最大似然分类的比较应用魏尔斯特拉斯变换前后

Muhammad Shoaib、Zaka Ur Rehman、Muhammad Qasim

摘要— 本文旨在通过定性和定量方法，对多光谱数据进行最大似然（ML）分类。最大似然是一种基于经典贝叶斯定理的监督分类算法。它利用判别函数，将像素分配到似然最高的类别。类别均值向量和协方差矩阵是该函数的关键输入，可根据特定类别的训练像素进行估算。最大似然算法应用于数据前需要满足一些假设。本文将比较应用魏尔斯特拉斯变换前后最大似然分类（ML）的结果，并考察高分辨率 Quickbird 卫星图像训练像素的准确率差异。我们还使用主成分分析（PCA）进行降维，并检查波段中的变化。结果表明，在决策空间中，各类别均值之间的分离程度是影响分类准确率的主要因素；应用魏尔斯特拉斯变换后，最大似然（ML）的分类准确率高于未应用该变换时的准确率。

索引词— 最大似然分类（ML）、魏尔斯特拉斯变换、贝叶斯定理、监督分类、主成分分析（PCA）。

I. 引言

卫星图像用途广泛。随着卫星技术的发展，各国都希望借助卫星图像解决自身面临的问题。卫星图像可应用于多个学科领域，例如气象学、环境科学、地理信息系统、地质学、地球物理学、工程与建筑、国防与情报等，不胜枚举。

卫星图像包含大量关于地球的信息，而主要问题在于如何从卫星图像中提取有用信息。研究人员可能面临多个问题，例如如何对土地覆盖类型进行分类、如何确定大气中气溶胶粒子的深度、森林气候以及洪水预测等。研究人员通过分析卫星图像来解决这些问题。

本文的主要目标是使用最大似然分类算法对土地覆盖类型进行分类。首先，我们在不使用魏尔斯特拉斯变换（WT）的情况下对土地覆盖类型进行分类，然后通过应用魏尔斯特拉斯变换（WT）比较结果。

最大似然（ML）是一种源自贝叶斯定理的监督分类方法。贝叶斯定理指出，后验分布 P(i|$\boldsymbol{\theta}$)，即特征向量为 $\boldsymbol{\theta}$ 的像素属于类别 i 的概率，表示为：

$$p(i|\theta) = \frac{p(\theta|i) * p(i)}{p(\theta)} \quad \rightarrow \quad (1)$$

其中，P($\boldsymbol{\theta}$|$\boldsymbol{i}$) 是似然函数，P(i) 是先验信息，即类别 i 出现在研究区域中的概率；$\boldsymbol{P(\theta)}$ 是观测到该概率，表示为：

$$P(\theta) = \sum_{i=1}^{M} P(i|\theta)p(i) \quad \rightarrow \quad (2)$$

其中，M 为类别数量。$\boldsymbol{P(\theta)}$通常被视为归一化常数，以确保 ∑P(i|$\boldsymbol{\theta}$) 的总和为 1。根据以下规则，将像素 x 分配到类别 i：

$$x \in i \quad if\ P(i|\theta) > P(j|\theta)\ for\ all\ i \neq j \quad \rightarrow \quad (3)$$

每个像素都会被分配到似然最高的类别；如果所有概率值都低于用户设定的阈值 [1]，则将其标记为未分类。

最大似然（ML）假设给定类别 i 内的数据服从多元高斯分布。

首先，我们在不检验正态性假设的情况下实施最大似然（ML）分类算法，然后对图像数据应用魏尔斯特拉斯变换后再进行最大似然（ML）分类；我们发现，应用魏尔斯特拉斯变换后，数据服从正态分布。应用魏尔斯特拉斯变换（WT）后的准确率有所提高，这是因为数据的分布呈正态分布，而机器学习分类要求数据服从正态分布。

II. 研究区域

我们使用 Quickbird 卫星图像进行分析。Quickbird 是由 DigitalGlobe 所有并于 2001年发射的高分辨率商业卫星。Quickbird 使用 Aerospace全球成像系统2000（BGIS 2000）。Quickbird卫星采集了分辨率为61厘米的全色（黑白）图像，以及分辨率为2.44米（轨道高度为450千米）至1.63米（轨道高度为300千米）的多光谱影像；随着任务寿命末期轨道高度降低，分辨率也随之变化 [2]。

我们用于分析的 Quickbird 卫星图像是意大利罗马的样本图像，带有坐标信息，包含四个波段：蓝色（450-520 nm）、绿色（520-600 nm）、红色（630-690 nm）和近红外（760-900 nm）。

$12^{o}34'21.43"E, 41^{o}52'25.49"N$

III. 方法论

方法论是对应用于某一研究领域的方法进行的系统性理论分析。它包含对某一知识分支相关的方法体系和原则的理论分析（维基百科）。

详细的方法论见图1。

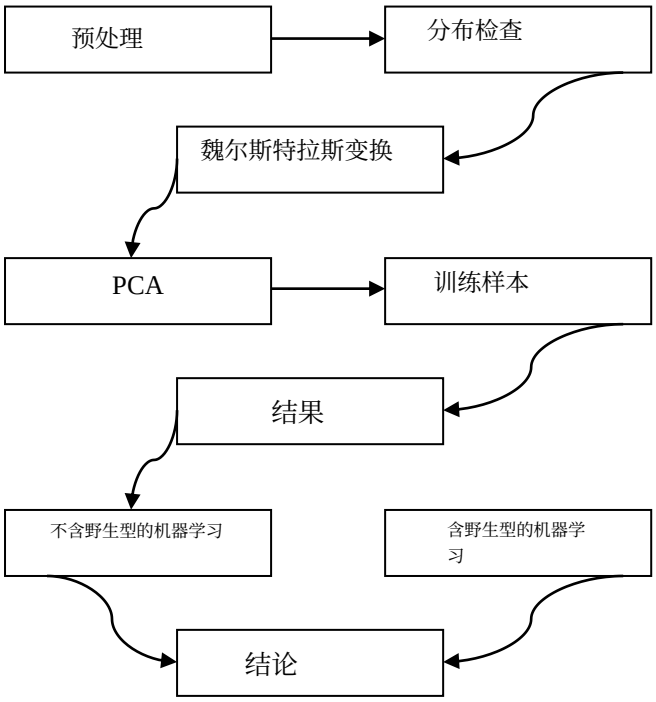


图1. 数据处理流程图

IV. 预处理

预处理是对图像进行操作的统称，其抽象层次最低，输入和输出均为强度图像。预处理的主要目的是改善图像数据，以抑制不必要的失真，或增强对后续处理至关重要的某些图像特征。

本文需要进行分类，因此不需要进行辐射校准或其他处理。但我们需要检查每个波段的分布是否为正态分布，这是机器学习分类的主要假设。

A. 分布检查

我们使用图像直方图来检查每个波段的分布。“图像直方图”是一种直方图，用于以图形方式表示数字图像中的色调分布。它绘制每个色调值对应的像素数量。通过查看特定图像的直方图，观察者可以一目了然地判断其整体色调分布。我们为每个波段分别制作直方图，以检查图像中各波段的分布。

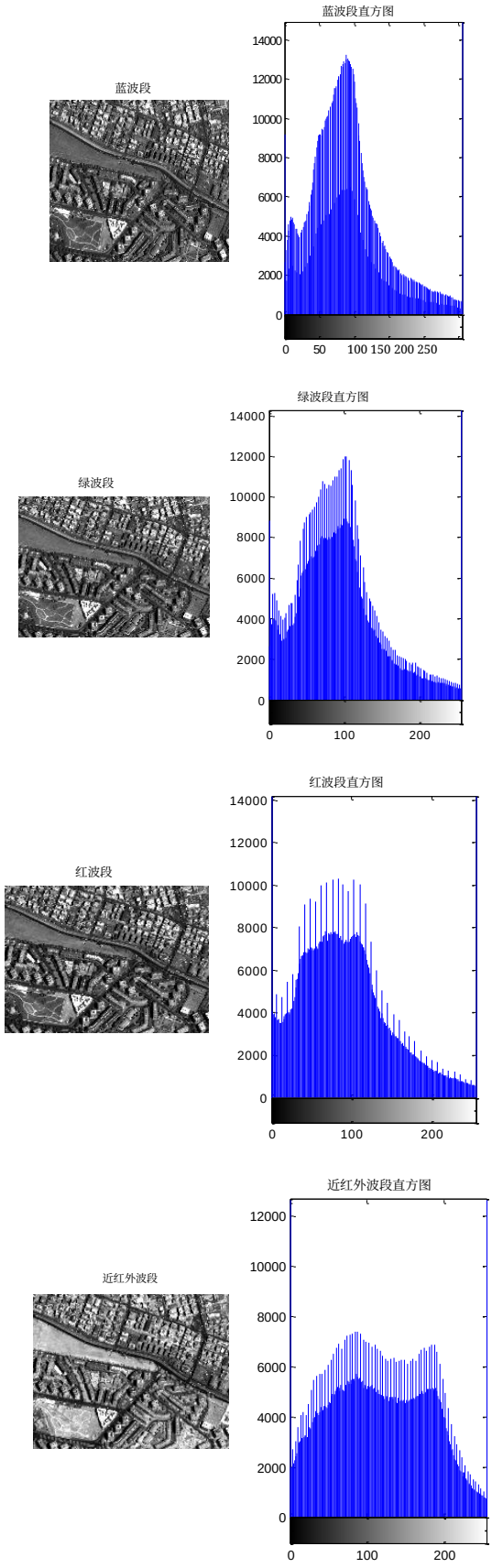


图 2. 直方图及其对应的波段

我们可以在图2中看到，每个波段的分布都不是高斯分布，而我们知道，对于机器学习分类算法，每个波段的分布都必须是高斯分布。在这种情况下，前3个波段的分布呈正偏态，而波段 4 的直方图看起来像高斯分布，但并不完全是高斯分布。因此，我们的目标是在使数据呈高斯分布之前和之后，检查分类的准确率。我们使用魏尔斯特拉斯变换使分布呈高斯分布。

B. 魏尔斯特拉斯变换

正如我们所见，我们的数据并不服从高斯分布，因此，要使数据呈高斯分布，我们必须进行变换，使像素值呈高斯分布。

研究人员使用各种类型的变换使数据呈高斯分布。本文使用一种称为魏尔斯特拉斯变换的变换。

魏尔斯特拉斯变换也称为高斯变换、高斯-魏尔斯特拉斯变换和希勒变换，与一维流动的热方程解密切相关。它也是卷积变换的一个特例，但被认为是Hirschmann和Widder提出的卷积变换理论中的奇异情形 [3]。

函数 f(y) 的魏尔斯特拉斯变换 F(x) 定义如下：

$$F(x) = \frac{1}{\sqrt{4\pi}} \int_{-\infty}^{\infty} f(y) e^{-\frac{(x-y)^2}{4}} dy$$

$$= \frac{1}{\sqrt{4\pi}} \int_{-\infty}^{\infty} f(y) e^{-\frac{y^2}{4}} dy$$

即 f 与高斯函数的卷积。除 F(x) 外，我们也将其写作 $W[f](x)$。请注意，F(x) 不一定对每个实数 x 都存在，因为定义积分可能不收敛 ***[3]***。

当高斯滤波器使用高斯函数修改输入信号时，这称为魏尔斯特拉斯变换。

高斯函数的二维数学表达式如下：

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right) \rightarrow (1)$$

其中，ρ 表示 X 和 Y 之间的相关性，$\sigma_x>0$ 和 $\sigma_y>0$。在这种情况下，

$$\mu = \begin{pmatrix} \sigma_x \\ \sigma_y \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}.$$

在二元情形下，多元正态性的第一个等价条件可以放宽：只需验证可数个不同的线性组合 X 和 Y 均服从正态分布，即可得出向量 ***[X Y]'*** 服从二元正态分布。

高斯滤波

用于模糊图像并去除噪声和细节。高斯滤波器使用上述高斯函数，该函数见公式 1。

应用高斯变换或魏尔斯特拉斯变换后，我们可以比较原始假彩色合成（FCC）图像与变换后的假彩色合成（FCC）图像。



（a）原始假彩色合成（FCC）图像



（b）变换后的假彩色合成（FCC）图像 图 3。

在图3中，我们可以清楚地看到原始假彩色合成（FCC）图像与变换后的假彩色合成（FCC）图像之间的差异。应用变换后，可以看到变换后的图像比原始图像更模糊。

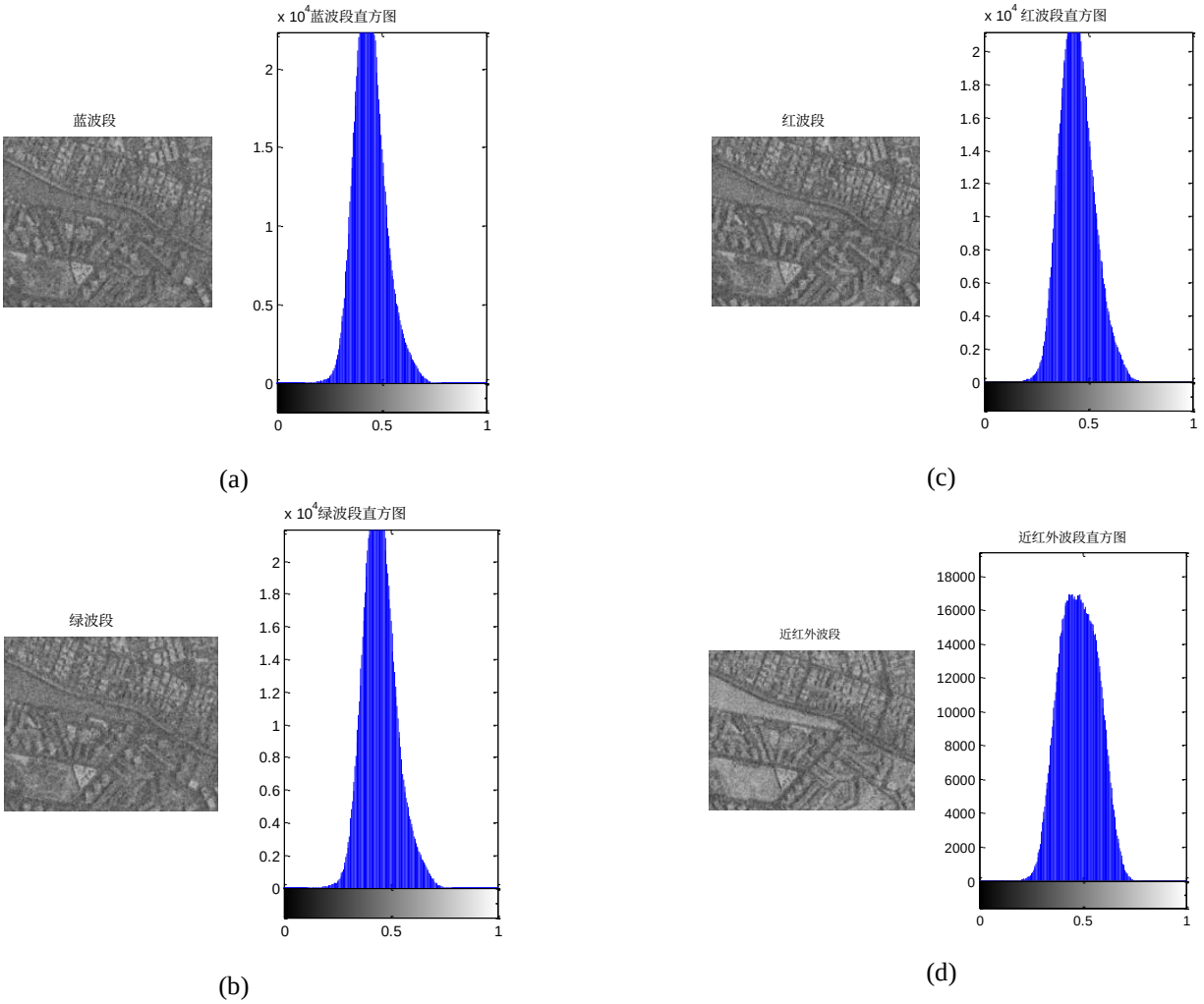


图4。（a）变换后的图像的波段1直方图，（b）变换后的图像的波段2直方图，（c）变换后的图像的波段3变换后的图像，（d）变换后的图像的波段4直方图

在图4中，观察变换后的假彩色合成（FCC）图像各波段的直方图，可以发现每个波段的分布都符合高斯分布。现在可以看到，图像的像素值符合标准高斯分布。

C. 降维

卫星图像处理中的降维是指减少信息本质相同的波段数量。当波段数量较多时，便会使用降维。

我们有一幅包含 4 个波段的 QuickBird图像，并通过主成分分析来判断应忽略哪个波段。我们对变换后的假彩色合成（FCC）图像进行主成分分析。

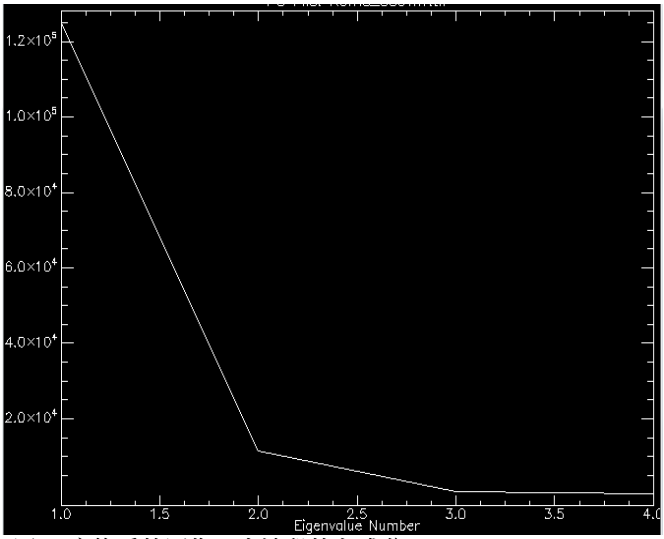


图 4. 变换后的图像四个波段的主成分。

在上图中，我们可以看到，波段 3（红色）和波段 4（近红外）包含相同的信息，因此在分析中应忽略波段 3 或波段 4。为此，我们将使用蓝色、绿色和近红外这三个波段进行分类。

D. 训练样本

通过为每个类别选择一个或多个多边形来确定训练区域。落在训练区域内的像素被视为该类别的训练像素。为某个类别选择合适的训练区域时，需要考虑的重要属性包括其均匀性，以及它们在整个图像中对同一类别的代表性[4]。

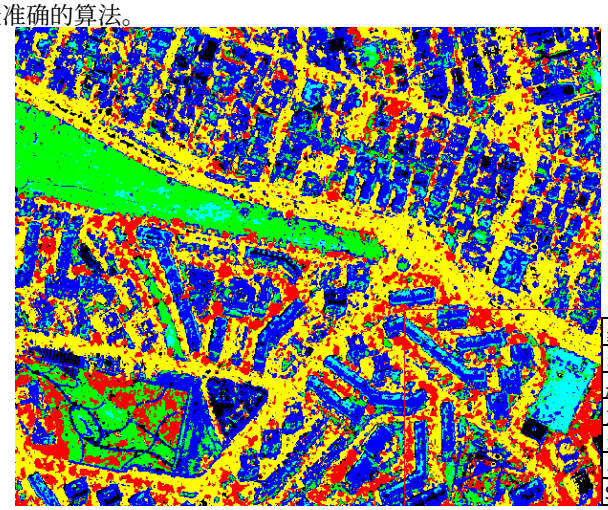
我们采用分层随机抽样来选择训练像素。分层随机样本是一种总体样本，要求将总体划分为更小的组，称为“层”。可以从每个层或组中抽取随机样本。

V. 结果

与平行六面体分类算法等经典算法相比，最大似然被认为是最准确的算法。但是，我们知道机器学习分类算法假设每个波段的分布都必须服从高斯分布。我们发现波段分布并非高斯分布，因此通过应用高斯变换或魏尔斯特拉斯变换使其成为高斯分布。现在，我们将检查分布服从高斯分布和不服从高斯分布时，分类类别的准确率会受到什么影响。

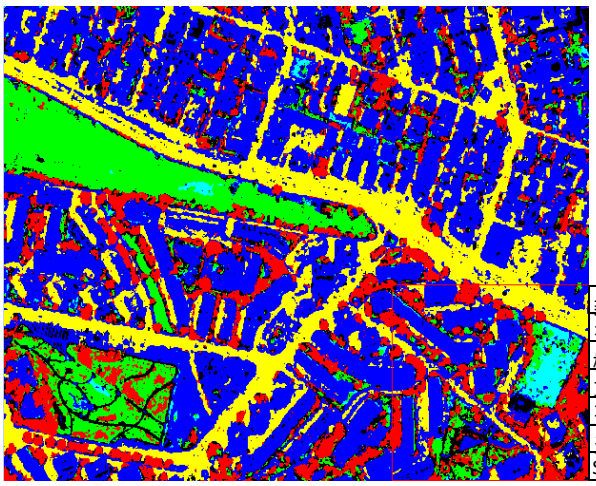
A. 魏尔斯特拉斯变换前的分类

将类别赋予合适的颜色后，机器学习分类的结果如图5 (a) 所示：草地（绿色）、城市区域（蓝色）、道路（黄色）、土壤（青色）、树木（红色）和未分类（黑色）。此外，还计算了各类别的面积（以百分比和平方米表示）；面积最大的类别是城市区域。在这种情况下，数据分布不是高斯分布。从图 5 (a) 中可以清楚地看到，一些建筑物被错误地分类为道路，这是因为我们对非正态数据应用了机器学习分类。



类别	颜色	面积占比	% 面积（平方米）	颜色
未分类：	[黑色]	6.44%	(26,783.6400 平方米)	
树木	[红色]	14.20%	(59,095.0800 平方米)	
草地	[绿色]	13.63%	(56,719.0800 平方米)	
城市区域	[蓝色]	31.45%	(130,868.2800 平方米)	
道路	[黄色]	30.31%	(126,133.9200 平方米)	
Soil	[青色]	3.97%	(16,532.6400 平方米)	

(a)



类别	颜色	面积（%）	面积（平方米））	颜色
未分类：	[黑色]	10.30%	(42,851.1600 平方米)	
树木	[红色]	13.75%	(57,232.0800 平方米)	
草地	[绿色]	10.34%	(43,005.2400 平方米)	
城市区域	[蓝色]	44.79%	(186,373.0800 平方米)	
道路	[黄色]	19.39%	(80,677.4400 平方米)	
Soil	[青色]	1.44%	(5,993.6400 平方米)	

(b)

图 5. (a) 最大值 原始图像的最大似然分类；（b）变换后的图像的最大似然分类

表 I. 原始分类图像的混淆矩阵

总体精度 = (18318/21394) = 85.62%							
卡帕系数 = 0.8135							
	地面真值（百分比）						
类别	颜色	树木	草地	城市区域道路		Soil	颜色
未分类	[黑色]	0.4	0.05	7.84	2.33	0	
树木	[红色]	94.38	8.06	0.94	5.28	0	
草地	[绿色]	1.64	88.71	2.66	0.03	3.41	
城市区域	[蓝色]	0	1.06	70.71	1.19	2	
道路	[黄色	3.58	0	4.61	91.17	0	
Soil	[青色]	0	2.11	13.23	0	94.59	
	总计	100	100	100	100	100	

表 II. 变换后分类图像的混淆矩阵

总体精度 = (20156/21394) = 94.2133%							
卡帕系数 = 0.9236							
	地面真值（百分比）						
类别	颜色	树木	草地	城市区域道路		Soil	颜色
未分类	[黑色]	0.85	0.36	4.36	0.52	0	
树木	[红色]	98.51	4.72	0.09	0.08	0	
草地	[绿色]	0.65	94.71	1.92	0	0.3	
城市区域	[蓝色]	0	0.2	86.5	1.33	0.52	
道路	[黄色]	0	0	3.19	98.07	0	
Soil	[青色]	0	0	3.94	0	99.19	
	总计	100	100	100	100	100	

机器学习分类的准确率评估通过混淆矩阵（有时称为误差矩阵）确定，该矩阵逐类别比较参考数据（地面真值）与相应分类结果之间的关系 [1]。

结果见表 1。表 1 中的对角元素表示正确分类的百分比，也称为生产者精度。生产者精度是衡量特定分类方案准确率的指标，表示某一特定地面类别被正确分类的百分比。

卡帕系数 κ 是衡量分类准确率的另一项指标，它综合考虑非对角元素和对角项，因此比总体精度能更稳健地评估准确率。

机器学习分类的总体精度为 85.62%，卡帕系数为 0.8135。

B. 魏尔斯特拉斯变换后的分类

当我们对数据应用高斯变换时，可以看到波段分布从偏斜对称变为近似高斯分布。

当我们对变换后的图像应用机器学习分类算法时，可以看到总体精度从85.62%提高到94.21%，这是因为图像数据的像素值遵循高斯分布，同时卡帕系数的值κ 也从0.8135提高到0.9236。在表2中，我们可以看到应用魏尔斯特拉斯变换或高斯变换后的分类结果。因此，当数据遵循高斯分布时，可以说机器学习分类效果更好。

VI. 结论

本研究对罗马热带土地覆盖类型的机器学习分类进行了详细分析，并得出若干结论。当分布符合高斯分布时，机器学习能够以较高的一致性对研究区域内的类别进行分类；当分布不符合高斯分布时，总体精度会下降近 9%。

。机器学习将研究区域分为

5　　88　　κ= 个类别，准确率为 % 0.8425（。）。机器学习根据每种覆盖类型的已知属性对像素进行分类，但生成的类别在统计上可能无法区分。

致谢

作者谨此感谢 Waqas A. Qazi 博士提供的宝贵意见和想法，并感谢 Muhammad Qasim 协助实施魏尔斯特拉斯变换。

参考文献

[1] T.M. Lillesand、R.W. Kiefer 和 J.W. Chipman，《遥感与图像解译》，John Wiley & Sons，霍博肯，新泽西州，美国，2004年。[2] DigitalGlobe.com。DigitalGlobe。2014年2月12日。检索于2014年6月19日。[3] Ahmed I. Zayed，《函数与广义函数变换手册》，第18章，第322–324页。[4] J.B. Campbell，《遥感导论》，Taylor & Francis，伦敦，2002年。

Muhammad Shoaib 是巴基斯坦伊斯兰堡空间技术研究所空间科学系的遥感与地理信息系统硕士研究生。

Zaka Ur Rehman 是空间技术研究所（巴基斯坦伊斯兰堡）空间科学系遥感与地理信息系统专业的硕士生。

Muhammad Qasim 获得了巴基斯坦阿伯塔巴德 COMSATS 信息技术学院的数学硕士学位。