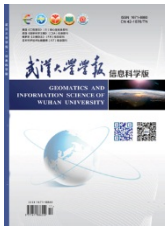


武汉大学学报(信息科学版)
Geomatics and Information Science of Wuhan University
ISSN 1671-8860,CN 42-1676/TN

《武汉大学学报(信息科学版)》网络首发论文

题目：适用于训练样本选择的斜交因子模型研究
作者：虞欣，郑肇葆，李林宜
DOI：10.13203/j.whugis20200631
收稿日期：2020-12-10
网络首发日期：2021-06-16
引用格式：虞欣，郑肇葆，李林宜．适用于训练样本选择的斜交因子模型研究．武汉大学学报(信息科学版). https://doi.org/10.13203/j.whugis20200631

DOI:10.13203/j.whugis20200631

适用于训练样本选择的斜交因子模型研究[1]

虞欣[1]　郑肇葆[2]　李林宜[2]

1 北京石油化工学院，北京,102617
2 武汉大学 遥感信息工程学院，湖北 武汉,430079

摘　要:训练样本的质量直接影响到训练阶段的训练质量（或效果），进而在一定程度上影响到测试阶段的分类精度。训练样本的代表性和典型性则反映出训练样本质量的一个重要方面。对于当前非常流行的深度学习模型研究，如何尽可能地减少训练样本的数量，一方面成为一个非常“棘手”的问题，另一方面从实际应用的角度来看，这也上升为一个经济或成本方面的问题。文章提出一种适用于训练样本选择的斜交因子模型方法，该方法松弛了 Q 型因子分析和对应分析对于公因子之间独立的假设条件，并在斜交参考解的基础上提出一种适合训练样本选择的近似求解斜交旋转的方法。实验与分析表明:提出的方法是可行、有效的。与基于正交因子模型的方法相比，它可以更好地描述或逼近现实的真实情况，可以选择出更合理、更具有代表性的典型训练样本，并且还可以取得满意的分类精度。适用于训练样本选择的斜交因子模型方法优于基于正交因子模型的训练样本的选择方法，被选择的训练样本，其分布相对更分散、更合理，而且总的分类精度平均提高 3%左右。提出的方法，在理论方面可以为优化样本的采集提供一种理论支持，在实际应用方面对样本的采集具有指导或参考意义。

关键词:　斜交因子模型；训练样本；影像分类；正交因子模型；样本选择

中图分类号:　P237　　**文献标志码:**　A

训练样本的质量和数量直接影响到训练阶段的训练质量（或效果），进而在一定程度上影响到测试阶段的分类精度[1-6]。如何选择训练样本一直困扰着影像分类领域的研究工作者。训练样本的典型性或代表性反映出训练样本质量的一个重要方面[7-10]。特别是对于当前非常流行的深度学习模型研究，如何尽可能地减少训练样本的数量，或者说如何才能使得所需的训练样本能够达到“少而精”的目标，这一方面成为一个非常“棘手”的问题，另一方面也上升为一个经济或成本方面的问题[11-22]。文献[23]提出一种基于 Q 型因子分析的训练样本的选择方法，利用 Q 型因子分析从一批样本中选出少数的具有一定典型性的样本作为训练样本。并且与基于人工随机地选择训练样本的方式进行了比较试验，试验结果表明该方法可以获得更好的分类精度。此后，考虑到计算量等问题，在文献[24]的基础上提出一种基于对应分析的训练样本的选择方法，以克服 Q 型因子分析计算量大的缺点。实际上，文献[23]和[24]所采用的因子模型中的公因子都是正交的（称为正交因子模型），而且进行的因子旋转也是正交旋转，这意味着在利用 Q 型因子分析或者对应分析时必须假设那些公因子（即所选择的代表性或典型性的“公共样本”）是相互独立的，然而在现实中对样本（或变量）都发生影响的公因子之间往往是相互联系的，即公因子之间是相关的，称这种相关的公因子为斜交公因子（简称斜交因子），大量实验数据证明了斜交因子是普遍的，而正交因子只是在少数范围内存在，或者作为斜交因子的一种近似[25-26]。鉴于此，本文提出一种适用于训练样本选择的斜交因子模型方法，该方法松弛了 Q 型因子分析和对应分析对于公因子之间独立的假设条件，并在斜交参考解的基础上提出一种适合训练样本选择的近似求解斜交旋转的方法。实验与分析表明，本文提出的方法是有效、可行的。与基于正交因子模型的方法相比，它可以更好地描述或逼近现实的真实情况，可以选择出更合理、更具有代表性的典型训练样本，并且还可以取得更加满意的分类精度。

收稿日期: 2020-12-10
项目资助: 国家重点研发计划课题（编号：2018YFC0407804）
第一作者: 虞欣，博士，教授，主要从事影像解译、人工智能和贝叶斯统计等研究工作。china_yuxin@163.com
通信作者: 李林宜，博士，副教授。lilinyi@whu.edu.cn

1 训练样本的选择方法

正交旋转通常是指在旋转的过程中因子之间互相正交，并且始终保持初始解中因子之间不相关的特性，这种模型称之为正交因子模型。然而，在实际的应用中，公因子之间常具有一定相关性。所以，在实际应用中，就需考虑通过斜交旋转得到斜交因子解。与正交因子解不同，在斜交因子解中，因子模型的公因子系数就不是变量与因子间的相关系数[22-24]。由此看来，一个完全的斜交因子解，除因子模型之外，还需要因子之间相关系数的相关矩阵和反映变量与因子间相关系数的因子结构。本小节先简单回顾一下正交因子模型，接着依次介绍斜交因子模型和基于斜交参考解的斜交旋转方法。

1.1 正交因子模型

通常对 n 个样本的 p 个特征依次进行观测，可以得到一个大小为 $p \times n$ 的原始观测数据矩阵 X 。对于每个特征通常包括 n 个样本的观测值，组成如下的随机观测向量 x

$$x = (x_1, x_2, \cdots, x_n)^T \tag{1}$$

式中，变量 x_i $(i = 1, \cdots, n)$ 表示某个特征的第 i 个样本的观测值。假设我们可以用少于 n 个样本（假设 m 个，$m < n$ ）来代表这组观测样本，在这种情况下某些样本可以视为是其它一些样本的线性组合。如此便可以减少样本的观测成本，进而也简化了观测系统。对于正交因子模型，x_i 可表示为如下的形式：

$$x_i = a_{i1} F_1 + a_{i2} F_2 + \cdots + a_{im} F_m + \varepsilon \tag{2}$$

式中，F_j $(j = 1, \cdots, m)$ 是公因子，它是每一个样本中都会出现的因子。a_{ij} 表示公因子 F_j 的因子载荷，也称为相对重要性（即权系数），ε_i 是个别样本所特有的一个特殊因子，a_i 是这一特殊因子的权系数。正交因子模型可表示为：

$$\underset{n\times 1}{x} = \underset{n\times m}{A} \times \underset{m\times 1}{F} + \underset{n\times n}{a} \times \underset{n\times 1}{\varepsilon} \tag{3}$$

式中，A 称为因子载荷矩阵，a 为特殊因子载荷，而 F 和 ε 分别为公因子和特殊因子。

因子载荷矩阵 A 中的元素 a_{ij} 表示第 i 个样本与第 j 个公因子的相关系数，依据其绝对值大小就可以判断样本的典型性（或代表性，相对重要性）。该值越大表明被选为典型性或代表性样本的可能性就越大。如果 $|a_{ij}|$ 的值越大，就表明第 i 个样本具有较大的载荷，比其它的样本具有更好的典型性或代表性，因此第 i 个样本就作为公因子 F_j 的典型性或代表性样本，这就是进行训练样本选择的理论基础或依据。

1.2 斜交因子模型

如果忽略了特殊因子，而且假设 x 已经标准化，正交因子模型可以表示为 $x = AF$ ，其中载荷矩阵 A 的元素 a_{ij} 刚好为 x_i 与 F_i 之间的相关系数。然而，如果公因子之间不是正交的，即公因子之间是相关的，斜交公因子将记为 T ，相应的载荷矩阵用 W 来表示。当忽略了特殊因子后，斜交因子模型可以写为如下形式

$$x = WT \tag{4}$$

或者写为

$$x_i = w_{i1} T_1 + w_{i2} T_2 + \cdots + w_{im} T_m \tag{5}$$

式中 $T_1, T_2, \cdots, T_m$ 为斜交公因子，它们可以视为斜交坐标的单位向量，x_i 视为斜交坐标系上的向量 P_i ，则 w_{ij} 为 P_i 在斜交因子轴 T_j 上的坐标，称它为斜交因子载荷。对于正交因子，因子模型就是因子解，这时因子模型和因子结构是一致的。但对于斜交因子，两者是有区别的。

实际上，斜交因子解是由正交因子解变换而来。

1)因子变换矩阵

所谓因子变换矩阵，就是从正交因子 F 变换（旋转）成斜交因子 T 的变换矩阵。如果把 $T_1, T_2, \cdots, T_m$ 看成斜坐标轴系的单位向量（即长度为 1），则 T_j 在 m 个正交因子 $F_1, F_2, \cdots, F_m$ 方向（正交坐标轴系方向）上的投影：$t_{1j}, t_{2j}, \cdots, t_{mj}$ （即 T_j 端点的正交坐标）就是斜坐标轴 T_j 相对于正交坐标因子轴 $F_1, F_2, \cdots, F_m$ 的夹

角余弦，其平方和等于 1，即

$$t_{ij} = |T_j| \cos(T_j, F_i) \tag{6}$$

$$T_j = (t_{1j}, t_{2j}, \cdots, t_{mj})^T \tag{7}$$

$$\sum_{i=1}^{m} t_{ij}^2 = 1 \quad (j = 1, 2, \cdots, m) \tag{8}$$

将公式（7）写成矩阵形式$T = (t_{ij})_{m \times m}$便是正交因子$F_1, F_2, \cdots, F_m$变换成斜交因子的变换矩阵。

2)斜交因子的相关系数矩阵

m 个斜因子轴既互不独立，必有相关，就要确定刻画 m 个斜交因子之间相关性的相关矩阵：$L = (l_{ij})$ $(i, j, = 1, 2, \cdots, m)$，其中元素$l_{ij}$表示斜因子$T_i$与$T_j$的相关系数$r_{T_iT_j}$，就是$T_i$与$T_j$之间的夹角余弦。从而斜交因子的相关系数矩阵为

$$L = T' T \tag{9}$$

3)因子结构矩阵（变量与斜交因子的相关系数矩阵）

$$S = (s_{ij}) \tag{10}$$

其中元素s_{ij} $(i = 1, 2, \cdots, n; j = 1, 2, \cdots, m)$表示第$i$个变量$x_i$的向量$P_i$在斜因子轴$T_i$上的投影。

因子结构矩阵S可以通过正交因子负荷矩阵A和相对应的因子变换矩阵T得到:

$$S = A T \tag{11}$$

4)斜交因子载荷矩阵

由式（5）斜交因子模型，于是有:

$$r_{iT_j} = E(x_i T_j) = w_{i1} r_{T_1T_j} + w_{i2} r_{T_2T_j} + \cdots + w_{im} r_{T_mT_j} \tag{12}$$

由（9）式，则有:

$$\begin{bmatrix} r_{1T_1} & r_{1T_2} & \cdots & r_{1T_m} \\ r_{2T_1} & r_{2T_2} & \cdots & r_{2T_m} \\ \vdots & \vdots & & \vdots \\ r_{nT_1} & r_{nT_2} & \cdots & r_{nT_m} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & \vdots & & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nm} \end{bmatrix} \cdot \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & & \vdots \\ l_{m1} & l_{m2} & \cdots & l_{mm} \end{bmatrix} \tag{13}$$

写成矩阵:

$$S = W L \tag{14}$$

把式（9）和式（11）代入式（14）式得:

$$AT = W[T'T] \tag{15}$$

从而有:

$$W = A[T']^{-1} \tag{16}$$

式（16）表示，当已求得正交因子负荷（已知A）时，只要知道T'就能求出斜交因子载荷矩阵W。但由于T不是对称矩阵，所以也可以用式（14）求W，即

$$S = W L^{-1} \tag{17}$$

至此，对于斜交因子解，它涉及到斜交因子的相关系数矩阵L、因子结构矩阵S以及斜交因子载荷矩阵W。无论要求L、S、W中的那一个都必须求出T，而由T的定义知道，它的第j列是斜交因子轴T_j在正交因子轴$F_1, F_2, \cdots, F_m$的坐标系中的方位余弦。那么如何求出T呢？为此，提出用斜交参考解来实现斜交旋转的方法。

1.3 基于斜交参考解的斜交旋转方法

在正交因子轴$F_1, F_2, \cdots, F_m$构成的坐标系中，斜交因子轴$T_1, T_2, \cdots, T_m$，是该坐标系中的m个单位向

量。在该坐标系中再引入 m 个单位向量：$\Lambda_1, \Lambda_2, \cdots, \Lambda_m$，称它们为斜交参考轴。

如果将 $T_1, T_2, \cdots, T_m$ 和 $\Lambda_1, \Lambda_2, \cdots, \Lambda_m$ 视为两组斜交因子轴，而 $T = (t_{ij})_{m\times m}$ 是正交因子轴 $F_1, F_2, \cdots, F_m$ 变换到斜交因子轴 $T_1, T_2, \cdots, T_m$ 的变换矩阵，它的元素 t_{ij} 刚好为 T_i 与 F_j 的夹角余弦。类似地，设 $\Lambda = (\lambda_{ij})_{m\times m}$ 是正交因子轴 $F_1, F_2, \cdots, F_m$ 变换到斜交参考轴 $\Lambda_1, \Lambda_2, \cdots, \Lambda_m$ 的变换矩阵，其中 λ_{ij} 是 Λ_i 与 F_j 的夹角余弦。现在来看 T 与 Λ 的关系，知道其中一个，如何求另一个？令

$$D = T'\Lambda = \begin{bmatrix} T_1' \\ T_2' \\ \vdots \\ T_m' \end{bmatrix} [\Lambda_1 \quad \Lambda_2 \quad \cdots \quad \Lambda_m] = \begin{bmatrix} T_1'\Lambda_1 & T_1'\Lambda_2 & \cdots & T_1'\Lambda_m \\ T_2'\Lambda_1 & T_2'\Lambda_2 & \cdots & T_2'\Lambda_m \\ \vdots & \vdots & & \vdots \\ T_m'\Lambda_1 & T_m'\Lambda_2 & \cdots & T_m'\Lambda_m \end{bmatrix} \tag{18}$$

D 是由斜主因子 $T_1, T_2, \cdots, T_m$ 与斜参考因子 $\Lambda_1, \Lambda_2, \cdots, \Lambda_m$ 的相关系数所组成的矩阵，由于 T_i 与 $\Lambda_j\ (i \neq j)$ 相互垂直，因此 $T_i'\Lambda_j = 0 (i \neq j)$，因而有：

$$D = diag(T_1'\Lambda_1, T_2'\Lambda_2, \cdots, T_m'\Lambda_m) \tag{19}$$

设 $\Lambda^{-1} = (\mu_{ij})_{m\times m}$，则：

$$T' = T'[\Lambda\Lambda^{-1}] = [T'\Lambda]\Lambda^{-1} = \begin{bmatrix} T_1'\Lambda_1\mu_{11} & \cdots & T_1'\Lambda_1\mu_{1m} \\ T_2'\Lambda_2\mu_{21} & \cdots & T_2'\Lambda_2\mu_{2m} \\ \vdots & & \vdots \\ T_m'\Lambda_m\mu_{m1} & \cdots & T_m'\Lambda_m\mu_{mm} \end{bmatrix} \tag{20}$$

因 T' 每一行元素平方和应为 1，即

$$[T_i'\Lambda_i]^2[\mu_{i1}^2 + \mu_{i2}^2 + \cdots + \mu_{im}]^2 = 1 \tag{21}$$

故

$$T_i'\Lambda_i = \frac{1}{\sqrt{\mu_{i1}^2 + \mu_{i2}^2 + \cdots + \mu_{im}^2}} = \frac{1}{(\sum_{l=1}^{m} \mu_{il}^2)^{\frac{1}{2}}} \tag{22}$$

由于所作斜交参考轴 Λ_i 垂直于斜因子轴 $T_l\ (i \neq l)$，所以 D 矩阵是一个对角阵，对角阵上的元素为 $T_i'\Lambda_i = 1/(\sum_{l=1}^{m} \mu_{il}^2)^{\frac{1}{2}}$。由此可以看出 T' 就等于 Λ 的逆（Λ^{-1}）按行“正规化（长度为 1）”而得到的矩阵。

1.4 训练样本的选择方法

实际上，斜交旋转是从某一正交因子解出发，经过变换（旋转）最终求出斜主因子解，这个解应包括斜主因子模型，因子结构矩阵和斜交因子相关矩阵。针对影像分类中训练样本选择的具体问题，本文提出了适合于实际应用的基于斜交因子模型的训练样本选择方法，现将步骤归纳如下。

设经方差极大旋转后正交因子模型为 A。

1)按文献[23]的方法求出初始载荷矩阵 A，再将 A 按行正规化得矩阵 A^*；

2)将 A^* 的各元素绝对值 k 次幂（k 为 >2 的适当正整数），并保留其原来的符号，得矩阵 H；

3)由已求出的 A^* 和 H，建立 A^* 对 H 的最小二乘拟合，即令

$$A^* C = H \tag{23}$$

其中 C 是 m 阶方阵，用 A^* 去左乘式（31）式两边，然后再用 $[A^{*'}A^*]^{-1}$ 左乘所得方程的两边，则可解得

$$C = [A^{*'}A^*]^{-1} A^{*'} H^* \tag{24}$$

4)将 C 按列正规化，得斜交参考矩阵 Λ；

5)将 Λ^{-1} 按行正规化得斜交变换矩阵 T'；

6)据式（9），式（14），式（16）式分别求得斜交因子解的相关系数矩阵 L、结构矩阵 S 以及斜交载荷矩阵 W，即：

$$L = T'T \tag{25}$$

$$S = AT \tag{26}$$

$$W = A[T']^{-1} \text{或} W = SL^{-1} \tag{27}$$

为了使所得的解更加“理想“，可让 k 为 2、3、4、…，依次进行旋转，比较各次的结果，k 值增大到因子相关矩阵相对稳定下来为止。根据实践经验，较为适当的 k 值通常在 2~4 之间。一般来讲，给定的一组变量之间相关关系越复杂，k 值也就越高。而 k 值应取多大，则需要具体问题具体分析。

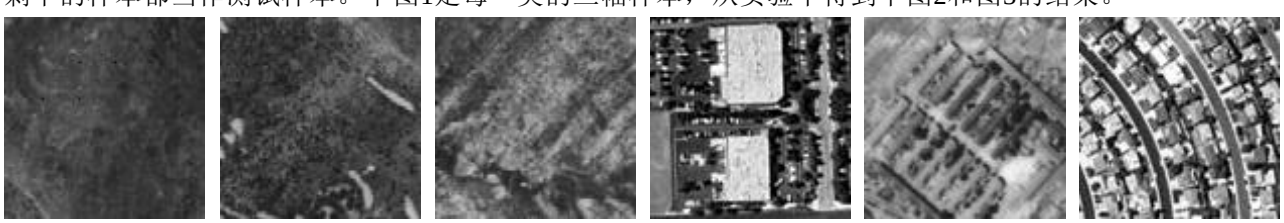
7）根据上述的斜交载荷矩阵 W，从原始的观测样本中选择出这类地物的代表性样本集（序号）Samples 为：$Samples = \{i \mid \max_{1\le i\le n}\{|w_{ij}|\}\ j = 1,\ldots,m\}$。$i$ 代表原始的观测样本的序号，而 $|w_{ij}|$ 表示对元素 w_{ij} 取绝对值。

8）重复上述1）～7），选出各类地物的代表性样本集，作为每一类的典型训练样本。

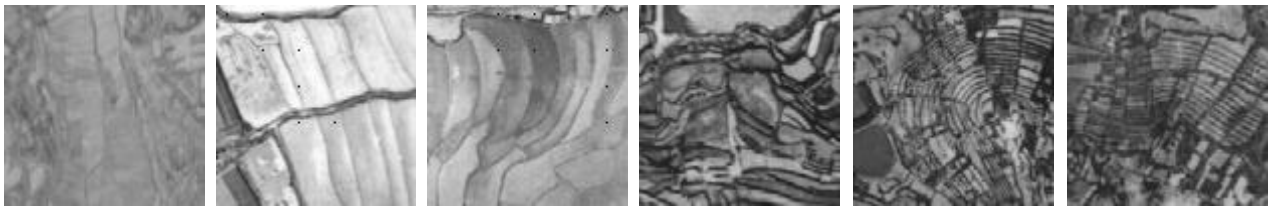
9）用上述方法选出的训练样本进行训练或学习，得到基于 Naive Bayes Classifiers 的分类器[27-30]，接着把所有采集的样本当作测试样本进行测试，最后统计总的分类精度。

2 实验与分析

为了验证本文提出方法的正确性和有效性，在本文的试验中，选取了10幅国内某个城市地区的23 cm×23 cm的黑白航空影像和澳大利亚某个地区的6幅23 cm×23 cm的黑白航空影像。根据野外调绘的结果，对这16幅大的航空影像人工分割为小块的684幅小图像，并将它们分成四类：山地（219幅）、水田（154幅）林区（154幅）和居民地（167幅），其中最小的为16×16像素，最大的为40×40像素。在实验中每一类的样本分为两组。一组是训练样本（样本序号在前50的样本作为训练样本），另一组是把被选为训练样本以外的剩下的样本都当作测试样本。下图1是每一类的三幅样本，从实验中得到下图2和图3的结果。



(a)灌木　　(b)居民地



(c) 水田　　(d) 山地

图 1 每一类中的三幅样本图像

Fig.1 Three samples for each class

2.1 所选训练样本分布情况分析

以居民地和林区两类为例，为了更好便于比较与分析，将它们显示在二维主分量特征平面上。每个样本的特征经过主分量变换，得到第一主分量和第二主分量。图 2 表示经过正交因子模型后被选中的具有代表性的典型样本。图中横坐标表示第一主分量，纵坐标表示第二主分量。图中的红色叉（×）号表示居民地，蓝色的（+）加号表示林区，圆圈（O）表示经过 Q 型因子模型后被选中的具有代表性的典型样本。而图 3 表示经过斜交因子模型后被选中的具有代表性的典型样本。

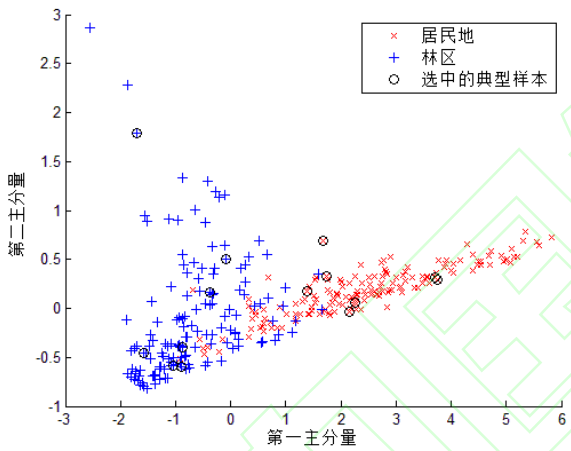


图 2 经过 Q 型因子模型后被选中的具有代表性的典型样本

Fig.2 selected samples of based on Q-factor model

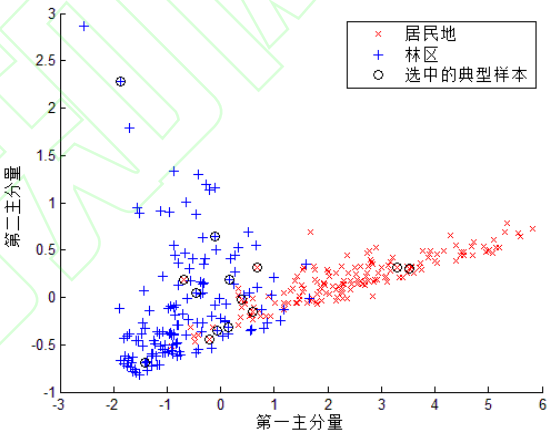


图 3 经过斜交因子模型后被选中的具有代表性的典型样本

Fig.3 selected samples of based on oblique factor model

从图 2 和图 3，不难发现，经过两种不同的因子模型，被选中具有代表性的典型样本有所不同。比较来看，经过斜交因子模型后被选中的具有代表性的典型样本比经过 Q 型正交因子模型后被选中的具有代表性的典型样本，分布相对分散一些。而在图 3 中，可以明显地发现，有两处地方，被选中的典型性样本在样本空间中“紧挨着”，似乎存在着一定的相关性。这很可能是因为正交因子模型忽视了公因子之间是相互独立的原因造成的，或者说由于忽略了现实中对样本（或变量）都发生影响的公因子之间往往是相互联系的（即公因子之间是相关）这个事实而形成的结果。

2.2 不同训练样本选择方法的精度

基于 R 型因子的分析方法是针对特征这个维度进行分析的，而基于对应分析的方法由于受到所选训练样本的数量必须小于等于特征提取数量的限制，所以在本文的比较实验中，只选用了基于 Q 型因子的分析方法与基于斜交因子模型的方法进行比较。图 4 是 4 种方法得到总的分类精度的比较结果，横轴表示被选中训练样本的数量，纵轴表示总的分类精度。图中米字形状的代表随机选取样本方法得到的结果，五角星形状的代表基于 Q 型正交因子模型得到的结果，正方形状的代表基于斜交因子模型得到的结果，圆形表示基于数据驱动选择的“最优”训练样本选择方法，搜索空间巨大。比如：在 684 个样本中选择 15 个“最优”样本，其组合有 C_{684}^{15}，所以在本文的试验中引入了遗传算法进行优化。下表 1 是四种方法在总的分类精度方面的数值比较，限于篇幅，只列出部分，训练样本数量 N 从 15 到 50，每增加 5 个训练样本列出总的分类精度值，最后两列分别是在训练样本数量在 15 至 50 之间每一种情况下统计所得总的分类精度的均值和

标准差。表 2 是 4 种方法在错分样本数量方面的比较结果。

表 1 四种方法的分类精度比较

Tab.1 the comparison of the four methods

N	15	20	25	30	35	40	45	50	mean	std
斜交因子	0.91	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.005
正交因子	0.88	0.89	0.88	0.90	0.90	0.90	0.90	0.90	0.90	0.006
随机选择	0.80	0.82	0.85	0.86	0.87	0.88	0.89	0.91	0.86	0.031
数据驱动	0.88	0.88	0.88	0.88	0.89	0.88	0.89	0.89	0.88	0.003

表 2 四种方法的错分样本数量比较

Tab.2 the comparison of the misjudged samples

N	15	20	25	30	35	40	45	50
斜交因子	58	47	45	43	45	41	43	43
正交因子	75	71	73	65	63	61	65	65
随机选择	128	112	96	86	80	77	69	59
数据驱动	73	70	75	75	68	70	68	68

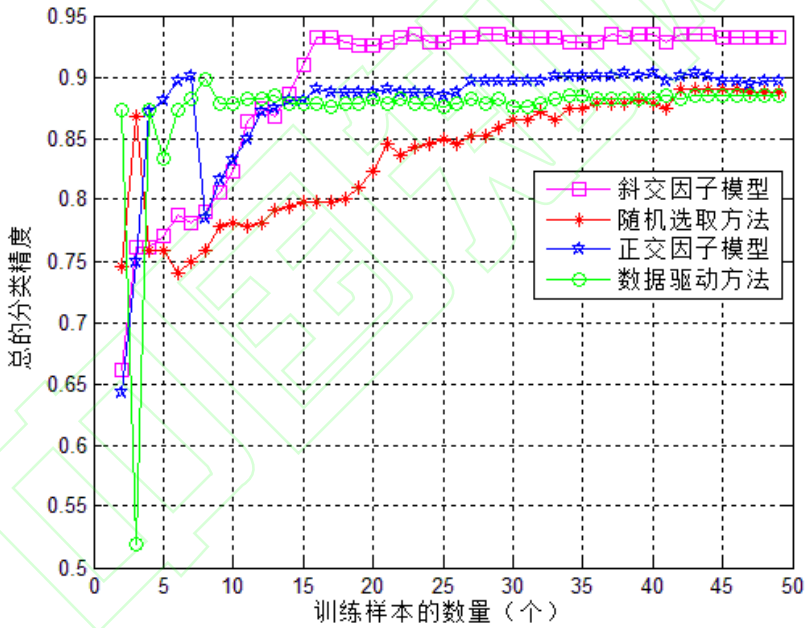


图 4 4 种方法的分类精度比较

Fig.4 The comparison of the four methods

从上图表中，可以明显地发现，总体上基于斜交因子模型的结果要好于基于 Q 型正交因子模型的结果，平均要高出 3%左右，而且总的分类精度的波动（即标准差）也从 0.006 1 下降到 0.004 6，这便可以表明所选的训练样本更具有典型性和代表性。当被选中的典型训练样本数较少时，两种方法的分类精度偏低，而且波动较大。随着所选典型样本数量的增多，两种方法的分类精度逐步提高，并一致趋向于稳定。这一方面说明，前期随着所选典型样本数量的增加，典型样本发挥着其自身典型性和代表性的作用。一旦达到典型性样本的数量或规模，或者说，这些典型性样本能够足够代表样本空间时，再继续增加样本，其作用或意义显得微乎其微。更为重要的是，由于松弛了 Q 型正交因子模型中对于公因子之间独立的假设条件，理论上更加符合实际的情况。而对于基于数据驱动的方法，很可能由于巨大的搜索空间更易陷入局部最优解或者受到初始解选择等原因，其结果表现一般。从实验的结果来看，在总的分类精度的波动方面，也可以明显地看到基于斜交因子模型比基于 Q 型正交因子模型要更好，更稳定一些，这也从侧面反映出基于斜交因子模型所选择出的典型样本比基于 Q 型正交因子模型所选择的典型样本更合理，而且更具有代表性或典型性。此外，从另外一个方面来看，这也为合理科学地确定训练样本的数量提供了较好的依据或参考。

3 结 语

在基于 Q 型正交因子模型的训练样本选择方法中，所采用的因子模型中的公因子都是正交的，而且进行的因子旋转也是正交旋转，这意味着在利用正交因子模型时必须假设那些公因子（即所选择的代表性或典型性的“公共样本”）是相互独立的。然而，在现实中对样本（或变量）都发生影响的公因子之间往往是相互联系（即公因子之间是相关的）。鉴于此，本文提出了一种适用于训练样本选择的斜交因子模型方法，该方法松弛 Q 型正交因子模型中对于公因子之间独立的假设条件，并在斜交参考解的基础上提出一种适合训练样本选择的近似求解斜交旋转的方法。实验与分析表明，本文提出的方法是可行，而且在分类精度方面要好于基于 Q 型正交因子模型的方法，而且所选的典型性样本更合理、也更具有代表性。此外，本文提出的方法也可以为合理科学地确定训练样本的数量提供了较好的依据或参考。

REFERENCES

[1] Adeli Ehsan, Li Xiaorui, Kwon Dongjin, Zhang Yong, Pohl Kilian M. Logistic Regression Confined by Cardinality-Constrained Sample and Feature Selection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(7): 1713-1728.

[2] Arnold Roger B., Wang Luke, Lopez Talle, James Sophie, Blute Nicole. Updating Lead and Copper Rule Sample-Site Selection: Best Practices From an Innovative Pilot Program [J]. Journal of American Water Works Association, 2020, 112(4): 22-31.

[3] Au Jessie, Youngentob Kara N, Foley William J, Moore Ben D, Fearn Tom. Sample selection, calibration and validation of models developed from a large dataset of near infrared spectra of tree leaves [J]. Journal of Near Infrared Spectroscopy, 2020. (Article in Press).

[4] Bellver Miriam, Salvador Amaia, Torres Jordi, Giro-i-Nieto Xavier. Mask-guided sample selection for semi-supervised instance segmentation[J]. Multimedia Tools and Applications, 2020. (Article in Press).

[5] da Silva, Marcus Vinicius Brito, de Carvalho, André Augusto Pacheco, Jacobs Arthur Selle, Pfitscher Ricardo José, Granville Lisandro Zambenedetti. Sample Selection Search to Predict Elephant Flows in IXP Programmable Networks [J]. Advances in Intelligent Systems and Computing, 2020, 1151: 357-368.

[6] Fernández Mariela, García Jesús E., Gholizadeh Ramin, González-López Verónica A. Sample selection procedure in daily trading volume processes [J]. Mathematical Methods in the Applied Sciences, 2020, 43(13): 7537-7549.

[7] He Kaixun, Wang Kai, Yan Yayun. Active training sample selection and updating strategy for near-infrared model with an industrial application [J]. Chinese Journal of Chemical Engineering, 2019, 27(11): 2749-2758.

[8] Kral, Jan, Gotthans Tomas, Marsalek Roman, Harvanek Michal, Rupp Markus. On feedback sample selection methods allowing lightweight digital predistorter adaptation [J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2020, 67(6): 1976-1988.

[9] Li Huiyong, Bao Weiwei, Hu Jinfeng, Xie Julan, Liu Ruixin. A training samples selection method based on system identification for STAP [J]. Signal Processing, 2018, 142: 119-124.

[10] Liu Jing, Zhu A-Xing, Rossiter David, Du Fei, Burt James. A trustworthiness indicator to select sample points for the individual predictive soil mapping method (iPSM) [J]. Geoderma, 2020, 373.

[11] Liu Xueqi, Zhu A-Xing, Yang Lin, Pei Tao, Liu Junzhi, Zeng Canying, Wang Desheng. A graded proportion method of training sample selection for updating conventional soil maps [J]. Geoderma, 2020, 357. (Open Access)

[12] Lu Qikai, Ma Yong, Xia Gui-Song. Active learning for training sample selection in remote sensing image classification using spatial information [J]. Remote Sensing Letters, 2017, 8(12): 1210-1219.

[13] Lu Wenbo, Ma Chaoqun, Li Peikun. Research on Sample Selection of Urban Rail Transit Passenger Flow Forecasting Based on SCBP Algorithm [J]. IEEE Access, 2020, 8: 89425-89438.

[14] Lu Yang, Ma Xiaolei, Lu Yinan. A cluster-based sample selection strategy for biological event extraction [C]. // Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering, 2019, p 72-77.

[15] Ma Jing, Hong Dezhi, Wang Hongning. Selective sampling for sensor type classification in buildings [C].//Proceedings - 2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks, IPSN 2020, p.241-252.

[16] Ng Wing W. Y., Jiang Xiaoxia, Tian Xing, Pelillo Marcello, Wang Hui, Kwong Sam Incremental hashing with sample selection using dominant sets. International Journal of Machine Learning and Cybernetics, 2020.(Article in Press)
[17] Rahimi Hamid. Considering factors affecting the prediction of time series by improving sine-cosine algorithm for selecting the best samples in neural network multiple training model [J]. Lecture Notes in Electrical Engineering, 2019, 480: 307-320.
[18] Tang Pengfei, Du Peijun, Lin Cong, Guo Shanchuan, Qie Lu. A Novel Sample Selection Method for Impervious Surface Area Mapping Using JL1-3B Nighttime Light and Sentinel-2 Imagery [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 3931-3941.
[19] Tran Nguyen, Abramenko Oleksii, Jung Alexander. On the sample complexity of graphical model selection from non-stationary samples [J]. IEEE Transactions on Signal Processing, 2020, 68:17-32.
[20] Varshavskiy Ilyas E., Dmitriev, Ivan A., Krasnova, Anastasiia I., Polivanov Vladimir V. Selection of Sampling Rate for Digital Noise Filtering Algorithms [C]. //Proceedings of the 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, 2020, p 932-935.
[21] Xu Xinzheng, Li Shan, Liang Tianming, Sun Tongfeng. Sample selection-based hierarchical extreme learning machine [J]. Neurocomputing, 2020, 377:95-102.
[22] YU Chong-wen, et al. Mathematical Geology and Application [M]. Beijing: Metallurgy Industry Press, 1980.(於崇文, 等.数学地质的方法与应用[M].北京：冶金工业出版社,1980.)
[23] YU Xin, ZHENG Zhaobao. Selcection of Training Samples Based on R-Q Factor Analysis [J]. Acta Geodaetica et Cartographica Sinica, 2007, 36(1)：67-71.(虞 欣, 郑肇葆. 基于Q型因子分析的训练样本的选择[J]. 测绘学报，2007, 36(1)： 67-71.)
[24] YU Xin, ZHENG Zhaobao. Selcection of Training Samples Based on Correspondence Analysis [J]. Acta Geodaetica et Cartographica Sinica, 2008, 37(2)：190-195.(虞 欣, 郑肇葆. 基于对应分析的训练样本的选择[J]. 测绘学报，2008, 37(2)：190-195.)
[25] Zhang Chenxiao, Wu Yifeng, Guo Mingming, Deng Xiaobo. Training sample selection for space-time adaptive processing based on multi-frames. Journal of Engineering [J], 2019, 20: 6369-6372.
[26] Zhang Xiwen, Seyfi Tolunay, Ju Shengtai, Ramjee Sharan, Gamal Aly El, Eldar Yonina C. Deep Learning for Interference Identification: Band, Training SNR, and Sample Selection [C] //IEEE Workshop on Signal Processing Advances in Wireless Communications, SPAWC, July 2019.
[27] YU Xin, ZHENG Zhaobao, TANG Ling, YE Zhiwei. Aerial Image Texture Classification Based on Naive Bayes Classifiers[J]. Geomatics and Information Science of Wuhan University, 2006, 31(2): 108-111.（虞欣, 郑肇葆, 汤凌, 叶志伟. 基于Naive Bayes Classifiers的航空影像纹理分类[J]. 武汉大学学报 ·信息科学版, 2006, 31(2): 108-111.）
[28] YU Xin, ZHENG Zhaobao, YE Zhiwei, TIAN Liqiao. Texture Classification Based on Tree Augmented Naive Bayes Classifier[J]. Geomatics and Information Science of Wuhan University, 2007, 32(4): 287-289.（虞欣, 郑肇葆, 叶志伟, 田礼乔. 基于 Tree Augmented Naive Bayes Classifier 的影像纹理分类[J]. 武汉大学学报·信息科学版, 2007, 32(4): 287-289.）
[29] ZHENG Zhaobao, PAN Li, ZHENG Hong. A Method of Image Texture Texton Classification with Markov Random Field[J]. Geomatics and Information Science of Wuhan University, 2017, 42(4): 463-467. doi: 10.13203/j.whugis20150615（郑肇葆, 潘励, 郑宏. 图像纹理基元分类的马尔柯夫随机场方法[J]. 武汉大学学报·信息科学版, 2017, 42(4): 463-467. doi: 10.13203/j.whugis20150615）
[30] ZHENG Zhaobao, ZHENG Hong. Image Classification Based on Data Gravitation[J]. Geomatics and Information Science of Wuhan University, 2017, 42(11): 1604-1607. doi: 10.13203/j.whugis20160457（郑肇葆, 郑宏. 利用数据引力进行图像分类[J]. 武汉大学学报·信息科学版, 2017, 42(11): 1604-1607. doi: 10.13203/j.whugis20160457）

Research on Oblique Factor Model for Selecting Training Samples

YU Xin[1] *ZHENG Zhaobao*[2] *LI Linyi*[2]

1 Beijing Institute of Petrochemical Technology, Beijing 102617, China,

2 School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079,China

Abstract: Objectives: Researchers notice that the quality of training samples will impact the effective of training

phase and then further will have an influence on the overall classification accuracy in the testing phase. In fact, representativeness or typicalness of training samples is able to reflect the quality of training samples in a way. Especially for the currently popular deep learning methods, it has needed thousands or millions of training samples. Therefore, how to reduce the number of training samples for deep learning method becomes a very important problem. In another hand, from the actual application angle, it is also very expensive. Therefore, we research one method of reducing the training samples as less as possible based on the representativeness or typicalness of training samples. **Method:** selection of training samples based on oblique factor model is proposed and it relaxes the independent condition among common factors in the orthogonal factor model, which is able to better describe the real world. **Results:** Experimental results show the proposed method is feasible and effective and it is able to select more representative training samples than the method of selection of training samples based on orthogonal factor model and achieve better performance in the overall classification precision and stability. Experimental results show that selection of training samples based on oblique factor model outperforms selection of training samples based on orthogonal factor model. And the distribution of selected samples becomes more decentralized and reasonable and the overall classification accuracy averagely improves about 3%. **Conclusions:** the proposed method, not only supports how to optimize capturing data in the theory, but also is able to guide how to effectively capture data in the actual application.

Key words: Oblique factor model; training samples; image classification; orthogonal factor model; samples selection

First Author: YU Xin, Ph.D., professor. His research interests include photogrammetry and remote sensing, image interpretation and artificial intelligence. Email: china_yuxin@163.com

Corresponding author: LI Linyi, Ph.D., associate professor. His research interests include remote sensing image interpretation, and intelligence computing. Email: lilinyi@whu.edu.cn

Foundation Support: the National Key Research and Development Program of China (No. 2018YFC0407804).