

测量数据信息的模式识别方法^①

靳奉祥 王同孝 独之行 郭金运

(山东矿业学院地球科学系, 泰安 271019)

摘 要 以模式识别理论为基础, 以测量数据统计分析为应用背景, 提出了测量数据信息的模式分析和模式识别的概念与方法。在最小误判概率的原则下, 应用 J_D 和 J_C 统计量, 分别对均值和方差的差异性检验建立了无水平判决的方法, 为测绘信息模式识别理论的发展奠定了基础。

关键词 母体均值 母体方差 模式识别方法

中图法分类号 P207

随着模式识别理论和方法的日益完善, 它的应用几乎遍及了各个学科领域, 测绘领域也不例外。统计假设检验理论在测量数据处理中解决了很多问题, 也占有重要位置, 但它固有的缺点显而易见: 存在着人为因素作用^[1-3], 处理结果中包含着很大的主观性。作者认为无论是统计检验还是模式识别最终解决的都是分类问题, 只是它们分类的基点不同而已。将模式识别的理论方法及基本思想应用于该方面的研究, 就会克服假设检验理论的缺点, 增加结果的可靠性。

在测量数据处理范畴内的统计检验中, 需要解决的根本问题无外乎两个: 其一为测量数据的母体均值位移显著性问题; 其二为测量数据的母体方差差异性问题。若用模式识别的思想方法能完全解决这两个问题, 应该说是测量数据处理理论的巨大贡献。本文对此作了一些研究工作, 取得了一些阶段性的成果, 但还需进一步完善, 在此提出来, 与同行讨论。

1 初步解决的问题及思想方法

在数理统计学理论中, 观测值总是被看成是随机变量, 由于抽样的随机性, 观测值中必

然存在着偶然误差; 由于观测外部条件及人为因素等, 观测结果中也常受系统误差的影响, 甚至有观测粗差的干扰, 这就造成了母体均值位移显著性问题或母体方差差异性问题。

1.1 母体均值位移显著性问题

设 $\bar{x}$ 和 $\bar{y}$ 分别是对同一量两次观测所得子样的均值, 且 $\bar{x}$ 概率性趋于 ζ_1 , $\bar{y}$ 概率性趋于 ζ_2 , 即子样均值为母体均值的无偏估计。子样均值隐含母体均值大小的信息和观测方差大小的信息, 是一定观测方差条件下子样的函数。若顾及观测方差大小这一条件, 能够处理两次子样均值的差异性问题, 显然可以解决实际应用问题。巴尔达 (Baarda) 所提出的数据探测法 (Data Snooping) 和统计检验法都较成功地解决了该问题, 并在相当长的时段内, 占据数据处理的主导地位^[2-4]。但这些方法所具有的弱点和局限性限制了它们的应用范围和解决问题的彻底性和合理性, 模式识别是解决分类问题的数学工具, 既可以处理两类问题的分类, 又可以处理多类问题的分类; 既能处理一维问题, 又能处理多维问题。因此, 用最为成熟的统计模式识别的理论和方法完全可以提取出测量数据中所隐含的母体均值位移显著性的信息, 从而更为合理地解决了该问题, 如变形观

① 山东省青年科学基金资助项目 收稿日期: 1998-03-04; 修回日期: 1999-02-04

靳奉祥, 男, 37岁, 博士, 教授, 博士生导师

测位移显著性的解决。

1.2 母体方差差异性问题

在表示某项观测工作的精度时,通常采用的是体现精度的母体方差(或中误差),而母体方差(或中误差)不知时,用其估计的子样方差(或子样中误差)来代替,换句话说,子样方差(或子样中误差)包含着某项工作的精度信息^[5-7]。如本文1.1节所述,用模式识别理论可以从中将精度信息提取出来,因此问题得以解决,如测量数据的周期性误差的解决。

1.3 测绘模式识别的思想方法

模式识别的实质问题,实际上是科学有效的分类问题。受模式识别处理问题基本思想的启发而形成测绘模式识别的思想方法,在处理母体均值位移显著性和母体方差差异性问题时,具体表述为:根据不同问题利用适当的准则函数(如概率距离准则、类内类间距准则),以贝叶斯决策理论的最小错误判决规则或最大似然比判决规则为基础,确定出判决基准,进行判识;在给出判识结果的同时,给出判识结果的可靠性(错误率),判识步骤可用图1表示。这种处理方法是传统处理方法的补充和发展,具有诸多优点,主要有:①克服了统计检验理论选择显著性水平的主观性;②依据贝叶斯决策理论的最小错误判决规则或最大似然比判决规则,使得误判概率总是最小;③既可根据确定的判识基准作有水平判决,又可据误判概率的大小作无水平判决;④在给出结论的同时给出判识的可靠性。

2 模式分析与判识

2.1 模式类描述与特征变量

本文所讨论的问题用模式识别方法来处理时,必然首先给出可能的模式类别。一组观测数据对于母体均值位移显著性问题来讲,虽然事先并不知道这批数据的母体均值位移是否显著,但总有可能的两种情况,即显著或不显著,因此可以将模式类描述为数据的母体均值位移显著类和数据的母体均值位移不显著类;同样,对于母体方差差异性问题来说,也可以将模式类描述为数据的母体方差差异性显著类和数据的母体方差差异性不显著类。

正如统计检验一样,必须找到能够反映实际问题的统计量,才能进行分类。模式识别中,通称这种统计量为特征变量。特征变量应具有以下特点:①它是观测子样的统计量,确能反映所研究问题的实质;②具有固定的概率密度分布形式,隐含数据的可分性信息;③能够确定其维数。

设 X 为 $N(\zeta_1, \sigma_1)$ 变量, Y 为 $N(\zeta_2, \sigma_2)$ 变量,两者的观测精度相同或不同。现抽得两组子样: $x_1, \dots, x_{n_1}$ 和 $y_1, \dots, y_{n_2}$ 。由抽得的子样可分别算得两子样均值和子样方差,其计算式为

$$\left. \begin{aligned} \bar{x} &= \frac{1}{n_1}(x_1 + \dots + x_{n_1}) \\ \bar{y} &= \frac{1}{n_2}(y_1 + \dots + y_{n_2}) \end{aligned} \right\} \quad (1)$$

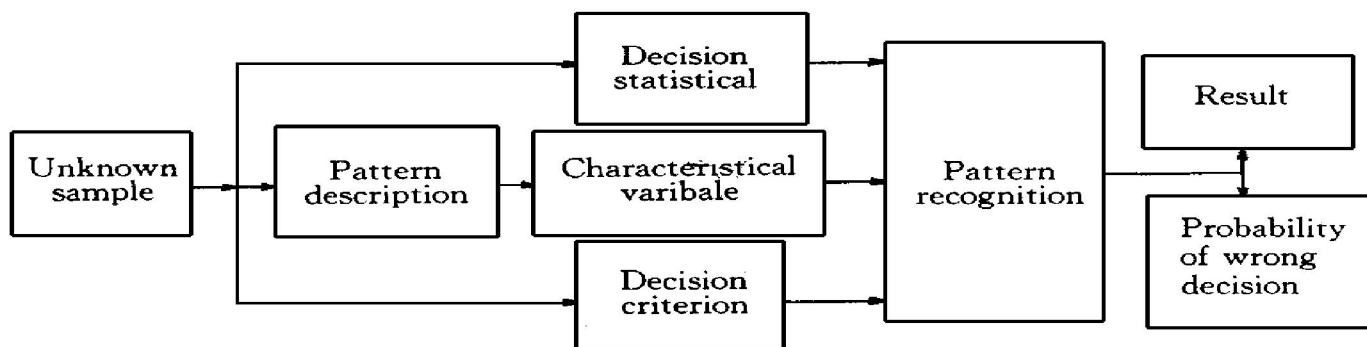


图1 测绘模式判识过程图

Fig. 1 Procedure of pattern recognition

$$\left. \begin{aligned} \bar{s}_x^2 &= \frac{1}{n_1 - 1} \sum_1^{n_1} (x_i - \bar{x})^2 \\ \bar{s}_y^2 &= \frac{1}{n_2 - 1} \sum_1^{n_2} (y_i - \bar{y})^2 \end{aligned} \right\} \quad (2)$$

根据数理统计知识和特征变量的要求, $\bar{x}$ 和 $\bar{y}$ 可以作为母体均值位移显著性问题中的特征变量, $\bar{s}_x^2$ 和 $\bar{s}_y^2$ 可以作为母体方差差异性中的特征变量。它们的可分性可概括为三种情况: ①两个特征变量的分布密度函数完全没有重叠, 即完全可分; ②特征变量的分布密度函数完全重叠, 即完全不可分; ③一般情况下, 两随机变量的分布密度函数总是部分地重叠, 这种情况下, 可以在一定判识基准下给出可分或不可分的结论。如图2、图3表示 $\bar{s}_1^2$ 和 $\bar{s}_2^2$ 可分性情况^[5, 6, 8]。

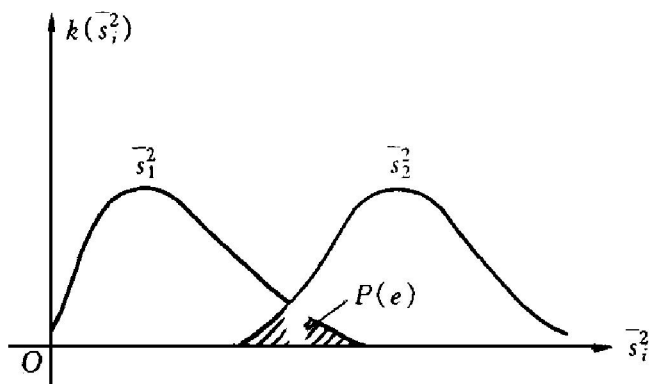


图2 在一般情况下的分布重叠情况

Fig. 2 Partly overlapping of two distribution functions of $\bar{s}_1^2$ and $\bar{s}_2^2$

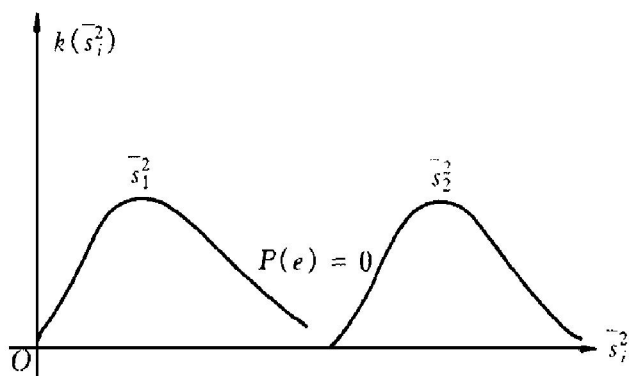


图3 $\bar{s}_1^2$ 和 $\bar{s}_2^2$ 完全可分情况

Fig. 3 Without overlapping of two distribution functions of $\bar{s}_1^2$ and $\bar{s}_2^2$

2.2 准则函数与错误率

由于概率距离准则直接反映类概率密度的分布情况, 且与分类错误率密切相关, 与模式类别的可分性直接相关联, 为类可分性的一种度量, 因此, 可以用概率距离准则作为准则函数来研究方差差异性和均值位移显著性问题。经研究可以用 J_C 和 J_D 准则来处理所提出的问题。

(1) 假设随机模式的特征空间是 d 维的, $p(\mathbf{x}/w_i)$ 和 $p(\mathbf{x}/w_j)$ 分别表示两类的类概率密度, E_d 为 $\mathbf{x}$ 的值域空间, 则 J_C 准则可以写成^[8]:

$$J_C = \ln \int_{E_d} [p(\mathbf{x}/w_i)]^{1-s} \cdot [p(\mathbf{x}/w_j)]^s d\mathbf{x} \quad (3)$$

式中 s 在闭区间 $[0, 1]$ 内取值。当类概率密度服从正态分布时, 有

$$J_C = \frac{1}{2} s (1-s) (\mu_i - \mu_j)^T \cdot [(1-s) \Sigma_i + s \Sigma_j]^{-1} (\mu_i - \mu_j) + \frac{1}{2} \ln \frac{\| (1-s) \Sigma_i + s \Sigma_j \|}{\| \Sigma_i \|^{1-s} \| \Sigma_j \|^s} \quad (4)$$

式中 μ_i, μ_j 为两类的均值矩阵, Σ_i 和 Σ_j 为两类的方差矩阵。当 $\Sigma_i = \Sigma_j = \Sigma$, $s = 0.5$ 时, 则有

$$J_C = \frac{1}{8} (\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j) = \frac{1}{8} r_{ij}^2 \quad (5)$$

式中 r_{ij} 为马氏距离。

考虑 J_C 准则, 相应地可以得出错误率上界的公式为

$$P(e) \leq [P(w_i)]^{1-s} [P(w_j)]^s \exp(-J_C) \quad (6)$$

式(6)右端是基于贝叶斯决策理论最小错误率判决规则处理 w_i/w_j 两类问题时的错误率上界, $P(w_i)$ 和 $P(w_j)$ 为两类的先验概率。

当为正态分布, 且两类协方差矩阵相同, 即 $\Sigma_i = \Sigma_j = \Sigma$ 时, 则有

$$P(e) \leq \sqrt{P(w_i) P(w_j)} \exp\left(-\frac{1}{8} r_{ij}^2\right) \quad (7)$$

(2) 据(1)的假设, J_D 准则有^[5]

$$J_D = \int_{E_d} [P(\mathbf{x}/w_i) - P(\mathbf{x}/w_j)] \cdot \ln \left[\frac{P(\mathbf{x}/w_i)}{P(\mathbf{x}/w_j)} \right] d\mathbf{x} \tag{8}$$

若类概率密度 $P(\mathbf{x}/w_i)$ 的分布服从 $N(\mu_i, \Sigma_i)$, $P(\mathbf{x}/w_j)$ 的分布服从 $N(\mu_j, \Sigma_j)$, 即有

$$J_D = \frac{1}{2} \text{tr} [(\Sigma_i - \Sigma_j)(\Sigma_j^{-1} - \Sigma_i^{-1})] + \frac{1}{2} \text{tr} [(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j)(\mu_i - \mu_j)^T] \tag{9}$$

当模式样本 $\mathbf{x}$ 的各维特征之间是相互独立的, 则有

$$J_D = \sum_{k=1}^d \left(\frac{\mu_{ik} - \mu_{jk}}{\sigma_k} \right)^2 \tag{10}$$

考虑 J_D 准则, 一般情况下, 其相应错误率计算较难, 但当特征变量为正态分布时, 且属两类一维问题时, 其错误率解析计算为

$$P(e) = \int_{\frac{1}{2}\sqrt{J_D}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \tag{11}$$

无论是 J_C 准则, 还是 J_D 准则, 相应计算出的错误率都是在贝叶斯决策理论的最小错误率判决或最大似然比判决规则的条件下作出的, 如图 4。这样根据判识过程就可进行模式

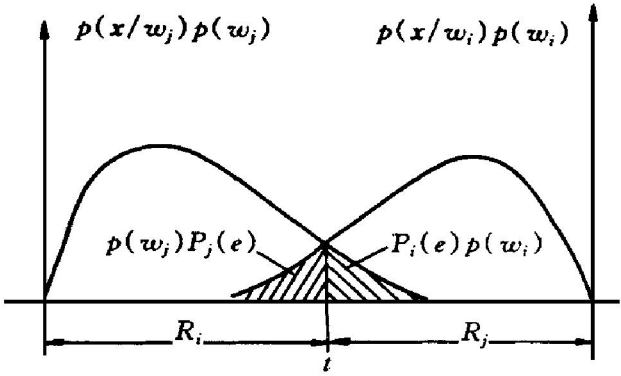


图 4 最小错误率判决规则的错误率 (一维)

Fig. 4 Minimum probability of wrong decision (one dimension)

判识。

3 实例及分析

例 1^[9]对某一方向观测自由网进行了第二期复测, 按消去定向未知数进行了伪逆平差, 平差结果列于表 1。以 ΔX , X_{II} 为特征变量, 以 $\frac{1}{2}\sqrt{J_D}$ 作为判识统计量, 记为 θ_D , 结果见表 2。

为了与 t 检验法作比较^[8], 以错误率 $\alpha = 0.05$ 时的 θ_D 为基准, 从结果上看各坐标位移

表 1 坐标平差成果

Table 1 Adjusted coordinates

	X_A	Y_A	X_B	Y_B	X_C	Y_C	X_D	Y_D	$T^T P V$
X_I	- 0. 4	0. 6	0. 9	- 0. 2	- 0. 5	- 0. 3	0	- 0. 1	0. 046
X_{II}	2. 8	0	- 0. 1	- 0. 7	0. 7	- 0. 9	- 2. 5	1. 6	0. 345
$Q_I = Q_{II}$	0. 114	0. 079	0. 091	0. 119	0. 087	0. 101	0. 105	0. 125	
ΔX	3. 2	- 0. 6	- 1. 9	- 0. 5	1. 2	- 0. 6	- 2. 5	1. 7	

表 2 以 J_D 为基础的各点的位移统计量及相应的错误率

Table 2 Statistical value and wrong decision probability of every coordinate

	X_A	Y_A	X_B	Y_B	X_C	Y_C	X_D	Y_D
$\theta_D = \frac{1}{2} \sqrt{J_D}$	21. 6	4. 8	14. 3	3. 2	9. 3	4. 5	17. 6	10. 9
$P(e)$	0	0	0	0. 0007	0	0. 0003	0	0
$\theta_D _{P(e)=\alpha=0.05}$	1. 6	1. 6	1. 6	1. 6	1. 6	1. 6	1. 6	1. 6

Note: when $P(e) < 10^{-5}$, $P(e)$ is regarded as zero.

表 3 原始观测数据表

Table 3 Observation data

Observation	<i>AB</i>	<i>AD</i>	<i>BD</i>
First result:			
Mean distance $\bar{x}/\text{m}$	15 287. 689	20 795. 463	13 869. 854
Measuring number in sets n_1	11	4	10
Mean square error $\bar{s}_1/\text{m}$	0. 007	0. 011	0. 010
Second result:			
Mean distance $\bar{y}/\text{m}$	15 287. 676	20 795. 456	13 869. 892
Measuring number in sets n_2	5	6	4
Mean square error $\bar{s}_2/\text{m}$	0. 011	0. 007	0. 001

均属显著类，与 t 检验结果相同，判决结果的可靠性大小反映在错误率上。例如在 Y_A 和 Y_C 位移量的判决中，虽然两者的位移量相当，但通过错误概率可以看出，前者的判决结果要比后者可靠。所以也可直接根据错误率作无水平判识。

例 2^[2]在某地高精度三角网上试验某类型的激光测距仪。在网中三角形 ABD 的三个边上试测的结果及有关计算见表 3。现欲判识 AB ， AD 和 BD 两次观测的母体方差是否相等。

以 $\bar{s}_1^2$ 和 $\bar{s}_2^2$ 作为特征变量，取 $\alpha=0.05$ 时的 J_C 值作为判识基准 θ_C ，结果如表 4。

为了与 F 检验法作比较^[2]，以错误率 $\alpha=0.05$ 时的 θ_C 为基准，从结果看 AB ， AD 两边差异不显著， BD 边方差差异显著，与 F 检验结果相同，差异显著程度反映在错误率上界上。也可直接根据错误率作无水平判识。

从例 1 中可以看出： Y_A 与 Y_C 的两期差是相等的，然而其 $\frac{1}{2}\sqrt{J_D}$ 和 $P(e)$ 是不等的，后者的 J_D 小于前者，后者的 $P(e)$ 大于前者，说明前者较后者更可靠，同理在例 2 中， AB 和 AD 的 J_C 较小而 $P(e)$ 较大，而 BD 的 J_C 较大， $P(e)$ 较小，这有两种解释，第一， $P(e)$ 可认为是判决的可靠性指标， $P(e)$ 越小，则判决的可靠性就越高，反之亦然；第二， $P(e)$ 较大，说明判决风险大或判决难度大，或难以区分判决，所以可不作差异的显著性检验。

综上所述，我们认为模式识别理论用于测

表 4 以 J_C 为基础的两组方差差异性及其相应的错误率上界

Table 4 Statistical value and super limit of wrong decision probability

	<i>AB</i>	<i>AD</i>	<i>BD</i>
J_C	0. 532	0. 401	2. 781
θ_C ($J_C/\alpha=0.05$)	2. 30	2. 30	2. 30
$P(e)$ super limit	0. 293 7	0. 334 8	0. 031
Test result	Not significant	Not significant	Significant

量数据信息的处理方法有着非常突出的优点，是对传统理论与方法的完善，测绘模式识别技术将会随着模式识别理论的完善而发展起来。

REFERENCES

1 Tao Benzao (陶本藻). Statistical Analysis on Surveying Data (测量数据统计分析). Beijing: Surveying and Mapping Publishing House, 1992.

2 Li Qinghai (李庆海) and Tao Benzao (陶本藻). Principle of Probability and Statistics and Its Application in Surveying (概率统计原理和在测量中的应用). Beijing: Surveying and Mapping Publishing House, 1982: 159.

3 Jin Fengxiang (靳奉祥), Wang Tongxiao (王同孝) and Du Zhixing (独知行). The Chinese Journal of Nonferrous Metals (中国有色金属学报), 1997, 7 (3): 18.

4 Li Deren (李德仁). Error Process and Reliability Theory (误差处理和可靠性理论). Beijing: Sur-

- veying and Mapping Publishing House, 1988: 187. 240.
- 5 Caspary W F. Concepts of Network and Deformation Analysis. School of Surveying, University of New South Wales, 1987. 8 Li Jinzong (李金宗). Introduction of Pattern Recognition (模式识别导论). Beijing: High Education Publishing House, 1994: 234– 390.
- 6 Pope A J. NOAA Technical Report No. 65 NGS, Geodetic Survey, Rockville Md. 1976. 9 Tao Benzao (陶本藻). Adjustment of Free Network and Deformation Analysis (自由网平差与变形分析). Beijing: Surveying and Mapping Publishing House, 1984: 156.
- 7 Ethrog U. Statistical Test of Significance for Testing Outlying Observations. Survey Review, 1991: 31:

PATTERN RECOGNITION ON SURVEYING DATA INFORMATION ANALYSIS

Jin Fengxiang, Wang Tongxiao, Du Zhixing and Guo Jinyun

*Department of Geoscience, Shandong Institute of Mining and Technology,
Tai'an 271019, P. R. China*

ABSTRACT Based on the pattern recognition theory and the practical background of surveying data statistical analysis, some ideas and methods of pattern analysis and pattern recognition have been proposed, for surveying information analysis. Considering the minimum wrong decision probability, the deciding approach for the statistical non-level test of the divergence among the means and among the variances was established by the statisticals J_D and J_C . All of these give a good basis for surveying pattern recognition theory.

Key words population mean population variance pattern recognition

(编辑 何学锋)