方法介绍

一类新的最优分割法及其应用

沈建法　黄叶芳
（华东师大）　（上海师大）

摘要　本文对最优分割法的目标进行了改造，提出了一类新的最优分割法，包括六种最优分割法 M1—M6。通过实例计算，发现新方法的分类性能指标有了明显改进。最后，把离散多目标规划方法与最优分割法结合起来，提出了一种寻求最优分类的协调型最优分割法。

一、问题的提出

最优分割法是对有序样本进行最优分类或分组的有效方法。单序列最优分割问题的提法是：设有一个有序序列 $x_1, x_2, \cdots, x_n$，要分成 K 类，记 $\bar{x}_{ij}$ 为 $x_i, x_{i+1}, \cdots, x_j$ 的平均，即

$$\bar{x}_{ij} = \frac{1}{j-i+1}\sum_{t=i}^{j} x_t, \quad \begin{matrix} 1 \leqslant i \leqslant n, \\ 1 \leqslant i \leqslant j. \end{matrix} \tag{1}$$

D_{ij} 为 $x_i, x_{i+1}, \cdots, x_j$ 的离差平方和。

$$D_{ij} = \sum_{t=i}^{j} (x_t - \bar{x}_{ij})^2, \quad \begin{matrix} 1 \leqslant j \leqslant n, \\ 1 \leqslant i \leqslant j, \end{matrix} \tag{2}$$

所谓最优分割就是要找出 $k-1$ 个分割点 $t_2, \cdots, t_k$，把序列分割成 k 段，使各段离差平方和为最小，即

$$\min_{t_2 \cdots t_k} D = \sum_{i=1}^{k} D_{t_i, t_{i+1}-1}, \tag{3}$$

这里 $t_1 = 1, t_{k+1} - 1 = n; t_i < t_{i+1}, i = 2, 3, \cdots k$。

上述最优分割法的目标是使组内总离差平方和为最小，故可称为最小离差平方和最优分割法。这一方法已在各个方面得到广泛应用，我们也曾用此方法研究了中国经济地

表 1　最小离差平方和最优五分割结果

组　号	样本数	平均数	最大值/最小值	最大值−最小值	离差平方和	离差平方和/平均数
1	1	6.02496	1	0	0	0
2	2	3.08386	1.075	0.222	0.024618	0.008
3	2	1.53569	1.284	0.382	0.0729711	0.048
4	6	0.996571	1.424	0.360	0.0908339	0.091
5	18	0.593514	2.086	0.395	0.199108	0.335

带的划分问题，效果较好，但同时也发现了一些问题。

用人均工农业产值为指标，对我国 29 个省市区作最优分类，表 1 给出了最优五分割的主要结果。

为便于分析，表 1 也列出了一些反映分类效果的指标。当进行最优五分割时，组内离差平方和仅为总离差平方和的 1.03%。但由表中可见，各组的组内差异性是极不平衡的，第 5 组的组内离差平方和为 0.199108，占 51%，组内数据的相差倍数第 2 组为 1.075，第 5 组为 2.086。为此，我们对最优分割法进行了改进，提出了一些新的最优分割法，并同离散多目标规划方法结合起来，提出了协调型最优分割法。

二、最优分割法的改进

最优分割法的核心是最优分割的目标函数，上述最优分割法存在的问题实际是由其使离差平方和最小这一目标决定的。所以必须改造最优分割法的目标函数。

我们首先考虑的方法是如何在计算离差平方和时消除各组数量级之间的差异对最优分割的影响。定义距离函数为：

$$D^1_{ij}=\sum_{e=i}^{j}\frac{(x_e-\bar{x}_{ij})^2}{\bar{x}_{ij}},\quad \begin{matrix}1\leqslant j\leqslant n,\\ 1\leqslant i\leqslant j.\end{matrix} \tag{4}$$

目标函数为

$$\min_{t_2,\cdots t_k} D^1=\sum_{i=1}^{k} D^1_{t_i,t_{i+1}-1}, \tag{5}$$

这里 $t_2,\cdots, t_k$ 为最优分割点。这一方法简记为 $M1$

引入参数 $\alpha(0\leqslant\alpha\leqslant 2)$ 可把上述方法与最小离差平方和最优分割法统一起来

$$D^2_{ij}=\sum_{e=i}^{j}\frac{(x_e-\bar{x}_{ij})^2}{\bar{x}^{\alpha}_{ij}},\quad \begin{matrix}1\leqslant j\leqslant n,\\ 1\leqslant i\leqslant j,\end{matrix} \tag{6}$$

目标函数仍同 (5)。当 $\alpha=0$ 时，即为最小离差平方和最优分割法，**当 $\alpha=1$ 时为方法 $M1$**。**当 $\alpha=2$ 时**，距离函数为常用的离差系数的平方的 $j-i+1$ 倍。当 α 取其它数值时，还可得到其它中间方法。**这一方法记为 $M2$.**

我们还试验了其它四种更为直观的方法。

方法 $M3$ 距离函数定义为组内最大值与最小值之差，即：

$$D^3_{ij}=\max_{e=i,\cdots j} x_e-\min_{e=i,\cdots j} x_e,\quad \begin{matrix}1\leqslant j\leqslant n,\\ 1\leqslant i\leqslant j,\end{matrix} \tag{7}$$

目标函数同 (5)。

方法 $M4$ 距离函数同方法 $M3$，即 (7) 式。目标函数为使各组最大差异最小，即

$$\min_{t_2,\cdots t_k} D^4=\max_{i=1,\cdots k} D^3_{t_i,t_{i+1}-1}, \tag{8}$$

方法 $M5$ 距离函数定义为组内最大值与最小值之比，即

$$D^5_{ij}=\frac{\max\limits_{e=i,\cdots j} x_e}{\min\limits_{e=i,\cdots j} x_e},\quad \begin{matrix}1\leqslant j\leqslant n,\\ 1\leqslant i\leqslant j,\end{matrix} \tag{9}$$

目标函数同 (5)

方法 M6 距离函数同方法 $M5$，即(8)式。目标函数使各组最大比最小，即

$$\min_{t_1\cdots t_k} D^6 = \max_{i=1,\cdots k} D^5_{t_i,t_{i+1}-1}, \tag{10}$$

上述各种方法均可以用一种类似于动态规划的递推算法求解，其基本思路是先从前 m 个数据($3 \leqslant m \leqslant n$)开始，分别对其作最优二分割、三分割，…，$k^*$ 分割($k^* = \min(m-1, k)$)，找出各最优分割的最后一个分割点，依次类推，每一步均利用前一步的最优分割结果，直至找出 n 个数据序列最优 k 分割的分割点 $t_2\cdots t_k$。

记 $E_{m,i}$ 为前 m 个数据作最优 i 分割时的目标值，$T_{m,i}$ 为其最后一个最优分割点(分割点分别为各组最后一个数据号)，对于方法 $M1$、$M2$、$M3$、$M5$，递推公式为：

$$\begin{cases} E_{m,i} = \min\limits_{i-1\leqslant j^*\leqslant m-1}\{E_{j,i-1} + D_{j+1,m}\}, \\ T_{m,i} = j^*, \end{cases} \quad \begin{matrix} i = 2, 3, \cdots k^*, \\ m = 3, 4, \cdots, n. \end{matrix} \tag{11}$$

对于方法 $M4$，$M6$，递推公式为：

$$\begin{cases} E_{m,i} = \min\limits_{i-1\leqslant j^*\leqslant m-1} \max(E_{j,i-1}, D_{j+1,m}), \\ T_{m,i} = j^*, \end{cases} \quad \begin{matrix} i = 2, 3, \cdots k^*, \\ m = 3, 4, \cdots, n. \end{matrix} \tag{12}$$

(11)、(12)式的初始条件均为

$$\begin{cases} E_{m,1} = D_{1,m},\ T_{m,1} = m \quad m = 1, 2, \cdots, n, \\ E_{m,m} = 0,\ T_{m,m} = m-1,\ m = 2, 3, \cdots, n. \end{cases} \tag{13}$$

表 2、表 3、表 4 分别给出了上述例 $M1$、$M2$、$M6$ 的最优五分割结果。

由表 2、表 3、表 4 可见，分类性能指标较原最优分割法有所改善，尤其是方法 6 ($M6$) 最为显著，第 3、4、5 组最大值与最小值之比均在 1.6 左右，反映了组内差异各组基本接近，这时组内离差平方和仅为总离差平方和的 1.46%，仅较原最优分割结果增加 0.43%。表 5 给出了根据方法 6 最优五分割结果作出的我国 29 个省市自治区分类情况，基本上反映了我国经济发展情况的省际差异。I、II 为我国三个经济最发达的大城市，人均工农业产值在 2970 元以上，III 类为我国东部最发达的省，人均工农业产值在 1060 至 1730 元之间，IV 类包括沿海和内地较发达的 13 个省区，人均工农业产值在 610 至 1040 元之间，V 类包括我国内地 9 个最不发达的省区，人均工农业产值在 350 至 600 元之间。

表 2 $M1$ 最优五分割结果

组 号	样本数	平均数	最大值/最小值	最大值 − 最小值	离差平方和	离差平方和/平均数
1	1	6.025	1	0	0	0
2	2	3.084	1.075	0.222	0.025	0.003
3	3	1.427	1.427	0.516	0.144	0.101
4	7	0.897	1.424	0.317	0.092	0.103
5	16	0.573	1.961	0.349	0.141	0.246

表3 M_2 最优五分割结果（$a=2$）

组 号	样本数	平均数	最大值/最小值	最大值−最小值	离差平方和	离差平方和/平均数
1	1	6.025	1	0	0	0
2	3	2.631	1.85	1.468	1.253	0.476
3	7	1.046	1.582	0.494	0.194	0.186
4	14	0.64	1.382	0.189	0.05	0.078
5	4	0.43	1.323	0.12	0.011	0.025

表4 M_6 最优五分割结果

组 号	样本数	平均数	最大值/最小值	最大值−最小值	离差平方和	离差平方和/平均数
1	1	6.025	1	0	0	0
2	2	3.084	1.075	0.222	0.025	0.008
3	4	1.337	1.619	0.66	0.241	0.180
4	13	0.75	1.678	0.418	0.220	0.293
5	9	0.514	1.631	0.229	0.063	0.122

表5 我国省区分类

类 型	省、市、自治区(以人均工农业产值为序)
I	上海市
II	天津市、北京市
III	辽宁、江苏、黑龙江、吉林
IV	浙江、湖北、山东、山西、河北、广东、新疆、湖南、陕西、内蒙古、宁夏、甘肃、福建
V	青海、四川、河南、江西、安徽、广西、云南、贵州、西藏

三、协调型最优分割法

前面我们对最小离差平方和最优分割法进行了改进，提出了6种新的最优分割法，试验结果表明不同分割方法得到的最优分割结果是不同的，且分类性能指标也不同，结果往往具有帕累托最优解(非劣解)的特性，无法确定唯一最优解。实际上，对于不同的具体分类问题，对分类性能指标的要求也可能是不同的。有的要求组内绝对差别不要太多，有的则要求组内相对差别不要太大，因此，要根据分类要求找出协调最优解(满意解)。这里用离散多目标规划方法求解，其基本思想如下：

设有N个最优分割方案，每个方案有K个分类性能指标，指标值越大则表示分类效果越好。P_{ij} 表示第 i 个方案第 j 个指标值，$i=1,2,\cdots N$; $j=1,2,\cdots K$。

1. 对各性能指标进行极差标准化处理，消除单位、数量级的影响，记标准化数据为

V_{ij}，则

$$V_{ij} = \frac{P_{ij} - \min_i P_{ij}}{\max_i P_{ij} - \min_i P_{ij}} \tag{14}$$

这里 $\max_i P_{ij}$ 为所有方案中 j 指标最大者，称为理想值，$\min_i P_{ij}$ 为所有方案中 j 指标最小者，称为最小解。

2. 定义协调解为与理想方案总差异最小的方案，即找出方案 i 使：

$$d_i = \min_i d_i = \left(\sum_{j=1}^{N} (1 - V_{ij})^p\right)^{1/p}, \quad P \geq 1 \tag{15}$$

则方案 i 即为协调解，对应的 K 个性能指标为 P_{ij}，$j = 1 \cdots K$。

3. 协调解满意则结束，否则输入要改善的性能指标集合 A。定义满足以下条件的方案 i 为新的最优分割方案集合

$$P_{ij} \geq P_j, \quad j \in A$$

4. 对于新的最优分割方案集合重复步骤 1—3，直至协调解满意或无法改善为止。

对前述最优分割例，我们计算了 12 种最优四分割方案，每组用离差平方和百分比（O_1，O_2，O_3，O_4）、离差平方和/平均数（O_5、O_6、O_7、O_8）、最大值-最小值（O_9、O_{10}、O_{11}、O_{12}）、最大值/最小值（O_{13}、O_{14}、O_{15}、O_{16}）四个分类性能指标，共 16 个目标，用离散多目标规划方法求解，结果见表 6。第一步协调解为方案 1（最小离差平方和最优分割法）、第二步协调解为方案 10（方法 $M2$，$a = 2$），O_{16} 由 2.575 降至 2.086，第三步协调解为方案 6（方法 $M6$），O_{16} 由 2.086 降至 1.807，接着要求改善目标 O_{15}，但直到第六步一致收敛时协调解仍为方案 6，说明已无法改进，故协调最优解为用 $M6$ 方法得到的最优四分割结果。上述方法把多种最优分割法与离散多目标规划结合起来，可称为协调型最优

表 6　协调型最优分割法迭代结果

目标	第一步			第二步			第三步		
	理想值	最小解	协调解（方案 1）	理想值	最小解	协调解（方案 10）	理想值	最小解	协调解（方案 6）
O_1	0	0	0	0	0	0	0	0	0
O_2	0	76	3	3	76	76	3	76	71
O_3	0	73	41	12	73	12	12	73	22
O_4	7	100	56	7	56	12	7	24	7
O_5	0	0	0	0	0	0	0	0	0
O_6	0	0.476	0.008	0.008	0.476	0.476	0.008	0.476	0.476
O_7	0	0.53	0.247	0.186	0.53	0.186	0.186	0.53	0.411
O_8	0.213	3.162	0.673	0.213	0.673	0.335	0.213	0.335	0.213
O_9	0	0	0	0	0	0	0	0	0
O_{10}	0	1.468	0.222	0.222	1.468	1.468	0.222	1.468	1.468
O_{11}	0	0.876	0.693	0.494	0.876	0.494	0.494	0.876	0.632
O_{12}	0.294	1.363	0.573	0.294	0.573	0.395	0.294	0.395	0.294
O_{13}	0	0	0	0	0	0	0	0	0
O_{14}	0	1.85	1.075	1.075	1.85	1.85	1.075	1.85	1.85
O_{15}	0	2.031	1.671	1.582	2.031	1.582	1.582	2.031	1.886
O_{16}	1.807	4.747	2.575	1.807	2.575	2.086	1.807	2.086	1.807

分割法。

四、结　语

本文对最小离差平方和最优分割法进行了改进，提出了6种新的最优分割法，能显著改善分类性能指标。最后，引入了离散多目标规划方法，建立了协调型最优分割法。

参　考　资　料

[1] 沈建法，最优分割法在空间分析中的应用，经济地理，2(1987).
[2] 华东师大数理统计系，统计模型与方法(讲义)，1984.12.
[3] Peter Nijkamp, Multidimensional Spatial Data and Decision Analysis, John Wiley & Sons, 1979.
[4] R. Gnanadesikan, Methods for Statistical Data Analysis of Multivariate Observations, John Wiley & Sons, 1977.

二阶线性微分算子的分解及其应用

黎　耀　善

(商丘师范专科学校)

摘要　本文给出由二阶线性微分算子的分解式求解二阶线性微分方程和二维线性微分方程组的方法，并由此得到它们的一些可积类型与可积的充要条件。

n 阶线性微分算子的一般形式为

$$D^n + a_{n-1}(x)D^{n-1} + a_{n-2}(x)D^{n-2} + \cdots + a_1(x)D + a_0(x),$$

直接求它的逆算子是困难的，在此我们采用分解式的方法。

为方便，本文只讨论二阶线性微分算子的分解(且设本文中出现的各函数均为连续函数)并应用它去解二阶线性微分方程和二维线性微分方程组，得到它们的一些可积类型与可积的充要条件。

一、二阶线性微分算子的分解

由微分算子的定义容易验证下列等式

$$[D + A_1(x)][D + A_2(x)] = D^2 + (A_1(x) + A_2(x))D + A_1(x)A_2(x) + A_2'(x), \quad (1)$$

故若

$$D^2 + p(x)D + q(x) = [D + A_1(x)][D + A_2(x)], \quad (2)$$

则有

$$\begin{cases} p(x) = A_1(x) + A_2(x), \\ q(x) = A_1(x)A_2(x) + A_2'(x), \end{cases} \text{即} \quad \begin{cases} A_1(x) = p(x) - A_2(x), & (3) \\ A_2'(x) = q(x) - p(x)A_2(x) + A_2^2(x). & (4) \end{cases}$$