有序样品的最优分割算法及其在 Matlab 中的实现

一、 有序样品聚类——最优分割的概念

地质数据中，有些样品有一定的排列顺序，如沿地层剖面采集的岩石标本，由钻孔取得的岩芯样品，由测井曲线所得的数据，由岩体中心到围岩的蚀变剖面的样品等，它们是有序地质变量，在对这些有序样品进行分类时，不能打乱样品的前后次序。所以， 一些不考虑样品排列顺序的数学处理方法，对此并不适用。有序样品的聚类分析就是对有序样品进行分段的统计方法。对 n 个有序样品进行分割，就可能有 2^{n-1} 种划分方法，这每一种分法成为一种分割。在所有的这些分割中，有一种分割使得各段内部之间差异性最小，而短语段之间差异性最大。这种对 n 个样品分段并使组内离差平方和最小的分割方法，成为最优分割法。

这类问题的提法如下：

设有一批（N 个）按一定顺序排列的样品，每个样品测得 p 项指标，其原始资料矩阵:

$$X（p\times N）=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pN} \end{bmatrix}$$

其中元素 x_{ij} 表示第 j 个样品的第 i 个指标的观测值。现在要把此 N 个样品按顺序（不破坏序列的连续性）进行分割（分段或者分类）。其所有可能的分割法共有

$$C^{1}_{N-1}+C^{2}_{N-1}+C^{3}_{N-1}+\ldots+C^{N-1}_{N-1}=2^{N-1}-1$$

种。现在要求在所有分割中找出一种分割法，这种分割法使得各段内样品之间的差异最小，而各段之间的差异最大。

各段内部差异最小，即各段内数值变化最小，段内数值变化可用变差或者极差来表示，比如样品段｛x_i、x_{i+1}、x_{i+2}、…、x_j｝:

变差：

$$d_{ij}=\sum_{\alpha=i}^{j}[x_\alpha-\bar{x}(i,j)]^2$$

$$\bar{x}(i,j)=\frac{1}{j-i+1}\sum_{\alpha=1}^{j}x_\alpha$$

d_{ij} 表示样品段｛x_i、x_{i+1}、x_{i+2}、…、x_j｝内样品间的差异情况，d_{ij} 小表示段内各样品之间数值比较接近，反之，d_{ij} 大表示段内各样品数值之间的差异大。

极差：

$$d_{ij}=\sum_{\alpha=1}^{p}(\max_{i\le\beta\le j}x_{\alpha\beta}-\min_{i\le\beta\le j}x_{\alpha\beta})$$

对于单指标情况，则

$$d_{ij}=(\max_{i\le\beta\le j} x_\beta - \min_{i\le\beta\le j} x_\beta)$$

要各段内部的差异最小，即所分成各段变差的总和（即段内离差平方和，称为总变差）为最小。

总变差分解公式：

$$S_{总}=S_{段间}+S_{段内}$$

$$S_{总} = \sum_{l=1}^{m} \sum_{j=1}^{n_l}(x_{jl} - \bar{x})^2$$

$$=\sum_{l=1}^{m} \sum_{j=1}^{n_l}[(x_{jl} - \bar{x}_l) + (\bar{x}_l - \bar{x})]^2$$

$$=\sum_{l=1}^{m} \sum_{j=1}^{n_l}(x_{jl} - \bar{x}_l)^2 + \sum_{l=1}^{m} \sum_{j=1}^{n_l}(\bar{x}_l - \bar{x})^2 + 2\sum_{l=1}^{m} \sum_{j=1}^{n_l}(x_{ij} - \bar{x}_l)(\bar{x}_l - \bar{x})$$

$$=\sum_{l=1}^{m} \sum_{j=1}^{n_l}(x_{jl} - \bar{x}_l)^2 + \sum_{l=1}^{m} \sum_{j=1}^{n_l}(\bar{x}_l - \bar{x})^2$$

$$=S_{段内}+S_{段间}$$

其中：

$$D = 2\sum_{l=1}^{m} \sum_{j=1}^{n_l}(x_{ij} - \bar{x}_l)(\bar{x}_l - \bar{x})$$

$$= 2\sum_{l=1}^{m}(x_{ij} - \bar{x}_l)\sum_{j=1}^{n_l}(\bar{x}_l - \bar{x})$$

$$= 2\sum_{l=1}^{m}(\bar{x}_l - \bar{x})(n_l\bar{x}_l - n_l\bar{x}_l)$$

$$= 0$$

$S_{段内} = \sum_{l=1}^{m} \sum_{j=1}^{n_l}(x_{jl} - \bar{x}_l)^2$ 为段内离差平方和

$S_{段间} = \sum_{l=1}^{m} \sum_{j=1}^{n_l}(\bar{x}_l - \bar{x})^2$ 为段间离差平方和

所以

$$S_{段间}=S_{总}-S_{段内}$$

对给定的N个样品，$S_{总}$是个固定的量。若使段内离差平方和为最小，则段间离差平方和必为最大。所以，使总变差（段内离差平方和）为最小的分割法就是最优的分割法。

二、 有序样品聚类的最优分割意义

最优分割在地质研究中是一个非常有用的手段，只要地质体的某些地球化学特征存在规律性的差异，采用最优分割的数学树立方法，就能按顺序在最理想的地方进行分段。通过对地层中采集的具某些地球化学特征样品的最优分割，能在地层的划分对比中发挥重要的辅助作用；在找矿过程中，该方法更显得天独厚的优势，它能进行蚀变、矿化及矿体的准确分带；对岩浆岩相带划分及演化序列的研究也十分有效。第四纪地层岩相和厚度变化很大，给第四纪地层划分和对比地层带来了困难，但第四系地层是不同阶段和不同的环境条件下形成的，如重矿物、微量元素等地球化学特征的分布规律与一定的沉积阶段和沉积环境相对应，因此，

最优分割法的数据处理，可能是第四纪地层划分和对比的有效方法。

三、 最优分割的计算步骤及其计算公式

1. 数据正规化

设原始资料矩阵为

$$X（p\times N）= \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pN} \end{bmatrix}$$

将矩阵 X 中的元素 x_{ij} 变换为：

$$z_{ij} = \frac{x_{ij} - \min_{1\le j\le n}\{x_{ij}\}}{\max_{1\le j\le n}\{x_{ij}\} - \min_{1\le j\le n}\{x_{ij}\}}$$ (i=1、2、…、p；j=1、2、…、N)

而得矩阵

$$Z_{(p\times N)} = [z_{ij}]$$

Matlab 代码：

```
function [std]= std1(vector)
%对矩阵进行标准化
%vector 为待分割矩阵
max1=max(vector);     %对列求最大值
min1=min(vector);     %对列求最小值
[a,b]=size(vector); %矩阵大小,a 为行数,b 为列数
for j=1:b
    std(j)= (vector(j)-min1)/(max1-min1);
end
```

2. 计算极差（或变差）矩阵

由上述极差（或变差）计算公式得到矩阵：

$$D = \begin{bmatrix} d_{12} & d_{13} & \cdots & d_{1N} \\ & d_{23} & \cdots & d_{2N} \\ & & \ddots & \vdots \\ & & & d_{N-1N} \end{bmatrix}$$

Matlab 代码：

```
function [D,a,b]=range1(vector)
%D 返回计算所得的极差矩阵
[a,b]=size(vector); %求矩阵大小,a 为行数,b 为列数
k=a;%当只计算单指标数据时，k=a=1
for i=1:b
    for j=i:b
        d(i,j)=max(vector(k,i:j))-min(vector(k,i:j));
    end
end
D=d;
```

3. 进行最优二分割

由 D 矩阵计算全部分两类的各种分割相应的总变差，即对每一个 m（m = N、N-1、…、2），求出相应的总变差

S_m（2；j）（j = 1、2、…、m-1）

找出最小值，确定各子段的最优二分割点α_1（m），即

$$S_m(2;\alpha_1(m)) = \min_{1\le j\le m-1} S_m(2;j)$$

从而得出 N 个样品的最优二分割

$$\{x_1、x_2、\ldots、x_{\alpha 1(N)}\}\{x_{\alpha 1(N)+1}、\ldots、x_N\}$$

Matlab 代码：

```
function [S,alp]=divi2(vector,n)
%最优二分隔,S 为最优二分割各段的分割点，
%a1 记录了二分割的序号
[d,a,b]=range1(vector);
alp=ones(n-1,b);%al(i,j)表示前 i 个样品的第 j 次分割点
S=zeros(b,b);
for m=2:b
    for j=1:m-1
        s(m,j)=d(1,j)+d(j+1,m);
    end
   S_temp(m,1)=min(s(m,1:m-1));
    for j=1:m-1
        if S_temp(m,1)==s(m,j);
            alp(n-1,m)=j;
        end
    end

    for t=1:m
        S(t,alp(n-1,t))=S_temp(t,1);
    end
end
```

4. 进行最优三分割

对于 m=N、N-1、…、4、3，由 $S_j(\alpha;\alpha_{1(j)})$（j=2、3、4、…、m-1）及 D 矩阵分别计算：

$$S_m(3;\alpha_{1(j)},j)=S_j(2;\alpha_{1(j)})+d_{j+1,N}$$

(m=N、N-1、…、4、3)

然后求出最小值，即

$$S_m(3;\alpha_1(m),\alpha_2(m)) = \min_{2\le j\le m-1} S_m(3;\alpha_1(j),j)$$

从而得到 N 个样品最优三分割

$$\{x_1、x_2、\ldots、x_{\alpha 1(N)}\}\{x_{\alpha 1(N)+1}、\ldots、x_{\alpha 2(N)}\}\{x_{\alpha 2(N)+!}、\ldots、x_N\}$$

5. 最优 K 分割

完全类似 4 的办法，在最优三分割的基础上可以进行最优四分割，继而进行五分割，以此类推，如果已经作出最优 K-1 分割，则可以产生最优 K 分割。

Matlab 代码：

```
function [S,alp]=divi(vector,n)
%n 为要分割的段数
[d,a,b]=range1(vector);
alp=zeros(1,b);%al(i,j)表示前 i 个样品的第 j 次分割点
for m=n:b
    for j=n-1:m-1
        if n==2
            s(m,j)=d(1,j)+d(j+1,m);
        else
            [S,alp]=divi(vector,n-1);
            s(m,j)=S(j,alp(j))+d(j+1,m);
        end
    end
    S=zeros(b,b);
    S_temp(m,1)=min(s(m,n-1:m-1));
    for j=1:m-1
        if S_temp(m,1)==s(m,j);
            alp(m)=j;
        end
    end
    for t=1:m
        if alp(t)~=0
            S(t,alp(t))=S_temp(t,1);
        end
    end
end

function [array] = sect(vector,n)
%vector 为样品矩阵，当直接对样品矩阵进行分割时调用该函数
%array 返回样品最优 n 分割的分割点号
%n 为要分割的段数
[a,b]=size(vector);
for num=n:-1:2
    [S,alp]=divi(vector,num);
    if num == n
        array(num-1)=alp(1,b);
    else
        array(num-1)=alp(array(num));
    end
end
```

```
function [array]=fsect(filename,n)
%filename 为需要分割的样品数据的文件名，如'd:\temp.txt'
%要对数据文件进行分割时调用该函数
fid=fopen(filename,'r');
[A,count]=fscanf(fid,'%f');
vector=A';
[a,b]=size(vector);
for num=n:-1:2
    [S,alp]=divi(vector,num);
    if num == n
        array(num-1)=alp(1,b);
    else
        array(num-1)=alp(array(num));
    end
end
```

四、 实例分析

琼州海峡标志性钻孔古地磁样品的磁化率参数，是海底沉积物在地质历史上各时期沉积产物受磁化作用强弱的重要参数，与沉积环境的磁场变化、氧化还原条件等因素有关，可据此对地层沉积的阶段性进行划分。

以下是钻孔 0-10m 岩芯中取得的 12 个样品数据：

ID	磁化率	ID	磁化率	ID	磁化率
1	6.0	5	5.7	9	8.3
2	6.0	6	6.3	10	7.7
3	5.3	7	5.3	11	7.7
4	4.0	8	4.7	12	10.3

则样品数据矩阵为：

Vector=[6.0 6.0 5.3 4.0 5.7 6.3 5.3 4.7 8.3 7.7 7.7 10.3]

1、样品正规化

```
>>v = stdl(Vector)
v =
    0.3175   0.3175   0.2063      0   0.2698    0.3651   0.2063
0.1111   0.6825   0.5873   0.5873      1.0000
```

2、计算极差矩阵

```
>>[D,a,b]=rangel(v);
>>D
D=
```

0	0	0.1111	0.3175	0.3175	0.3651	0.3651	0.3651	0.6825	0.6825	0.6825	1.0000
0	0	0.1111	0.3175	0.3175	0.3651	0.3651	0.3651	0.6825	0.6825	0.6825	1.0000
0	0	0	0.2063	0.2698	0.3651	0.3651	0.3651	0.6825	0.6825	0.6825	1.0000
0	0	0	0	0.2698	0.3651	0.3651	0.3651	0.6825	0.6825	0.6825	1.0000
0	0	0	0	0	0.0952	0.1587	0.2540	0.5714	0.5714	0.5714	0.8889
0	0	0	0	0	0	0.1587	0.2540	0.5714	0.5714	0.5714	0.8889
0	0	0	0	0	0	0	0.0952	0.5714	0.5714	0.5714	0.8889
0	0	0	0	0	0	0	0	0.5714	0.5714	0.5714	0.8889
0	0	0	0	0	0	0	0	0	0.0952	0.0952	0.4127
0	0	0	0	0	0	0	0	0	0	0	0.4127
0	0	0	0	0	0	0	0	0	0	0	0.4127
0	0	0	0	0	0	0	0	0	0	0	0

3、最优 K 分割

设 K=4

```
>>[array]=sect(v,4)
array =

     7     8    11
```

即样品段 Vector=[6.0 6.0 5.3 4.0 5.7 6.3 5.3 4.7 8.3 7.7 7.7 10.3]的最优四分割的分割点为 8 和 11，即最优四分割方案为：

{6.0 6.0 5.3 4.0 5.7 6.3 5.3}{4.7}{8.3 7.7 7.7}{10.3}

五、 结论

本文首先对有序样品的最优分割发原理及算法进行了深入分析，在深入理解原理的基础上，设计了 Matlab 算法并独立完成了代码编写，使得这一较为复杂的数学计算问题能够在计算机程序中快速的计算，以得到满意的答案。

但同时，算法设计与程序代码中仍存在一些问题，现列如下：

1、 用以衡量段间和段内样品数据差异的指标因不同的应用环境而应采用不同的指标，本文中作为示例仅以极差作为衡量指标。

2、 由于程序中使用了递归算法，使得当分割段数较多的情况下计算量迅速增大，计算时间较长，算法有待改进。

3、 未能及时进行视窗设计，实现软件化，所以实用代码仅能在 Matlab 命令窗里调用计算。