出 $N_{cbest}=1176$,CLCA 循环 16 次收敛(是在 $N'_c=1175$ 聚类收敛的基础上进行进一步分割),此时 $L_{1176}=12115667>L_{1175}$。

另外由表 1 可知,$L_{1175}>L_{1100}=12112331>L_{1170}$,因此实际的 $N_{cbest}\in[1100,1170]$,取 $N_{cbest}\approx 1130$,此时 $L_{1130}=12110633$。可见根据式(8)预测的最小熵聚类数 $N_{cbest}$ 和实验测试的最佳聚类数相差不大。

(3) DPCM/D$^2$PCM 算法:

整数小波变换系数[6]取 $\alpha_{-1}=\beta_1=0$, $\alpha_0=\alpha_1=0.25$,用 DPCM 则有 $\sigma'_1=177$, $B'=14048912$ 字节,无损压缩率为 2.0525。因此 CLCA 算法比 DPCM 算法高效。而 D$^2$PCM 的 $\sigma'_1=190.957$,无损压缩率仅为 1.9963,可见谱向二次残差并不能提高无损压缩率,表 2 说明,LELCA 压缩算法效果最优,但耗时很长。

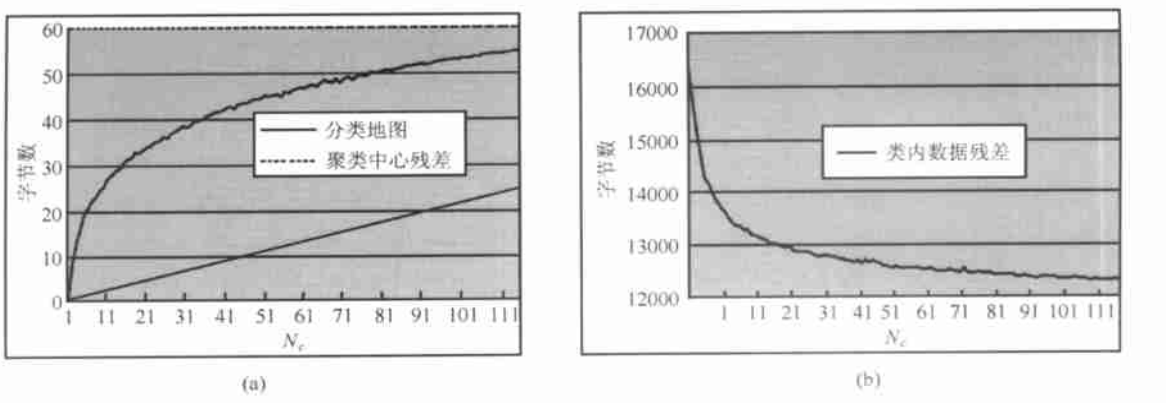


图 3　Sook Lake AVIRIS 图像压缩参数(纵坐标以千为单位,横坐标为聚类数):(a)聚类中心和分类地图,(b)类内残差编码

Fig 3　Sook Lake AVIRIS images compression parameters(vertical coordination is scaled by thousandth, horizontal coordination indicates class number):(a)clustering centers and classification map,(b)inter-class residue dada

表 2 无损压缩结果比较

Table 2 Compression results comparison

| 算法 | $D^2$PCM | DPCM | CLCA(200 类) | LELCA(1130 类) |
|---|---|---|---|---|
| 无损压缩率 | 1.9963 | 2.0525 | 2.3363 | 2.3810 |

## 6 结 论

基于聚类的无损压缩算法多次采用改进的 K-means聚类确定最优聚类数,充分利用历史聚类结果,有效解决初始聚类问题;同时又利用空间几何关系,像素只和近邻聚类中心进行聚类判断,明显提高迭代收敛速度。本文还通过残差数据概率分布模型的理论分析,论证了基于聚类的无损压缩算法比 DPCM 算法压缩效率更高。

## 参 考 文 献 (References)


[1] Zhang Rogn, Liu Zheng-kai. A Near-lossless Compression Technique of Multispectral Image Data[J]. *Journal of Image and Graphics*. 1998, **3**(10): 823－826. [张荣, 刘政凯. 一种多光谱遥感图象的近无损压缩方法[J]. 中国图像图形学报. 1998, **3**(10): 823－836.]

[2] Zhang Rogn, Liu Zheng-kai, Li Hou-qiang. Classification-based Lossless Compression of Multispectral Date[J]. *Journal of Image and Graphics*. 1998, **3**(2): 106－109. [张荣, 刘政凯, 李厚强, 基于分类的多波段遥感图像无损压缩方法[J]. 中国图像图形学报. 1998, **3**(3): 106－109.]

[3] Zhu Shu-long, Zhang Zhan-mu. Acquisition and Analysis of Remotely Sensed Data[M]. Beijing: Science and Technology Publishing Company, 2000. 162－165 [朱述龙, 张占睦. 遥感图像获取与分析[M]. 北京: 科学技术出版社, 2000. 162－165.]

[4] Wang Xue-liang. Lossless $D^2$PCM Compression Algorithm Based on Correlation of HRIS Spectral Image Sequence[J]. *Journal of Remote Sensing*, 2001, **5**(2): 119－121. [王学良. 基于 HRIS 光谱图像帧序列相关性的 $D^2$PCM 无损压缩方法[J]. 遥感学报, 2001, **5**(2): 119－121.]

[5] Cao Zhi-gang, Qian ya-sheng. Principles of Modern Communication [M]. Beijing: Tsinghua University Publishing Company, 1999. 11－26. [曹志刚, 钱亚生. 理代通信原理[M]. 北京: 清华大学出版社, 1999. 11－26.]

[6] Tang Yan, Mo Yu-long. Second Generation Wavelet Transform Applied to Lossless compression Coding of Image[J]. *Journal of Image and Graphics*, 2000, **5A**(8): 699－702. [汤炎, 莫玉龙. 第二代小波变换应用于图像的无损压缩编码[J]. 中国图像图形学报, 2000, **5A**(8): 699－702.]


# Fast Clustering Lossless Compression Algorithm for Hyperspectral Images


WANG Zhao-hui, ZHOU Pei-ling

(*Department of Electronic Science and Technology, USTC, Hefei* 230026, *China*)



**Abstract**: Every pixel in the super space is required by K-means algorithm to calculate Euclidean distance for clustering. When there are many class centers, this is a rather time consuming work. In this paper, an improved K-means clustering algorithm is presented to save initial clustering time by making initial division based on previous clustering results, to remain the stable relationship between classes, and to accelerate clustering process with more and more classes becoming stable by judging the centers nearest to the pixel. A new clustering lossless compression algorithm designed here can determine the best class number and the highest compression ratio by fully utilizing previous clustering results and converging quickly eliminating the inter-spectral redundancy and intra-intra-spectral redundancy through enhancing the intra-class pixel redundancy. The convergence of this algorithm and existence of the best parameters are also inferred by making a deep analysis of the probability distribution model of the residue data. Furthermore, the comparison with DPCM lossless compression algorithm in the entropy value of the probability distribution model and the experimental results show that this clustering algorithm is better than non-clustering compression algorithm. Several times clustering approach can forecast the best class number with the least entropy lossless compression.



**Key words**: Hyperspectral images; Lossless compression; Entropy; K-means clustering