# 马氏距离分类器

姓名：阮思琳 学号：2022302131290

##### 原理

1. **马氏距离**

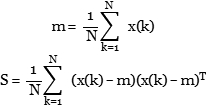
考虑一个由d维向量x组成的样本集合X，若该集合中的样本x服从均值为m，协方差矩阵为S的多元高斯分布，其表达式如下所示：

则：

马氏距离是用来表示数据的协方差距离，它是一种有效的计算两个未知的样本集的相似度的方法，考虑到了各个特性之间的联系。它的原理是通过计算向量x到每一个类均值向量的马氏距离，将x归于离它最近的均值所属的类。

1. **最小马氏距离[分类器](https://so.csdn.net/so/search?q=%E5%88%86%E7%B1%BB%E5%99%A8&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/yuyaweibest/article/details/_blank)**

假设服从多元高斯分布的样本集合  中有  个样本, 第  个样本为 , 则样本集合  的均值  ，协方差  的估计值由下面的公式给出：



其中,  均为  维向量,  为  的矩阵。

考虑  类样本类 , 均服从均值为  与协方差矩阵为  的多元高斯分布。令分类器函数为:



则



称其为最小马氏距离分类器。

1. **混淆矩阵**

混淆矩阵是用于评估分类器性能的工具，特别是在多类别分类中。对于马氏距离分类器，混淆矩阵的表格如下所示：

| **实际类别/预测类别** | **类别1** | **类别2** | **…** | **类别n** |
| --- | --- | --- | --- | --- |
| 类别1 | TP | FP | … |  |
| 类别2 | FN | TN | … |  |
| … |  |  |  |  |
| 类别n |  |  |  |  |

其中，

* TP（True Positive）：实际属于类别i且被正确预测为类别i的样本数量。
* FP（False Positive）：实际不属于类别i但被错误预测为类别i的样本数量。
* TN（True Negative）：实际不属于类别i且被正确预测为非类别i的样本数量。
* FN（False Negative）：实际属于类别i但被错误预测为非类别i的样本数量。

对于  分类模型，混清矩阵为一  阶方阵:



其第 行第  列元毒定义为:



即混清矩阵的第  行第  列为真实标签为  的样本被预测为  类的比例。

通过混淆矩阵，可以计算出一系列分类器的性能指标，如准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 F1 值等，它们的计算公式如下：

* 准确率（Accuracy）：
* 精确率（Precision）：
* 召回率（Recall）：
* F1 值：

##### 编程过程

1. ENVI影像数据导出

在地理空间数据云网站下载TM遥感影像，TM影像包含7个波段，，需要将这六个波段输出为一个.tif文件。

在ENVI软件中打开带有元数据的GeoTIFF，然后将实习中的要求的六个波段导出为一个tif文件，由于直接下载的图像太大，且不便于处理，我通过空间裁剪进行剪切为1024\*1024大小的图像。

1. ENVI Classic ROI 区域提取

接下来进行ROI区域提取

首先使用ENVI Classic工具打开导出的tmwuhan.tif文件，在弹出的窗口选择要假彩色增强的波段。

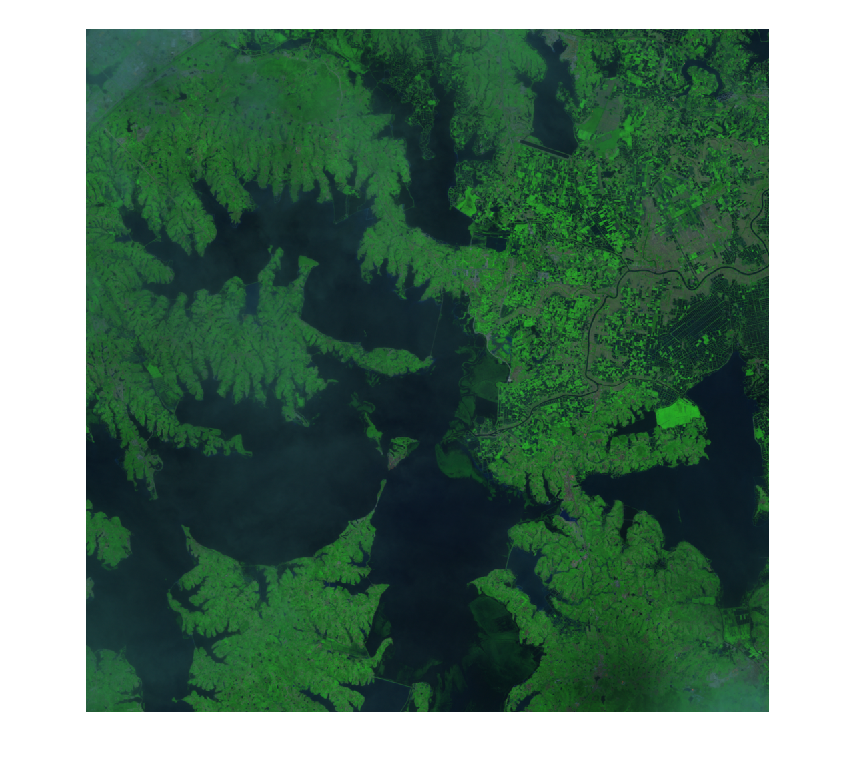
然后选择图像视图窗口菜单Region of Interest命令，通过ROI TOOL选择感兴趣区域绘制窗口并在ROI Name中定义其类型。

选择Options→Compute ROI Separability进行训练样本可分离性计算。在Select Input File for Separability窗口中选择计算训练可分离性的图像文件,查看训练样本分离性报告。(ENVI为每一个感兴趣区组合计算Jeffries-Matusita距离和Transformed Divergence。在窗口底部，根据可分离性值的大小，从小到大列出感兴趣区组合。这两个参数值为0 2.0，大于1.9说明样本之间可分离性好，属于合格样本；小于1.8，需要重选择样本；小于1，考虑将两类样本合成一类样本。)

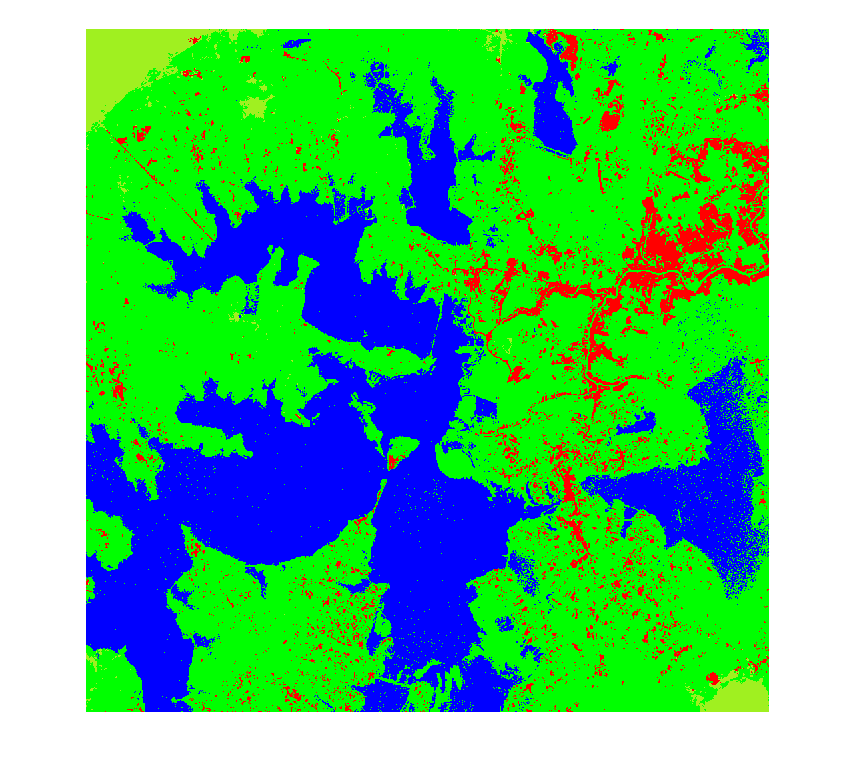
1. txt数据转mat数据
2. clear;clc;
3. filename="Roi.txt";
4. % 波段数
5. n=6;
6. % 打开文件
7. fid=fopen(filename);
8. % 读取头行
9. fgetl(fid);
10. % 读取分类数
11. NumOfROIs\_line=fgetl(fid);
12. NumOfROIs=str2double(NumOfROIs\_line(19));
13. % 读取图像大小
14. size\_line=fgetl(fid);
15. Size(1)=str2double(size\_line(19:22));
16. Size(2)=str2double(size\_line(26:29));
17. % 读取每个类别像素的数量
18. NumOfPerROIs=zeros(NumOfROIs,1);
19. for i=1:NumOfROIs
20. fgetl(fid);fgetl(fid);fgetl(fid);
21. line=fgetl(fid);
22. NumOfPerROIs(i)=str2double(line(13:length(line)));
23. end
24. % 关闭文件
25. fclose(fid);
26. % 读取矩阵数据
27. data=readmatrix(filename);
28. clear i line size\_line fid filename NumOfROIs\_line
29. end\_index=0;
30. new\_data=cell(1,NumOfROIs);
31. for i=1:NumOfROIs
32. begin\_index=end\_index+1;
33. end\_index=begin\_index+NumOfPerROIs(i)-1;
34. new\_data{i}=data(begin\_index:end\_index,8:(8+n-1));
35. end
36. data=new\_data;
37. clear begin\_index end\_index i ans new\_data;
38. % 保存结果
39. save("Roi.mat");
40. 马氏距离分类器
41. clear;clc;
42. load("Roi.mat")
43. % 划分训练集和测试集
44. test=[];
45. training=cell(1,NumOfROIs);
46. for i=1:NumOfROIs
47. test=cat(1,test,cat(2,data{i}(1:100,:),ones(100,1)\*i));
48. training{i}=data{i}(101:end,:);
49. end
50. % 使用训练集求各类别的均值向量以及协方差矩阵  ENVI导出的roi.txt
51. means=cell(1,NumOfROIs);
52. covs=cell(1,NumOfROIs);
53. for i=1:NumOfROIs
54. means{i}=mean(training{i})';
55. covs{i}=cov(training{i});
56. end
57. % 计算测试集的马氏距离
58. r2=zeros(400,NumOfROIs);
59. for i=1:NumOfROIs
60. du=(test(:,1:n)-repmat(means{i}',100\*NumOfROIs,1))';
61. inv\_cov=pinv(covs{i});
62. r2(:,i)=sum((du'\*inv\_cov.\*du'),2);
63. end
64. [~,classify]=min(r2,[],2);
65. % 使用混淆矩阵进行精度评定
66. ConfusionMatrix=zeros(NumOfROIs,NumOfROIs);
67. for i=1:100\*NumOfROIs
68. ConfusionMatrix(test(i,end),classify(i))=ConfusionMatrix(test(i,end),classify(i))+1;
69. end
70. precision=sum(diag(ConfusionMatrix))/(100\*NumOfROIs);
71. % 对整张图像进行分类
72. % 读取原始影像
73. tif=imread("tmpicture.tif");
74. [M,N,bands]=size(tif);
75. % 假彩色增强显示原始图像
76. imshow(cat(3,tif(:,:,6),tif(:,:,4),tif(:,:,3)))
77. tif=double(reshape(tif, M\*N,[]));
78. % 计算各类别马氏距离并分类
79. r2\_s=[];
80. for i=1:NumOfROIs
81. dpix=tif-means{i}';
82. r2=sum((dpix/covs{i}).\*dpix,2);
83. r2\_s=cat(2,r2\_s,r2);
84. end
85. % 根据类别赋予不同的颜色
86. [~,type]=min(r2\_s,[],2);
87. classify=reshape(type,M,N);
88. colors\_R=[0,255,0,160];
89. colors\_B=[255,0,0,32];
90. colors\_G=[0,0,255,240];
91. new\_tif=cat(3,colors\_R(classify),colors\_G(classify),colors\_B(classify));
92. new\_tif=uint8(new\_tif);
93. figure,imshow(new\_tif)
94. clear begin\_index end\_index i ans;

##### 结果分析

1. 假彩色图像



1. 分类处理图像



**分类类别：**

* Water（水域）——蓝色
* Land（陆地）——绿色
* Towns（城镇）——红色
* City（城镇）——黄色

**图像结果分析：**

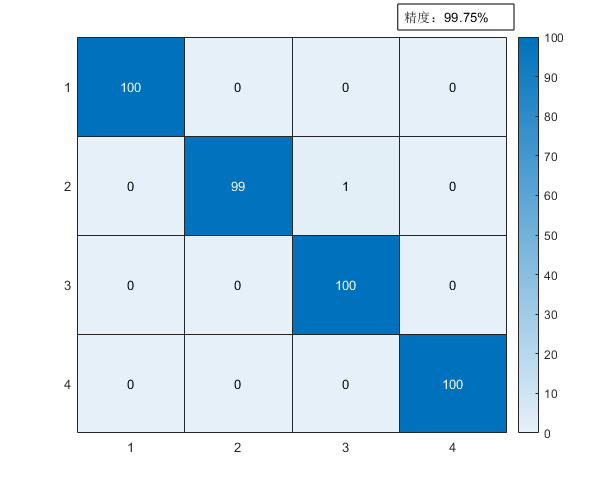
优点：

* 整体分类效果良好，分类性能较优；
* 图像中的水域和陆地区域被有效地识别并保留，轮廓和细节清晰；

缺点：

* 城镇（红色）分类精度较低，存在明显的噪声或伪影；
* 图像右下方的水域存在被误分类为陆地的情况；

1. 混淆矩阵及精度分析



* 精度值：99.75%，有一个类别3（城镇）的样本被误分类成类别2（陆地）
* 精度解释：根据混淆矩阵计算得到的总精度表明，模型正确分类了大约 99.75% 的样本。这个精度值反映了模型的整体性能良好。

##### 遇见问题与改进方法

* ROI区域训练集区域的样本提取不具备代表性：

解决方法：

通过ENVI Classic的Compute ROI Separability进行训练样本可分离性计算，可以查看训练样本分离性报告，ENVI为每一个感兴趣区组合计算Jeffries-Matusita距离和Transformed Divergence。其中根据可分离性值的大小，大于1.9说明样本之间可分离性好，属于合格样本；小于1.8，需要重选择样本；小于1，考虑将两类样本合成一类样本。