

浮游动物识别

戴嘉伦 武斌 朱亚菲 王如晨

2015 年 6 月

目录

1 ZooScan 扫描仪	2
1.1 操作步骤	2
1.1.1 创建项目	2
1.1.2 扫描背景	3
1.1.3 准备样品	4
1.1.4 扫描样品	5
1.1.5 回收样品	5
1.2 注意事项	6
2 ZooProcess	6
3 Plankton Identifier(PkID)	7
3.1 主窗口	7
3.2 Learning	8
3.3 Evaluation	11
3.4 Prediction	16
3.5 Validation	17
3.6 Compilation	20
4 评价方法	22
4.1 论文中采用的评价方法	22
4.2 论文中的评价结果	23
4.3 怎样得到混淆矩阵	23
4.4 混淆矩阵 Confusion matrix (CM)	24
4.5 交叉验证 Cross Validation (CV)	25

4.5.1 训练集和测试集	26
4.5.2 常见的交叉验证方法	27
5 识别结果分析	27
6 特征	28
6.1 位置特征	28
6.2 尺寸特征	28
6.3 灰度值特征	29
6.4 形状特征	29
6.5 自定义特征	30
7 优缺点分析	30

由法国国家科学研究院 Villefranche 海洋实验室的 Gorsky 等人发明 [?], Hydroptic 公司生产的 ZooScan 浮游动物图像扫描分析系统是一种实验室成像系统(即针对的是已经采集并固定保存的浮游生物样品), 主要用于对液体中的浮游动物样品进行计数、大小测量、种类鉴定以及生物量测定。ZooScan 系统采用非破坏性技术对液体浮游动物样本进行分析, 样品可重复使用。

ZooScan 系统是由 ZooScan、ZooProcess 和 Plankton Identifier(PkID) 等共同组成的, ZooScan 是硬件部分, 主要进行浮游动物样品扫描, 形成数字图像。ZooProcess 和 PkID 是软件部分, 分别以标准化的程序处理原始图像、对不同个体的形态参数进行自动测量和对图像中的浮游动物进行自动分类和计数。

1. ZooScan 扫描仪

ZooScan 扫描仪完成的任务:

- 扫描空白背景
- 扫描样品, 获得原始图片和元数据信息

1.1 操作步骤

1.1.1 创建项目

- 打开 “Image J”, 在软件界面中, 选择项目选项表最后的 “CREATE NEW PROJECT”, 创建一个新的项目。如图 1。

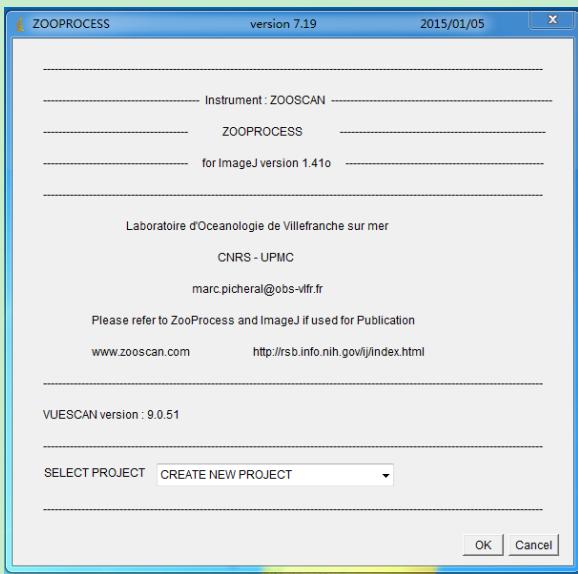


图 1: 创建项目

- 选择扫描选项。选择 2400dpi 分辨率和“Large”的扫描框 ($15cm \times 24cm$)。如图 2。

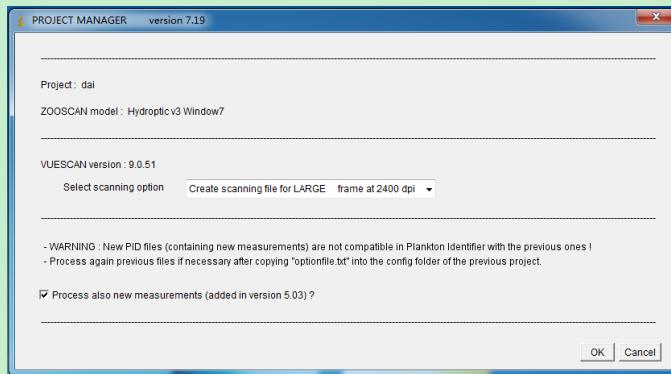


图 2: 选择分辨率和扫描框大小

- 输入样品的元数据 (metadata)。在后面处理过程中，可以通过 ZooProcess 中的“EDIT and MODIFY metadata”工具来修改元数据。如图 3。

1.1.2 扫描背景

背景图像是一张空白图像，用于图像分析过程中。在与样品相同环境下 (自来水或是过滤海水)，先扫描背景再扫描样品。最好是在每个扫描任务开始时，都扫描一次背景图像。

- 用清水清洁和冲洗 ZooScan 托盘和表面玻璃，时不时地检查并清除在玻璃和扫描框上的污点。
- 倒一些清水 (保持在室温) 没过托盘，它可以防止扫描框刮擦托盘。

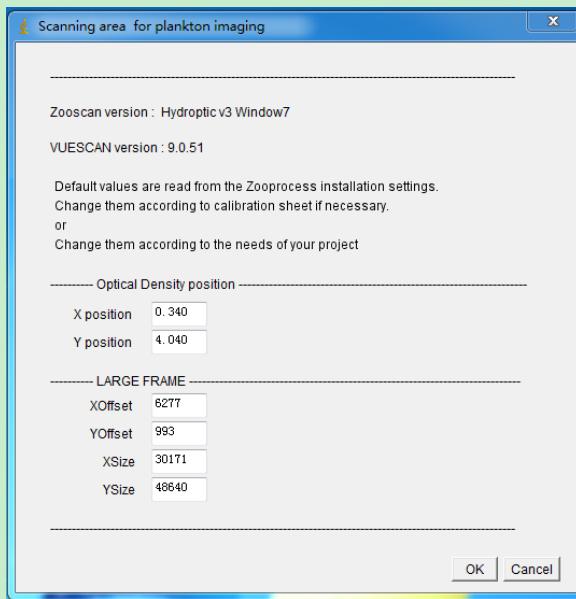


图 3: 输入扫描信息

- 放置扫描框 ($15\text{cm} \times 24\text{cm}$)，这取决于之前 ZooProcess 中的“创建项目”中的选项。
- 在预扫描和实际扫描之间，等待 30 秒。

1.1.3 准备样品

- 存储几升清水，保持在室温，用来为 ZooScan 注水。
- 用筛子（网格间隙为 $100\mu\text{m}$ ）过滤掉防腐剂和海水中物质。样品通过间隙为 1mm 和 $200\mu\text{m}$ 的两种网格，将浮游生物分成不同体型的两部分。
- 浮游生物被分为不同体型的两部分：一个为体型大的样品，另一个为体型小的样品。分别将分开后的样品，添加标签 $d1$ 和 $d2$ ，用于扫描后的数据处理过程中。如图 4。

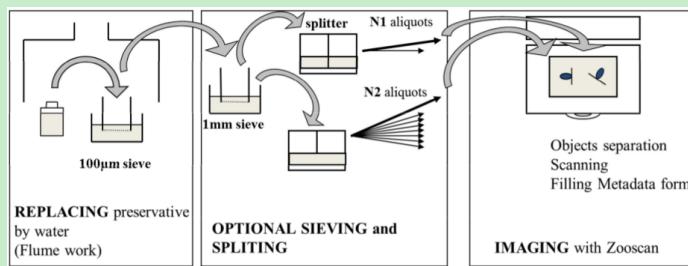


图 4: 准备样品

- 使上述两个分开后的样品中，保持只有 $1000 - 1500$ 个浮游生物。

1.1.4 扫描样品

- 在扫描托盘中加入量水，放置扫描框（ $15cm \times 24cm$ ），调整扫描框放置的位置（位置在扫描托盘中有标注）。
- 倒入样品，加入清水，直到没过扫描框的台阶。
- 将个体较大的浮游生物放在扫描框的中心，用木棒分离粘连浮游生物，避免浮游生物贴靠扫描框边缘。对于漂浮在水面的浮游生物，轻轻用木棒将浮游生物压入水中。如果存在无法没入水中的浮游动物，且数量不多，则将它们移出（[这一步是生成好的数据质量的关键步骤](#)）。如图 5。



图 5：用木棒分离粘连浮游生物

- 检查托盘是否有气泡，顶盖玻璃的表面是否冷凝。
- 加载 ZooProcess，选择项目，点击扫描样品，选择两部分样品中的一个样品（ $d1$ 和 $d2$ ），写入相关元数据。

1.1.5 回收样品

- 清洁托盘，避免下次实验污染了样品。
- 移走并冲洗透明的扫描框，回收所有的标本。
- 清洁干燥扫描托盘。

1.2 注意事项

- 清水可以是自来水或过滤海水，保持在室温是为了避免在 ZooScan 的托盘中产生气泡或者顶盖玻璃上出现冷凝。因为自来水管中的自来水和房间的温度存在一定温差。
- ZooScan 提供 1200dpi 和 2400dpi 两种分辨率扫描。分辨率限制在 2400dpi，是由于 ZooScan 设计的光路需要空气入水和由水进入玻璃两次穿过界面，使成像分辨率收到限制。
- ZooScan 的扫描框有两个尺寸： $11\text{cm} \times 24\text{cm}$ 和 $15\text{cm} \times 24\text{cm}$ 。推荐使用 $15\text{cm} \times 24\text{cm}$ 的扫描框，具体使用哪个扫描框由 ZooProcess 中的选项决定。
- 扫描空白背景不仅可以去除灯光产生的异质性的斑点等，而且可以检验系统的稳定性。
- 在准备样品阶段，通常情况下，将样品分为两个或以上的小样品进行扫描。
- 在准备样品阶段，如何将浮游生物样品分成不同的小样品，取决于原样品中浮游生物的种类多少与体型大小。
- 扫描的浮游生物必须保持不动的状态，使用固定剂或将样品麻醉。
- 扫描框上有 5mm 的小台阶，注入的水必须漫过这台阶的高度，避免在扫描后的图像边缘出现弯液面现象。如图 6。

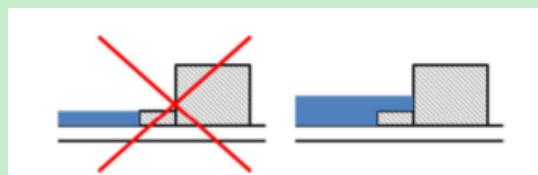


图 6: 框的台阶高度

- 在较大的框内 ($15\text{cm} \times 24\text{cm}$)，最多可以容纳 1000 – 1500 个浮游生物。
- 在 ZooProcess 中也可以对扫描图像中的粘连浮游生物进行分离，但是最好是在样品扫描之前用木棒进行分离。如图 7。

2. ZooProcess

ZooProcess 介绍详见<http://wbtwd2004.github.io/zooscan/2015/06/13/Zooprocess-Study.html>。

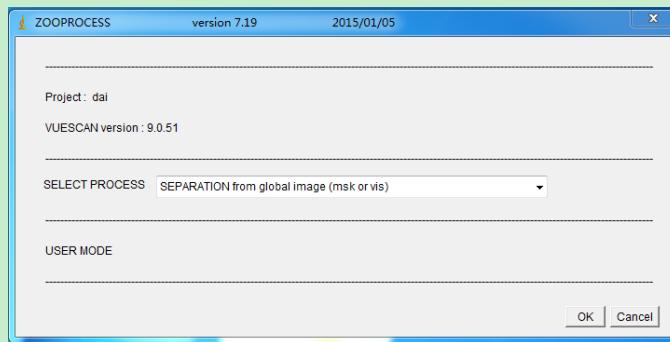


图 7: ZooProcess 的浮游生物分离

3. Plankton Identifier(PkID)

3.1 主窗口

打开 PkID 应用程序, 显示的界面上共有 5 个按钮: “Learning”、“Evaluation”、“Prediction”、“Validation” 和 “Compilation” (图 8)。这 5 个部分可以相互独立地运行, 但是会有运行的先后顺序。比如 “Evaluation” 和 “Prediction” 会用到 “Learning” 生成的文件, “Validation” 用到 “Prediction” 生成的文件, “Compilation” 用到 “Validation” 所生成的文件。因此, 当你第一次使用 PkID 时你应该首先运行 “Learning”。

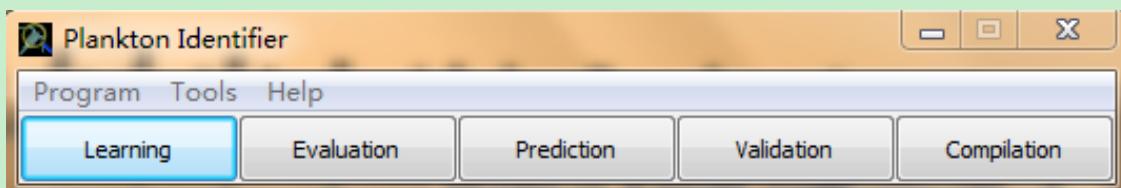


图 8: 主窗口

其他菜单:

Program > Settings 定义 Tanagra.exe 所在的路径以及存放缩略图、PID 文件和结果的默认文件夹路径。

Tanagra Path: 如果你安装了两个以上的 Tanagra 版本, 你可以选择想要使用的版本。点击 **Browse**, 浏览硬盘文件夹, 找到 Tanagra.exe, 点击 **OK**。

注: 如果你的 Tanagra 没有安装在 \Program Files\Tanagra 路径下, 或者你就根本没有安装 Tanagra, 这时当你运行 PkID 时就会自动弹出这个窗口 (图 9)。

Default folder: 默认文件夹是你第一次使用 PkID 执行一些步骤产生的文件所存放的地方。

Program > Exit 关闭 PkID。

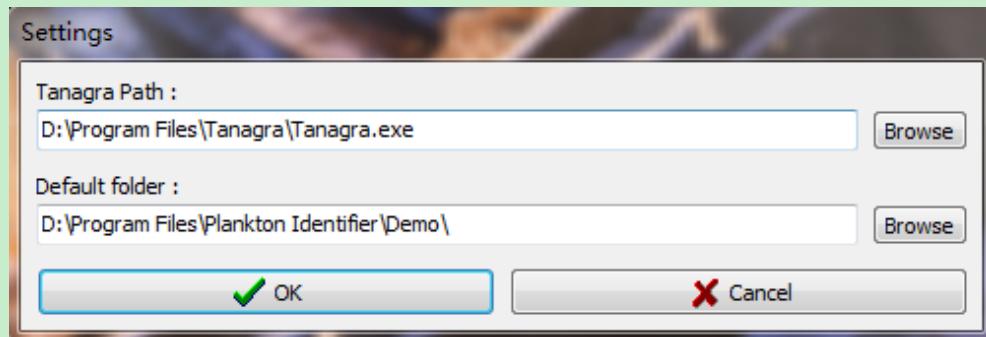


图 9: 设置窗口

3.2 Learning

这一步会生成一个学习文件以用作后续的自动识别。它对应于经专家鉴定的具有代表性的一些物体的子样本并且可以作为将来分析的参考。

1、文件夹选择窗口

当点击 Learning 按钮时，会出现如图 10 所示的文件夹选择窗口。

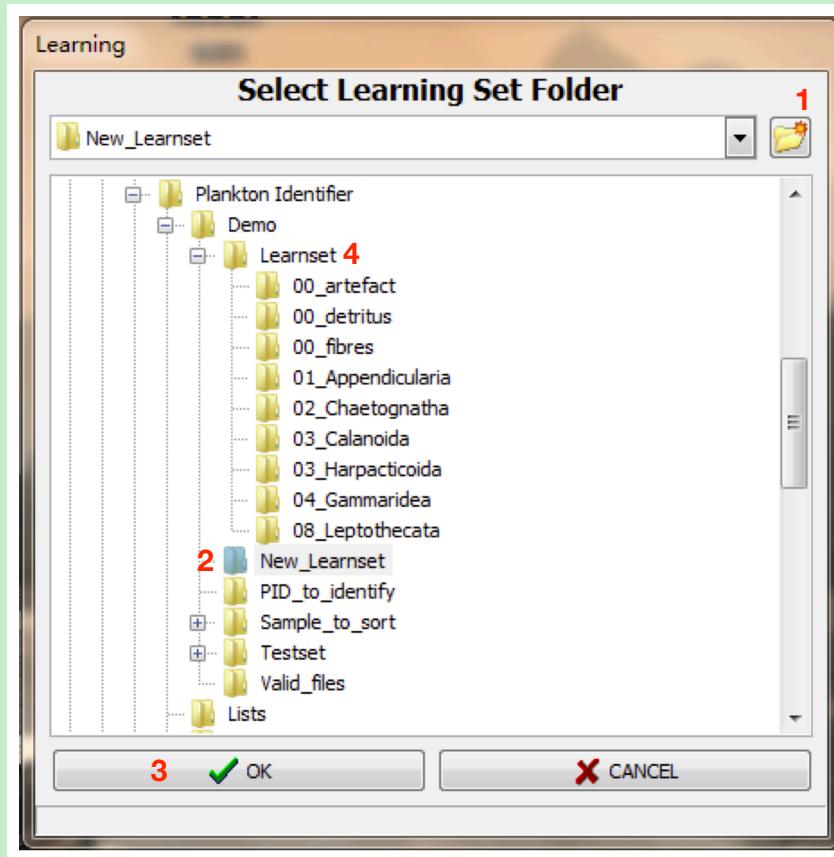


图 10: 训练集文件夹选择窗口

选择一个空的文件夹以用来创建新的训练数据集，你也可以通过点击右上角的按钮（图 10: 1）来

创建一个新的文件夹然后给它命名（图 10: 2），最后点击 OK 按钮（图 10: 3）。

你也可以选择一个已有的训练数据集（其中包含已分类好的子文件夹以及包含物体元数据的一些 PID 格式的文件，每个子文件夹代表一个种类，其中存放属于该类的 jpg 格式的缩略图），如图 10: 4。

注：如果此处选择的文件夹结构不符，或者包含了无效的数据，那么将会无法打开，并且在窗口的最下方会出现一行红色的警告信息以说明原因。

2、学习窗口

当选择了一个可以用来对缩略图进行分类的有效文件夹之后，点击 OK，会出现一个新的窗口（图 11）。其中左边部分（“Sample Set”）是要通过浏览硬盘文件夹来选择未被分类的样本。右边部分（“Learning Set”）是要将左边未被分类的样本拖到右边以完成分类（创建子文件夹）。

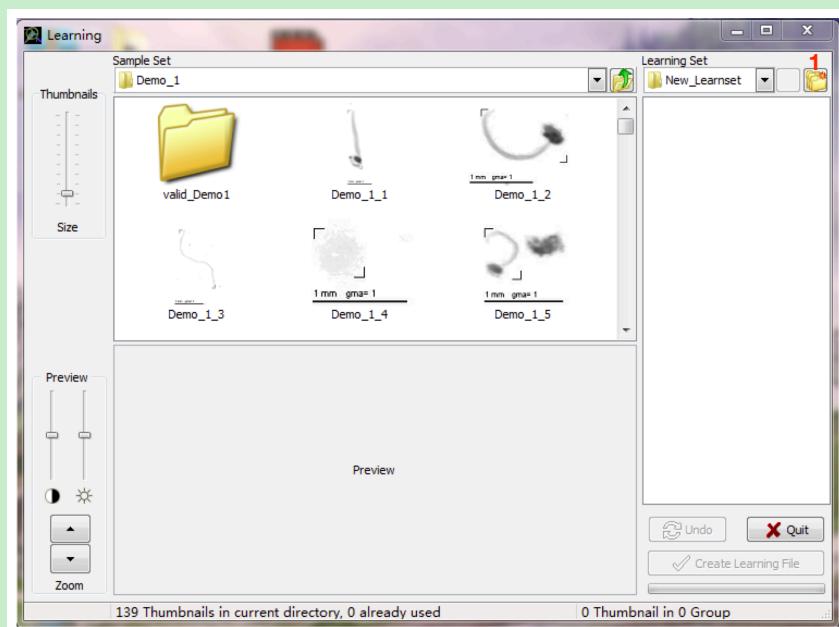


图 11: Learning 窗口

Sample Set

在窗口左边栏“Sample Set”处，浏览硬盘文件夹以打开一个包含未分类的缩略图和相应 PID 文件的文件夹。只有有有效名字（<Sample Name>_<Item Number>.jpg）的缩略图才会在左边栏中显示出来。如果名字是有效的但是 PID 文件中并没有包含该幅图像的一些数据，这时就会这个缩略图上方会出现一个问号并且当把鼠标放在问号上时会有相应的原因为解释，这幅图也是无法使用的。

注：如果此处选择的文件夹结构不符，或者包含了无效的数据，那么将会无法打开，并且在窗口的最下方会出现一行红色的警告信息以说明原因。

类别（子文件夹）创建

在窗口右边栏“Learning Set”处，为了把缩略图归到对应的类别，需要在步骤 1 所选择的文件夹

中创建一些子文件夹，每个文件夹代表一类。点击右上角的按钮（图 11: 1）创建新的文件夹，这时会出现一个新的窗口，可以给新创建的文件夹从给定的种类名字中选择一个（图 12）。如果给定的这些名字中没有合适的，你还可以选择另一个“Predefined Lists”或者先选择“New”作为名字，之后再在“Learning Set”这一栏中重新编辑命名。

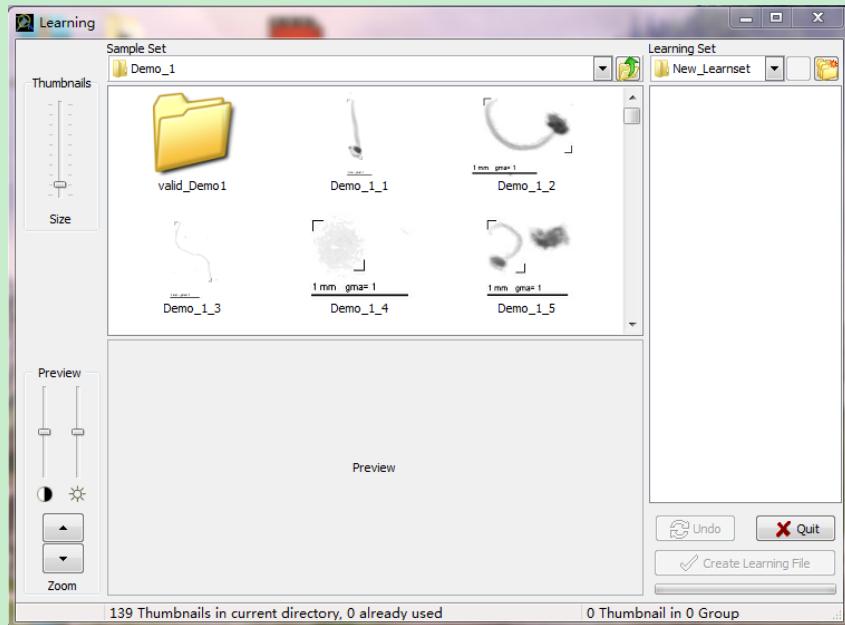


图 12: 创建训练集中分类类别的窗口

注 1: 所创建的子文件夹名字不能重复。已经用过的名字在“*Create Group Folders*”窗口中将不再显示。

注 2: 可以通过 *Tools>Import Name List* 菜单来自定义你自己的种类名字清单。

对缩略图进行归类

选中一个缩略图，下方会显示它的预览，如图 13: 1。你可以通过左侧的 zoom 按钮（图 13: 2）将预览结果放大，还可以通过调节左侧对比度和亮度的状态条来显示更多细节。

把缩略图拖到相应的子文件夹中（图 13: 3），这时缩略图会被复制到那个子文件夹中，而不是被剪切过去了。对应的 PID 文件也被复制到了右边文件夹中，但在这个窗口中是看不见的，以免视觉上的混淆。一旦一个缩略图已经用在了所创建的数据集中，上面就会出现一个红色的叉（图 13: 4），表示它不能再被使用了。

一次选中多个缩略图也是可以的，可以通过按住 Ctrl 键来完成。在选中多个缩略图的时候，如果你想看每一个缩略图的预览效果，可以从右下角往左上角选择。

每一个子文件夹中缩略图的数目会显示在这个子文件夹上，并且随着你的操作而更新。已经被归类了的缩略图数目以及还没有完成归类的缩略图数目都被显示在了窗口的最下方（图 13: 5）。

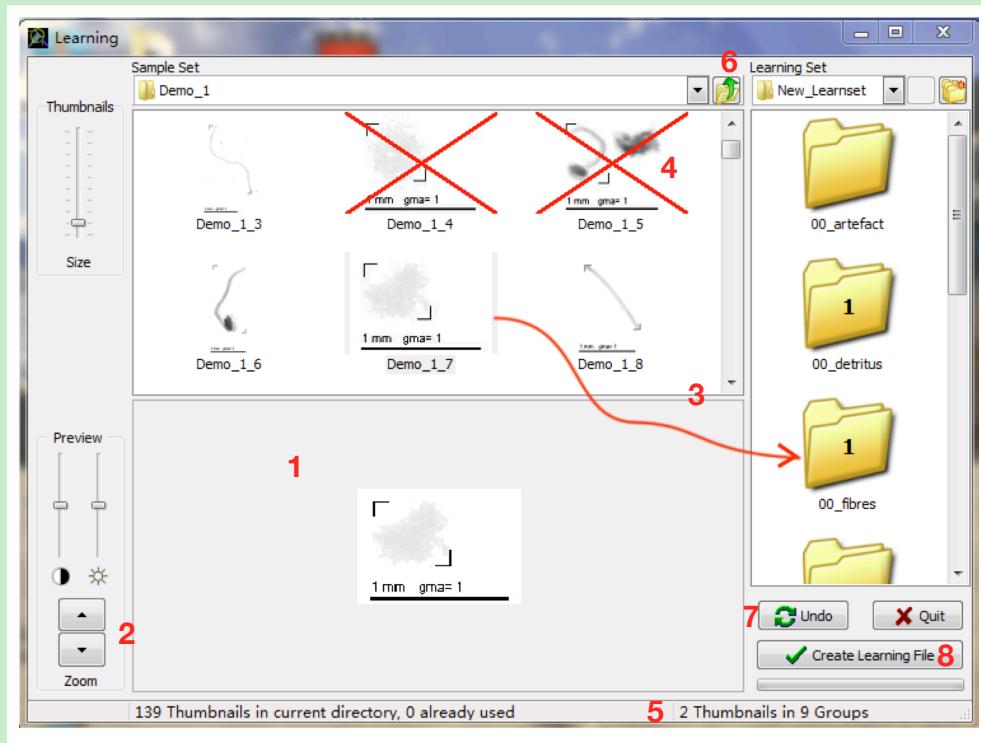


图 13: 对样本文件中的缩略图进行分类

你可以对多个样本集中的缩略图进行归类以创建自己的训练数据集。点击上面的一个按钮（图 13: 6）转到你想要操作的文件夹路径下。

Cancel action

当你在训练数据集的创建过程中执行了一些误操作之后，可以有以下两个办法取消：(1) 用 Undo 按钮（图 13: 7）(2) 打开右侧的子文件夹，选中缩略图然后用 DEL 键删除。其中 Undo 键可以用来取消删除操作或子文件夹删除操作，DEL 键可以用来删除一个所有缩略图都被清除了的空子文件夹。

创建学习文件

一旦你觉得你创建的每一个类别中已经归类了足够多的样本，你就可以点击“Create Learning File”按钮（图 13: 8）了，点击之后会出现一个保存对话框，上面显示了所要保存的目标文件夹路径以及学习文件的名字，默认的名字为 Learn_<number> 格式。点击“Save”按钮，所有的学习工作就完成了。这时又会出来一个对话框询问你是否要继续分类，如果你选择“No”，这个学习窗口就会被关闭回到主窗口。

3.3 Evaluation

这一步是要帮助你评估一下基于上一步中创建的训练数据集所建立的预测模型对训练数据集中的物体的识别率有多高。最后会生成一个包含识别结果的文本文件以及一个包含了数据分析信息的

html 报告。点击主窗口中的 Evaluation 按钮，会出现如下的界面（图 14）。

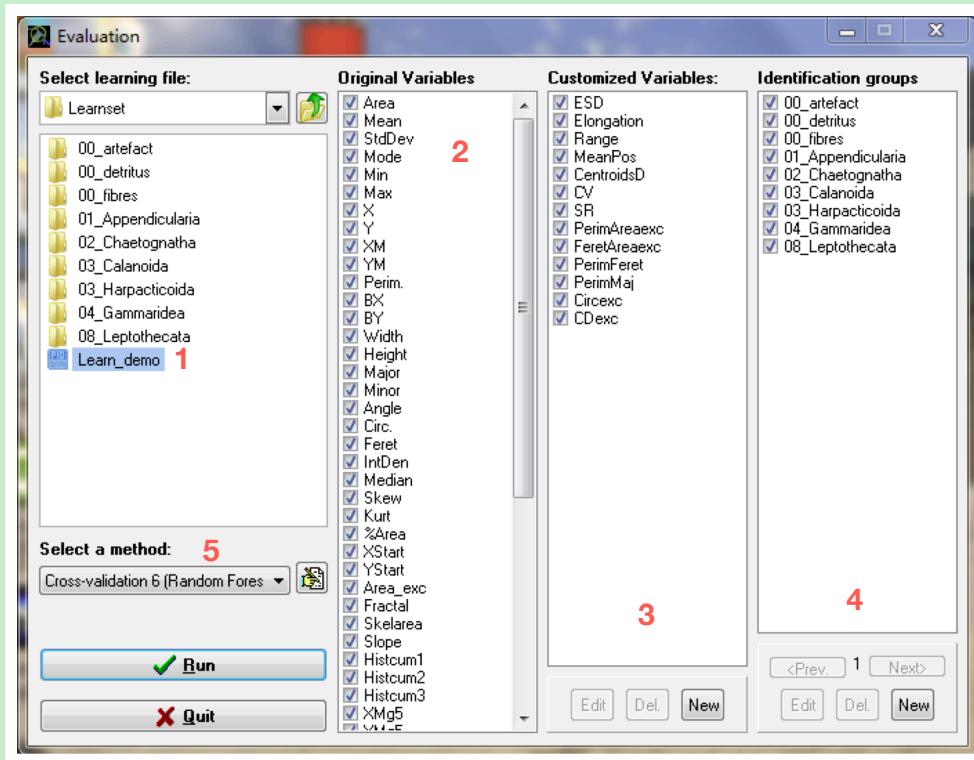


图 14: Evaluation 窗口

1、选取学习文件（图 14: 1）

浏览硬盘文件夹，选择你想要用来进行数据分析的学习文件。

注 1：选择学习文件后才能激活其它部分。

注 2：双击 PID 文件，将会自动在 PID viewer（如果已安装）或者文本编辑器（例如 Windows 下的 Notepad）中打开，从而可以校正文件内容。

2、初始变量（图 14: 2）

这里展示的是所选择的学习文件中的一些变量，你可以任意选取一些变量以用作分析，没有被选取的初始变量在计算时会被忽略，但不会从结果文件中移除。

注：对于那些用 ZooProcess 软件生成的 PID 文件，在计算时要被忽略的初始变量可以在附件中找到。

3、自定义变量（图 14: 3）

这一步是要根据已有的初始变量创建自定义变量。在你安装 PkID 之后会有 13 个自定义变量可供你选择是否要用它。没有被选取的初始变量在计算时会被忽略，并且不会在结果文件中显示。如果定义的某个变量不能从已经选取的初始变量计算，那么它会自动变成不可选取状态，并且显示成灰色。

要编辑一个已经存在的自定义变量，选中它，点击 Edit 以打开一个新窗口（15）。

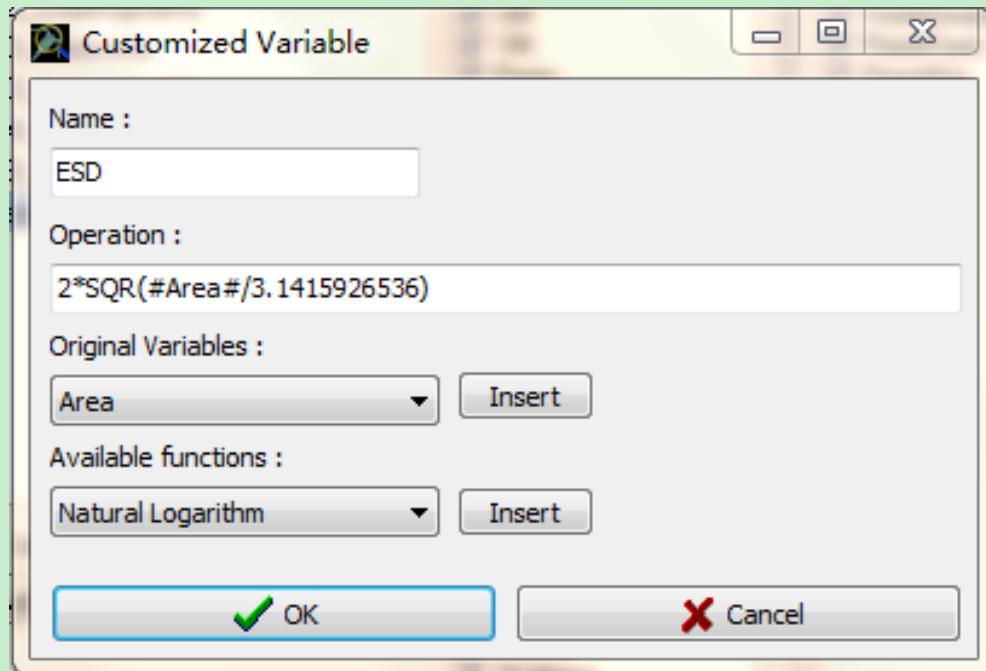


图 15: 自定义变量窗口

选中已存在的自定义变量，点击 Del 可以将其删除。

要新建一个自定义变量，点击 New，出现如图 15 的变量自定义窗口。

1. 对新建的变量命名（命名必须与已有变量不同）
2. 在“Operation”下面一栏中输入计算公式
3. 按“OK”键

注：在编写公式时，一些基本的运算符（例如 +, -, \, *, ^）、括号、数字都是跟平常在键盘上敲的一样。要插入一个初始变量的话，可以在“Original Variables”下面一栏选中它，然后点击“Insert”。在公式中用到初始变量时，需要在这个初始变量前后加上“#”号。在“Available functions”一栏中，可以选择一个函数进行插入。建议可以先看看已有的一些自定义变量是如何定义的，然后再编辑自己需要的公式。

4、Identification Groups (图 14: 4)

这里显示了在选取的学习文件中所定义的分类种类。默认的种类是不能被删除或编辑的，但是你可以通过将已有的类别合并来创建新的类别。

点击 New 可以创建新的类别，打开如图 16 所示的类别编辑窗口。

修改初始类别名字 (图 16: 1):

1. 在“Modified Group Names”一栏下方选择一个需要修改的名字 (图 16: 2)

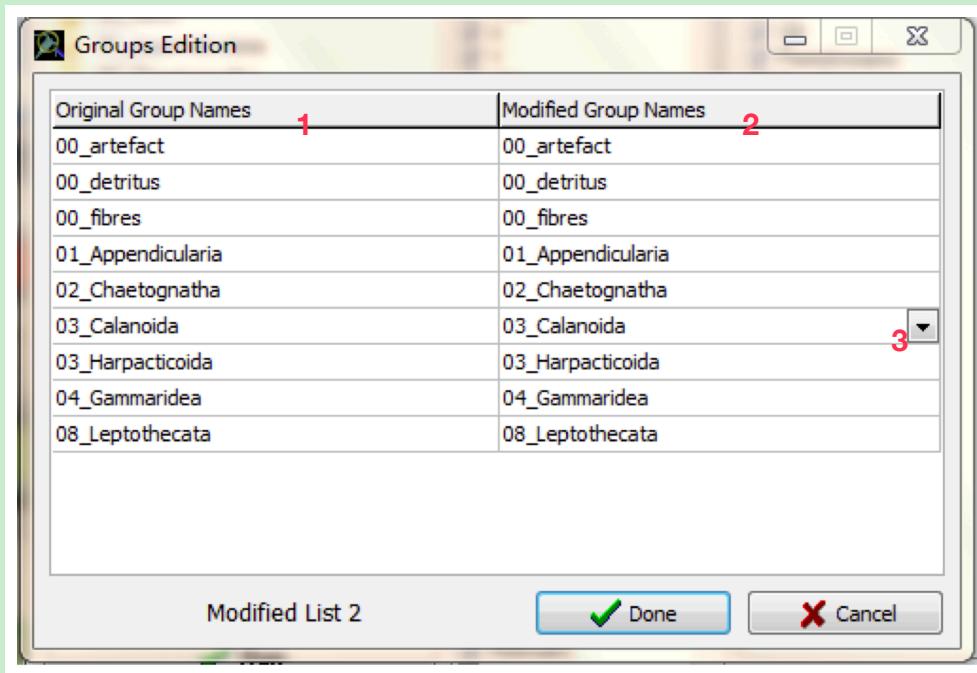


图 16: 分类类别编辑窗口

2. 给这个类别重新编辑命名，或者点击如图（图 16: 3）所示按钮，在名字列表中选择一个
3. 完成对类别的重新命名后，点击 Done

注：你可以创建很多的分类种类的列表，后续数据分析中用到的只会是当前这个列表（在 *Evaluation* 窗口点击 *Run* 时所看见的那个列表）。你可以用 *Prev.* 和 *Next>* 按钮来选择你想要的进行分析那个列表。在生成的结果文件中，初始名字会显示在“Ident1”这一列中，新的名字显示在“Ident2”这一列中。

5、选择一种方法（图 14: 5）

这里是要选择一种评价方法，用来检测用有监督学习方法和所选择的学习文件学习的模型的识别能力。PkID 中一共提供了两种评价方法。

k-fold cross-validation method: 用重采样技术来评价学习算法在学习文件上的识别准确率。初始的训练数据集被随机分为 k 个相同大小的子集。每一次循环过程中，将 $k - 1$ 个子集放在一起形成训练集，并且构造出预测模型，剩下的那 1 个子集被当成测试集来评价模型预测能力。一共循环 k 次。这样下来，每个子集都会有 1 次机会作为测试集， $k - 1$ 次机会作为训练集。将 k 次预测结果取平均，可以得到这个模型的最终预测能力。交叉验证过程会重复 n 次， n 次交叉验证的平均错误率在一个混淆矩阵中被计算得到。

PkID 一共实现了 8 种交叉验证方法，每一个都采用了不同的学习算法。

所有的交叉验证方法中所用的参数都是 $k = 2, n = 5$ 。要想改变这两个值：

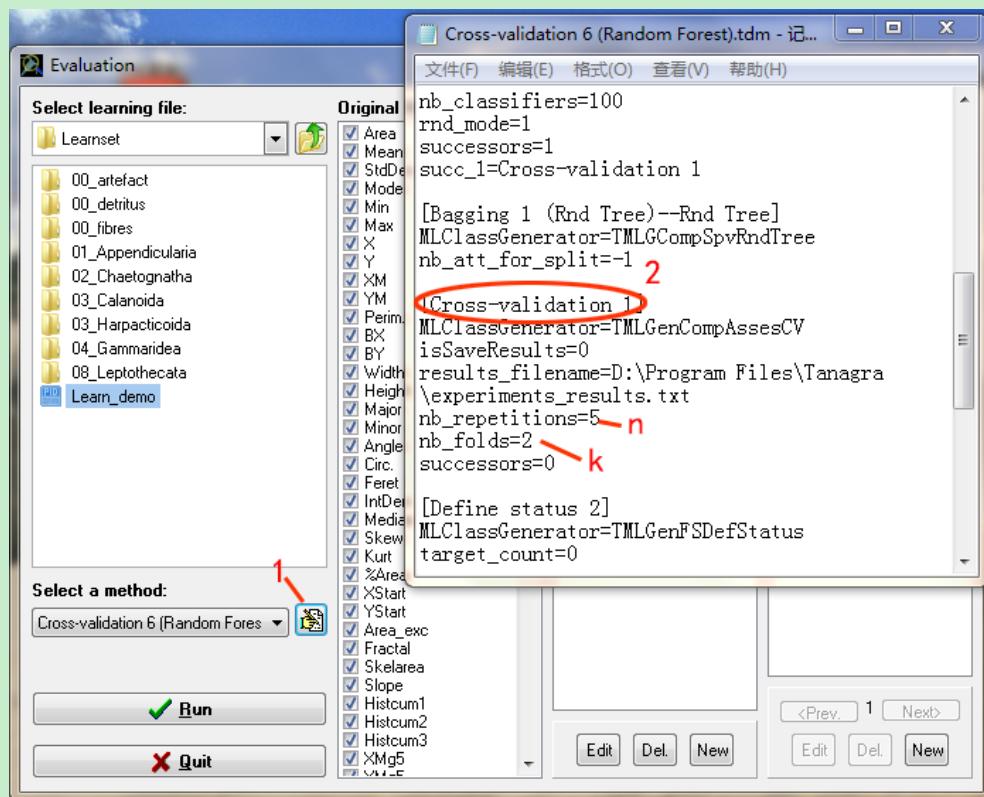


图 17: 分类类别编辑窗口

1. 点击编辑按钮 (图 17: 1), 打开相应的 tdm 文件
2. 在交叉验证部分修改 k 和 n 值图 17: 2
3. 在关闭 tdm 文件前保存修改

Test method: 在一个预先定义好的、独立的测试文件上对模型的准确率进行评价。在使用这个测试方法之前，需要创建一个特殊的文件，可以通过主窗口菜单中的 **Concatenate Learning Files** 来创建。相比于用两个不同的文件来作为训练集和测试集，我们更倾向于用一个文件来把它们联系起来，并且通过状态来表示所要扮演的角色，状态有“学习”和“测试”两种。

PkID 提供了两种测试方法。Test 1 可以通过在预先定义的测试文件上比较 8 个算法的准确率来选择最好的学习方法。Test 2 只用到随机森林算法。

注：通常来讲，训练集越大，所构造的分类模型越好，测试集越大，预测准确率越高。

Export to text file (no analysis): 生成一个文本文件，包含了所选的学习文件中所有初始变量和自定义变量，而不包含预测结果。这个文件还可以输入到任意的数据挖掘软件站进行分析。

6、开始分析

当所有文件、变量、类别都选择好之后，点击 **Run** 按钮，将会出现一个保存页面，选择要保存的文件夹、文件名，点击 **Save**。

分析完成后（可能需要几分钟，这取决于样本大小和所选方法），结果和 html 报告被保存在了所选择的文件夹中，并且 html 报告会自动在网页中打开。这时会有一个对话框询问你是否要退出 Evaluation 窗口，如果你选择“Yes”，Evaluation 窗口就会被关闭并返回主窗口。

与初始 PID 文件相比，评价文件包含了以下新的列：

1. 对应于自定义变量的列
2. 包含了学习文件中分类类别的一列 (Ident)
3. 包含了修改后的类别名字的一列 (Ident2)
4. 用来表示状态的一列 (Learning 或 Test)

3.4 Prediction

这一步是要根据选择的学习文件中的种类对样本进行自动识别。结束后也会生成一个包含了自动识别结果的文本文件和一个包含了数据分析信息 html 报告。点击主窗口中的 Prediction 按钮，会打开一个新的窗口（图 18）：

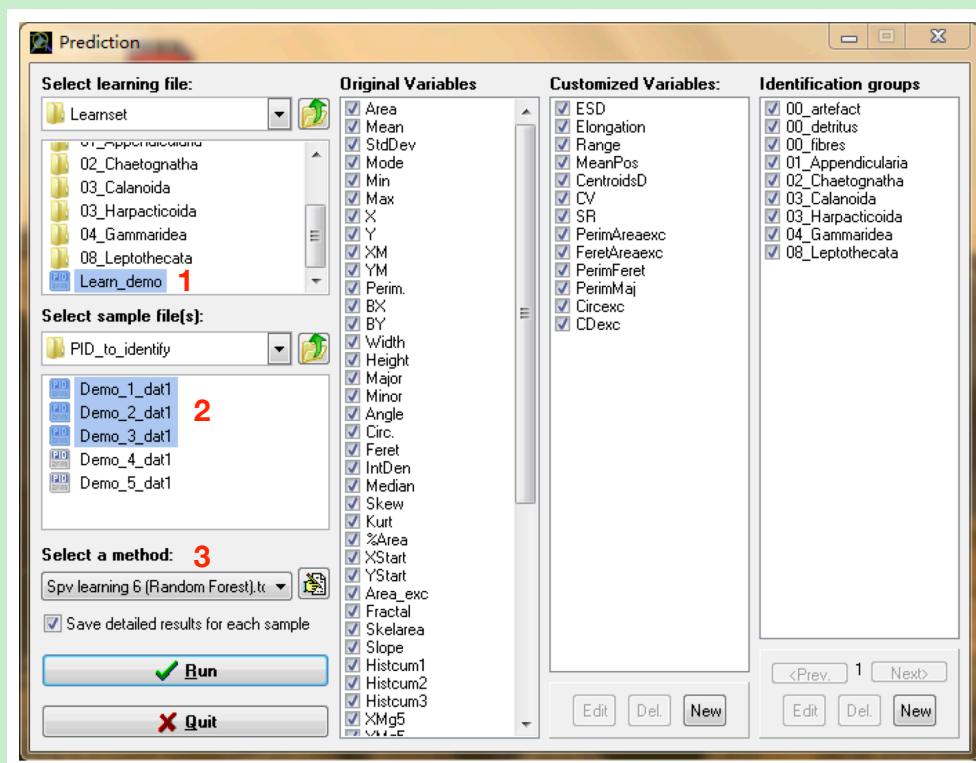


图 18: Prediction 窗口

- 1、选择学习文件（图 18: 1）

2、选择样本文件（图 18: 2）

3、选择学习算法（图 18: 3）

从 PkID 提供的 8 个学习算法中选择一个。

- Spv learning 1 (5-NN)
- Spv learning 2 (C-SVC linear)
- Spv learning 3 (C-SVC RBF)
- Spv learning 4 (BVM)
- Spv learning 5 (C4.5)
- Spv learning 6 (Random Forest)
- Spv learning 7 (PLS)
- Spv learning 8 (Multilayer Perceptron)

4、开始分析

当所有文件、变量和分类类别都选好之后，点击 Run 按钮，会出现一个保存对话框。

3.5 Validation

这一步是将上一步生成的预测文件可视化，并且人为地对识别结果进行校正。这里可以有两个选择（图 19）：1) 用 Prediction 那一步生成的 Pred_.txt 文件来将预测结果可视化，真正实现缩略图的自动分类，你还可以对自动识别的结果进行进一步检查和校正。2) 打开一个已有的校正集来继续一个校正或进行二次校正。

Visualize a prediction from a Pred_.txt file

1. 选择一个要用来校正的 Pred_.txt 文件（图 20: 1）

2. 选择一个包含未分类缩略图的文件夹作为“Sample Set”（图 20: 2）

3. 选择用来存放分类好了的缩略图的目标文件夹（图 20: 3）

4. 点击“Visual Validation”（图 20: 4）

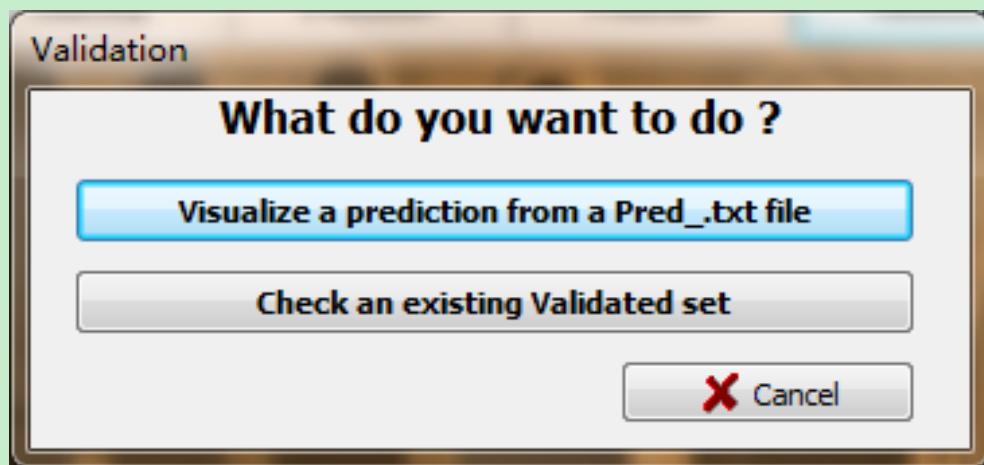


图 19: Validation 窗口：What do you want to do?

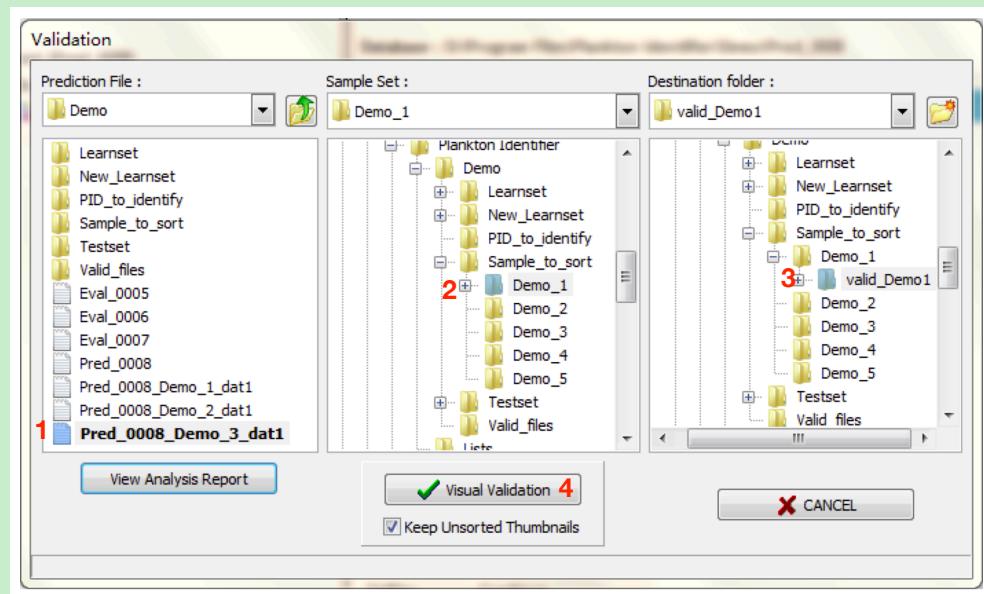


图 20: 校正选择窗口 1

注：在分类过程中，缩略图是被复制而不是剪切到了目标文件夹中。如果你不想保留未分类的缩略图，可以在图 ?? 处取消勾选 *Keep Unsorted Thumbnails*。

Check an existing Validated set

1. 选择你想要检查的 Valid_.txt 文件（图 21: 1）
2. 选择一个包含已分类好的校正后的缩略图的文件夹（图 21: 2）
3. 点击“Visual Validation”按钮（图 21: 3）

Visual Validation

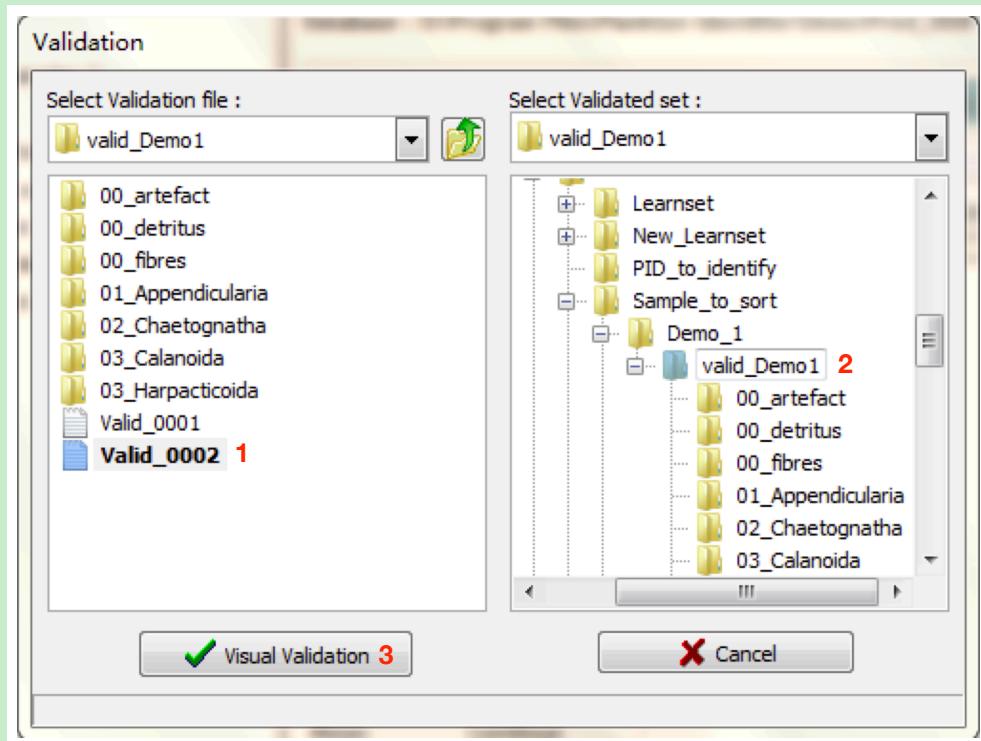


图 21: 校正选择窗口 2

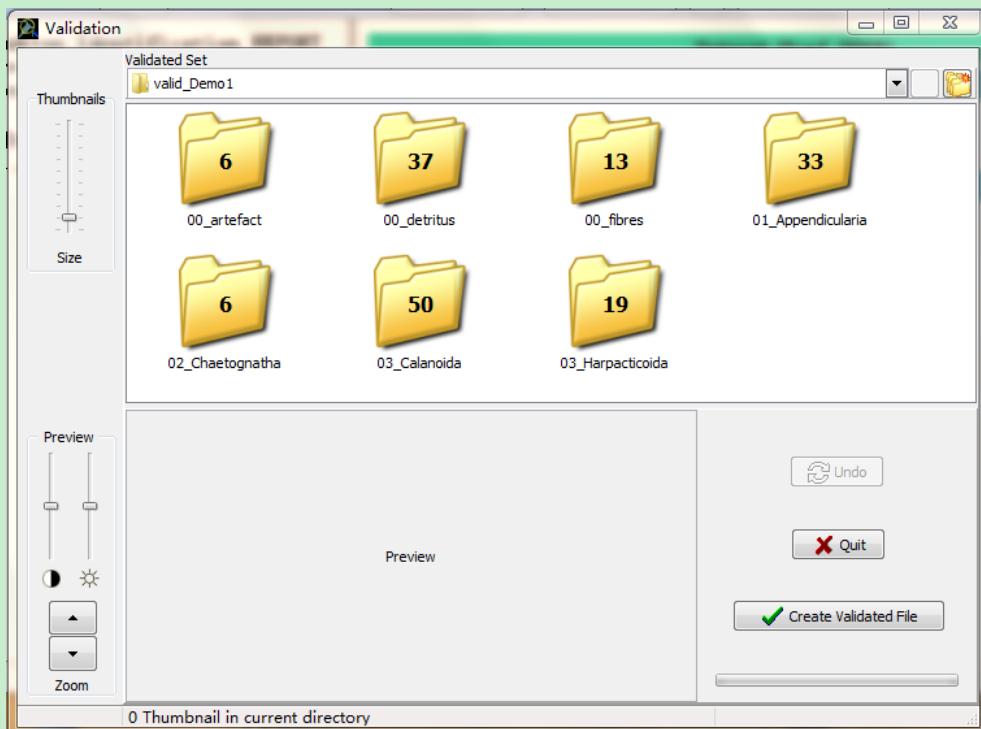


图 22: 校正窗口

前面的两个选择最终都会打开如图 22 所示的校正窗口：

利用预测模型已经将每幅缩略图放到了其对应的类别文件夹中，你可以打开每个文件夹查看是否被正确分类了。

Thumbnails moving

3.6 Compilation

这一步是用来将上一步的生成的多个校正文件连起来，并且计算每个类别中的物体数目。在主窗口中点击 Compilation，会出现如下窗口（图 23）：

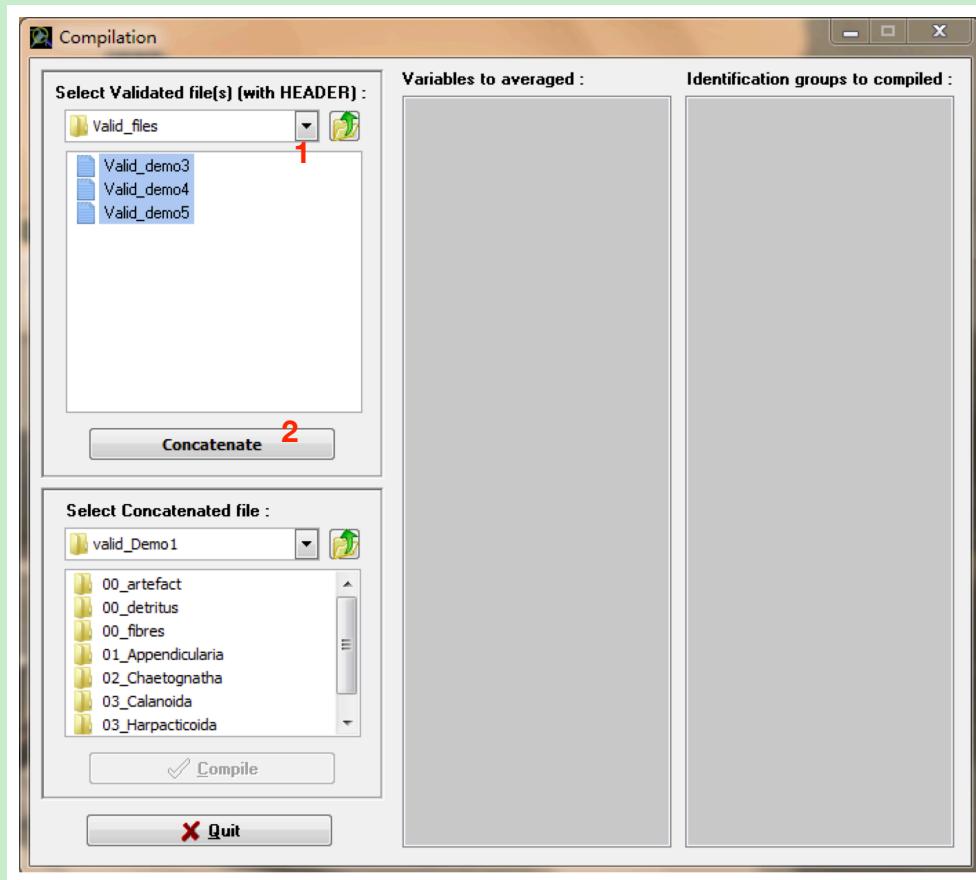


图 23: Compilation 窗口：连接

1、Create a concatenation file

1. 浏览硬盘文件夹，找到包含你所要连接的 Valid_.txt 文件的文件夹（图 23: 1）
2. 点击 Concatenate（图 23: 2），会出现一个保存会话框
3. 给连接后的文件进行命名，选择将其保存的文件夹路径
4. 点击 Save 按钮，连接就开始了（这可能需要几分钟，取决于你想要连接的文件数目）

注：要一次连接很多 *Valid_.txt* 文件，需要把它们放在同一个文件夹中，然后按“*Ctrl*”键同时选中它们。

Create a compilation file

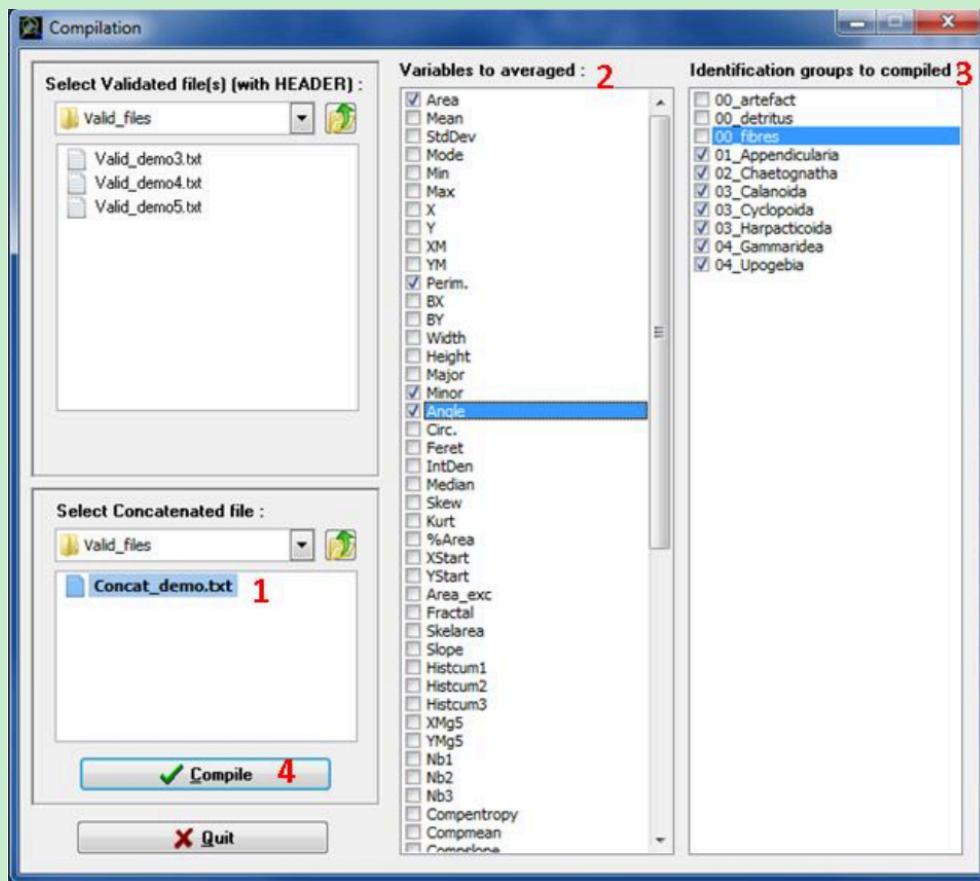


图 24: Compilation 窗口：编译

1. 选择一个已经连接好的 *Concat_.txt* 文件。最终的连接文件会以粗体显示（图 24: 1）。
 2. “Variables to averaged”一栏中可以勾选也可以不勾选相应的变量（图 24: 2），被打上勾的变量在编译文件中会被平均，没有打勾的会被去除。
 3. “Identification groups to compiled”一栏也可以挑选相应的类别打上勾（图 24: 3），只有打上勾的类别才会计算出它所包含的物体数目。
 4. 点击 **Compile** 按钮（图 24: 4），出现一个保存对话框
 5. 给编译后的文件进行命名，选择将其保存的文件夹路径
 6. 点击 **Save**，编译就开始了
- 编译后生成的文件如图 25。

COMPILED FILE											
Label	GroupName	FracId	SubPart	Count	Area	Perim.	Minor	Angle	ESD	Elongation	
Demo_3	00_detritus	F1		128	55	1936.4	240.064	31.888	90.373	44.96	2.349
Demo_3	03_Harpacticoida	F1		128	11	1284.364	177.813	27.623	88.205	40.362	2.155
Demo_3	03_Calanoida	F1		128	45	2233.889	315.354	37.825	82.176	52.158	1.992
Demo_3	02_Chaetognatha	F1		128	6	5797.167	747.159	31.192	83.094	84.364	7.822
Demo_3	00_fibres	F1		128	10	1789.4	420.543	15.493	65.26	45.885	9.942
Demo_3	03_Cyclopoida	F1		128	1	1544	243.078	28.343	42.376	44.338	2.447
Demo_3	00_artefact	F1		128	6	1263.5	270.969	35.605	74.971	39.386	1.239
Demo_3	01_Appendicularia	F1		128	30	3211.9	468.046	30.148	83.982	62.01	4.562
Demo_4	00_artefact	F1		128	5	1873.6	319.886	38.802	50.847	47.521	1.667
Demo_4	03_Calanoida	F1		128	65	3063.2	348.628	42.982	81.698	61.048	2.052
Demo_4	04_Gammaridea	F1		128	4	7820	583.645	80.585	62.359	97.558	1.47
Demo_4	00_detritus	F1		128	43	2168.86	237.041	37.311	86.007	48.898	2.085
Demo_4	01_Appendicularia	F1		128	26	5943.577	660.8	42.251	76.959	83.938	4.445
Demo_4	03_Harpacticoida	F1		128	4	1370.75	186.223	28.762	102.387	41.719	2.179
Demo_4	04_Upogebia	F1		128	2	12262	906.794	57.27	41.864	124.772	4.771
Demo_4	03_Cyclopoida	F1		128	7	2078.429	252.343	30.891	100.26	50.725	2.69
Demo_4	00_fibres	F1		128	3	1744.667	518.907	11.963	85.451	47.033	18.579
Demo_5	00_detritus	F1		256	48	1575.125	202.747	31.077	85.253	42.98	2.194
Demo_5	00_fibres	F1		256	15	1171.333	383.601	14.173	99.428	37.625	9.063
Demo_5	00_artefact	F1		256	11	1304.182	353.829	29.674	62.284	39.575	2.066
Demo_5	01_Appendicularia	F1		256	13	3227.846	471.515	32.683	73.519	63.513	4.298
Demo_5	03_Harpacticoida	F1		256	13	1535.308	215.014	30.055	102.575	44.162	2.188
Demo_5	03_Calanoida	F1		256	41	2562.976	351.819	39.317	90.851	55.844	2.089
Demo_5	03_Cyclopoida	F1		256	2	2413	270.877	35.866	44.807	54.734	2.334

图 25: 编译文件

4. 评价方法

4.1 论文中采用的评价方法

文中采用混淆矩阵 (CM) (混淆矩阵介绍见 4.4) 对分类器的分类效果进行评价。评价时计算的是 CM 的召回率 (recall 即 the rate of true positives) 和虚警率 (low contamination 即 the rate of false positives)。

论文中的具体介绍：

Evaluation of classifier performance requires the examination of a [Confusion matrix \(CM\)](#), which is a contingency table crossing true (manually validated) and predicted (assigned by the classifier) identification of objects. Correct interpretation of the CM requires the examination of each category separately, including [the rate of true positives](#) as well as [false positives](#).

- the rate of true positives (recall)

$$\text{true positives} = \frac{\text{number of objects correctly predicted}}{\text{total number actual objects}}$$

- the rate of false positives (low contamination)

$$\text{false positives} = \frac{\text{number of objects falsely assigned to a category}}{\text{total number of predicted objects}}$$

4.2 论文中的评价结果

论文中的评价结果如图 26:

	Percentage in learning set	浮游生物的种类																				实际每种浮游 生物的数量	Recall	1-Possitives rate(FP)	
		Aggregates	Aggregates_dark	Appendicularia	Bad focus	Bubbles	Chaetognatha	Cladocera	Copepoda_other	Oithona	Copepoda_small	Decapoda_large	Egglike	Fibers	Medusae	Nectophores	Thaliacea	Limacina	Scratch	Pteropoda_other	Radiolaria				
Aggregates	7.2	683	27	80	67	0	0	77	99	53	134	8	16	26	9	37	1	1	12	0	25	1355	0.50	0.42	
Aggregates_dark	5.7	3	671	0	0	39	0	18	38	0	4	0	27	0	0	0	0	0	0	8	43	1064	0.63	0.40	
Appendicularia	7.7	85	0	1243	0	0	22	0	2	2	0	5	0	73	0	18	1	0	0	2	2	1455	0.85	0.20	
Bad focus	7.5	44	0	5	1230	0	0	5	27	2	12	0	1	5	0	47	0	0	20	0	12	1410	0.87	0.10	
Bubbles	4.5	2	30	0	0	0	786	0	2	0	0	0	0	7	0	0	0	0	10	0	0	3	840	0.94	0.06
Chaetognatha	2.3	0	0	51	0	0	0	356	0	0	0	0	0	1	0	20	0	0	0	0	0	428	0.83	0.11	
Cladocera	6.2	27	19	2	1	0	0	1028	22	0	2	0	5	0	0	0	1	0	1	0	0	62	1170	0.88	0.16
Copepoda_other	11.3	74	32	4	8	0	0	14	1608	119	237	24	1	0	0	4	0	0	0	6	2	2133	0.75	0.24	
Oithona	7.5	24	0	28	0	0	0	0	61	1256	38	0	0	9	0	0	0	0	0	4	0	1420	0.88	0.17	
Copepoda_small	7.8	130	3	0	6	0	0	3	139	43	1141	0	0	0	0	0	0	0	0	0	0	1465	0.78	0.28	
Decapoda_large	2.4	2	0	7	0	0	0	0	23	0	0	410	0	0	0	1	0	0	0	2	0	445	0.92	0.12	
Egg-like	1.9	12	26	3	7	4	0	12	3	0	0	259	0	1	0	0	8	0	0	15	0	350	0.74	0.20	
Fibers	6.4	10	0	84	0	0	21	0	7	14	0	0	0	1034	0	0	0	0	12	28	0	1210	0.85	0.13	
Medusae	0.7	16	0	0	2	0	0	0	2	0	0	0	2	0	100	18	0	0	0	0	0	140	0.71	0.19	
Nectophores	6.1	26	0	21	32	0	1	3	5	0	0	0	1	1	10	1029	20	0	1	0	5	1155	0.89	0.19	
Thaliacea	1.4	8	0	5	6	0	2	0	0	0	0	0	0	0	3	108	123	0	0	0	0	255	0.48	0.17	
Limacina	4.5	2	267	0	0	8	0	9	0	0	0	0	4	0	0	0	558	0	0	0	7	855	0.65	0.28	
Scratch	1.6	2	0	0	5	0	0	0	1	0	0	0	0	6	0	2	4	0	0	285	0	305	0.93	0.14	
Pteropoda_other	2.3	9	0	2	1	0	0	3	55	23	23	20	0	11	0	0	0	0	0	291	2	440	0.66	0.15	
Radiolaria	5.1	23	45	11	8	0	0	45	12	1	3	0	1	0	0	3	0	0	0	0	800	955	0.84	0.18	
Total	100	1182	1120	1546	1373	837	402	1219	2124	1513	1594	468	324	1185	123	1268	149	774	330	341	978	18850	0.78	0.19	

Corresponding correct identifications (in bold) are in the diagonal.

检测到的每种浮游生物的数量

分类正确的数量 true positives rate(TPR)

图 26: Confusion matrix for the 20 categories in the learning set

4.3 怎样得到混淆矩阵

论文中提到三种得到 CM 的方法 (即采用什么训练集和测试集来生成):

1. Re-substitution CM

这个方法是采用的测试集和训练集为同一个数据集。在这个过程中，用产生的分类器对测试集进分类时，得到的分类结果错误较少甚至可能没有错误，采用 CM 进行评价时就会低估分类器的错误率。

2. Cross-validation CM

这个方法是采用交叉验证 (交叉验证介绍见4.5) 的方法。在这个过程中，将一个数据集分成 n 个相等的子集，用其中 n-1 个子集来训练产生分类器，用剩下的 1 个子集来进行测试，重复进行 n 次来构建 CM。

3. Uses two equivalent and independent learning files describing the same categories with different objects

这种方法是用两个相等且相互独立的数据集分别作为训练集和测试集，根据测试结果建立 CM。两个相等的数据集即在两个集合中浮游生物的种类相同，但是每种浮游生物中的个体是不同的。

4.4 混淆矩阵 Confusion matrix (CM)

在机器学习中，混淆矩阵 (CM) 是一种比较简单的对学习算法性能进行评价的评估准则，而且是大多数指标的基础。常用的算法评估准则有：Confusion Matrix、ROC、Lift、Gini、K-S 等等。

混淆矩阵 (CM)¹²：混淆矩阵是一种评估分类器可信度的方法。在图像精度评价中，主要用于比较分类结果和实际测得值，可以把分类结果的精度显示在一个混淆矩阵里面。特别用于监督学习，在无监督学习一般叫做匹配矩阵。

混淆矩阵是一个 n 行 n 列的矩阵，n 代表类别的数量。矩阵的每一列代表预测的每一类的数量，每一行代表实际的每一类的数量。对角线上表示分类正确的每一类的数量。

例如：有 150 个样本数据，这些数据实际分为 3 类，每类 50 个。分类结束后得到的混淆矩阵如图27。例如：第一行说明类 1 的 50 个样本有 43 个样本分类正确，5 个错分为类 2，2 个错分为类 3；第一列说明类 1 有 43 个样本分类正确，类 2 的 2 个样本被错分为类 1，类 3 没有样本被错分为类 1；对角线上的数据表示，类 1、2、3 分别有 43、45、49 个样本被分类正确。

	类1	类2	类3
类1	43	5	2
类2	2	45	3
类3	0	1	49

图 27: 混淆矩阵例子

根据混淆矩阵可以导出以下几个参数³:

- true positives (TP): 正样本被识别出的数量
- true negatives (TN): 负样本被识别出的数量

¹<http://baike.baidu.com/view/2781781.htm>

²http://en.wikipedia.org/wiki/Confusion_matrix

³http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html

- false positives (FP): 负样本被错误分为正样本的数量
- false negatives (FN): 正样本被错误分为负样本的数量
- The true positive rate (TPR) : 召回率, 就是正样本被识别出的概率。(文中使用的参数)

$$TPR = \frac{TP}{TP + FN}$$

- The false positive rate (FPR): 虚警率, 负样本被错误分为正样本的概率。(文中使用的参数)

$$FPR = \frac{FP}{FP + TN}$$

- The true negative rate (TNR): 负样本被识别出的概率

$$TNR = \frac{TN}{FP + TN}$$

- The false negative rate (FNR) : 漏警率, 正样本被错误分为负样本的概率。

$$FNR = \frac{FN}{FN + TP}$$

- Positive predictive value (PPV)

$$PPV = \frac{TP}{TP + FP}$$

- Negative predictive value (NPV)

$$NPV = \frac{TN}{TN + FN}$$

- False discovery rate (FDR)

$$FDR = \frac{FP}{FP + TP}$$

4.5 交叉验证 Cross Validation (CV)

交叉验证是用来验证分类器的性能一种统计分析方法, 基本思想是把在某种意义下将原始数据进行分组, 一部分做为训练集 (training set), 另一部分做为验证集 (validation set), 首先用训练集对分

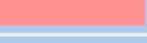
Attribute	Gini	Distribution			
		Values	Count	Percent	Histogram
Ident	0.8396	Appendicularia	545	5.76 %	
		Bubble	145	1.53 %	
		Chaetognatha	349	3.69 %	
		CladoceraPenilia	925	9.78 %	
		Copepoda	1474	15.58 %	
		Decapoda	557	5.89 %	
		Doliolida	285	3.01 %	
		Egg	344	3.64 %	
		Fiber	304	3.21 %	
		Gelatinous	602	6.36 %	
		Multiple	635	6.71 %	
		Nonbio	3078	32.54 %	
		Pteropoda	217	2.29 %	

图 28: 数据集中每种类别的数目及所占比例

类器进行训练，在利用验证集来测试训练得到的模型，以此来做为评价分类器的性能指标。

4.5.1 训练集和测试集

在模式识别与机器学习的相关研究中，经常会将数据集分为训练集跟测试集这两个子集，前者用以建立模型，后者则用来评估该模型对未知样本进行预测时的精确度，正规的说法是泛化能力。怎么将完整的数据集分为训练集跟测试集，必须遵守如下要点：

1. 只有训练集才可以用在模型的训练过程中，测试集则必须在模型完成之后才被用来评估模型优劣的依据。
2. 训练集中样本数量必须够多，一般至少大于总样本数的 50%。
3. 两组子集必须从完整集合中均匀取样。

注：其中最后一点特别重要，均匀取样的目的是希望减少训练集和测试集与完整集合之间的偏差，但却也不易做到。一般的作法是随机取样，当样本数量足够时，便可达到均匀取样的效果，然而随机

也正是此作法的盲点，也是经常是可以在数据上做手脚的地方。举例来说，当辨识率不理想时，便重新取样一组训练集和测试集，直到测试集的识别率满意为止，但严格来说这样便算是作弊了。

在 MATALB 中使用 cvpartition 对数据集进行随机拆分，完成交叉验证。

4.5.2 常见的交叉验证方法

- Hold-Out Method

将原始数据随机分为两组，一组做为训练集，一组做为验证集。

- Double Cross Validation (2-fold Cross Validation, 记为 2-CV)

将数据集分成两个相等大小的子集，进行两回合的分类器训练。在第一回合中，一个子集作为 training set，另一个便作为 testing set；在第二回合中，则将 training set 与 testing set 对换后，再次训练分类器。

- K-fold Cross Validation (K-折交叉验证，记为 K-CV) (实验中采用的交叉验证方法)

将原始数据分成 K 组，将每个子集数据分别做一次验证集，其余的 K-1 组子集数据作为训练集，这样会得到 K 个模型，用这 K 个模型最终的验证集的分类准确率的平均数作为此 K-CV 下分类器的性能指标。K 一般大于等于 2，实际操作时一般从 3 开始取。

- Leave-One-Out Cross Validation (记为 LOO-CV)

将每个样本单独作为验证集，其余的 N-1 个样本作为训练集，所以 LOO-CV 会得到 N 个模型，用这 N 个模型最终的验证集的分类准确率的平均数作为此下 LOO-CV 分类器的性能指标。相比于前面的 K-CV，LOO-CV 有两个明显的优点：

- 每一回合中几乎所有的样本皆用于训练模型，因此最接近原始样本的分布，这样评估所得的结果比较可靠。
- 实验过程中没有随机因素会影响实验数据，确保实验过程是可以被复制的。

5. 识别结果分析

采用交叉验证的方法来评价分类器的性能好坏，所选用的数据集中的类别如图 28 所示，所采用的学习算法是随机森林。将该数据集随机均分为 n 份，其中每个子集数据分别做一次验证集，其余的 n-1 组子集数据作为训练集，这样会得到 n 个模型，用这 n 个模型最终的验证集的分类准确率的平均数作为此分类器的性能指标。最终的评价结果如图 29。

Error rate			0.2155														
Values prediction			Confusion matrix														
Value	Recall	1-Precision	Appendicularia	Bubble	Chaetognatha	CladoceraPenilia	Copepoda	Decapoda	Doliolida	Egg	Fiber	Gelatinous	Multiple	Nonbio	Pteropoda	Sum	
Appendicularia	0.8051	0.2039	2194	0	111	8	1	19	0	0	57	1	157	177	0	2725	
Bubble	0.8359	0.1049		606	0	0	0	0	0	29	0	0	3	87	0	725	
Chaetognatha	0.8934	0.0983	137	0	1559	0	5	2	0	0	7	1	18	14	2	1745	
CladoceraPenilia	0.8830	0.1044		0	0	0	4084	22	0	7	0	0	7	19	486	0	4625
Copepoda	0.8419	0.2108		3	0	0	9	6205	145	0	0	0	1	234	745	28	7370
Decapoda	0.7989	0.2157		4	0	0	0	339	2225	0	0	0	0	87	126	4	2785
Doliolida	0.7474	0.1418		0	0	0	2	1	0	1065	0	0	156	7	194	0	1425
Egg	0.7424	0.1401		0	66	0	13	2	1	1	1277	0	28	3	320	9	1720
Fiber	0.7414	0.2069		59	0	18	0	0	0	0	1127	0	62	254	0	1520	
Gelatinous	0.6914	0.2042		6	0	0	72	10	5	93	39	0	2081	79	624	1	3010
Multiple	0.3238	0.5093		267	0	23	51	491	202	9	23	45	91	1028	941	4	3175
Nonbio	0.8406	0.2452		86	5	17	321	716	189	66	117	185	246	390	12937	115	15390
Pteropoda	0.6627	0.1848		0	0	1	0	70	49	0	0	0	3	8	235	719	1085
			Sum	2756	677	1729	4560	7862	2837	1241	1485	1421	2615	2095	17140	882	47300

图 29: 评价结果

6. 特征

6.1 位置特征

BX 能够包围物体，且平行于图像两条边的最小外界矩形的左上角顶点的 X 坐标

BY 能够包围物体，且平行于图像两条边的最小外界矩形的左上角顶点的 Y 坐标

Height 能够包围物体，且平行于图像两条边的最小外界矩形的高

Width 能够包围物体，且平行于图像两条边的最小外界矩形的宽

XStart 图像最左上角像素点的 X 坐标

YStart 图像最左上角像素点的 Y 坐标

6.2 尺寸特征

Area 包含物体的矩形面积

Perim 物体最外层边缘的长度

Major 物体内切椭圆的长轴

Minor 物体内切椭圆的短轴

Circ Circularity = $(4 * \pi * \text{Area}) / \text{Perim}^2$; 值为 1 时表示近似圆形，值趋于 0 时表示逐渐变得瘦长

Feret Maximum feret diameter (最大费雷特径) , 沿物体边缘任意两个点的最长距离

6.3 灰度值特征

Min 物体内部所有像素点的最小灰度值 (0 = black)

Max 物体内部所有像素点的最大灰度值 (255 = white)

Mean 物体内的平均灰度值; 物体中所有像素点的灰度值的总和除以总的像素个数

IntDen Integrated density (总密度)。物体内像素点的灰度值的总和 (i.e. = Area*Mean)

StdDev 物体内像素的灰度值的标准差

Mode Modal grey value within the object

Skew 灰度直方图的偏斜度

Kurt 灰度直方图的峰值

Mean_exc Average grey value excluding holes within the object (= IntDen /Area_exc)

Median 物体内像素的灰度值的中位数

Slope 归一化的灰度累计直方图的斜率

Histcum1 灰度累计直方图的值为 25% 时所对应的灰度值

Histcum2 灰度累计直方图的值为 50% 时所对应的灰度值

Histcum3 灰度累计直方图的值为 75% 时所对应的灰度值

6.4 形状特征

Fractal 物体边界的分形维数 (Berube and Jebrak, 1999)

Skelarea 骨架像素的表面积

6.5 自定义特征

$$\text{ESD} \quad 2 \times \sqrt{\frac{\text{Area}}{\pi}}$$

$$\text{Elongation} \quad \frac{\text{Major}}{\text{Minor}}$$

$$\text{Range} \quad \text{Max} - \text{Min}$$

$$\text{MeanPos} \quad \frac{\text{Mean} - \text{Max}}{\text{Max} - \text{Min}}$$

$$\text{CentrodisD} \quad \sqrt{(\bar{X}M - X)^2 + (\bar{Y}M - Y)^2}$$

$$\text{CV} \quad 100 \times \frac{\text{StdDev}}{\text{Mean}}$$

$$\text{SR} \quad 100 \times \frac{\text{StdDev}}{\text{Max} - \text{Min}}$$

$$\text{PerimAreaexc} \quad \frac{\text{Perim}}{\sqrt{\text{Area_exc}}}$$

$$\text{FeretAreaexc} \quad \frac{\text{Feret}}{\sqrt{\text{Area_exc}}}$$

$$\text{PerimFeret} \quad \frac{\text{Perim}}{\text{Feret}}$$

$$\text{PerimMaj} \quad \frac{\text{Perim}}{\text{Major}}$$

$$\text{Circexc} \quad \frac{4 \times \pi \text{Area_exc}}{\text{Perim}^2}$$

$$\text{CDexc} \quad \frac{\sqrt{(\bar{X}M - X)^2 + (\bar{Y}M - Y)^2}}{\sqrt{\text{Area_exc}}}$$

7. 优缺点分析