《数据挖掘》实验

班级 计科11606 姓名 张仁兵 成绩 \_\_\_\_\_\_\_

3. 实验三、聚类构建与簇评测

3.0 实验内容与目的

（1）以鸢尾花数据集为基础，使用Weka平台，构建层级式聚类和k均值聚类。

（2）对聚类结果进行可视化，采用基于类属性的评测方式，对不同聚类方法所得结果进行评估。

操作方法与要点如下：

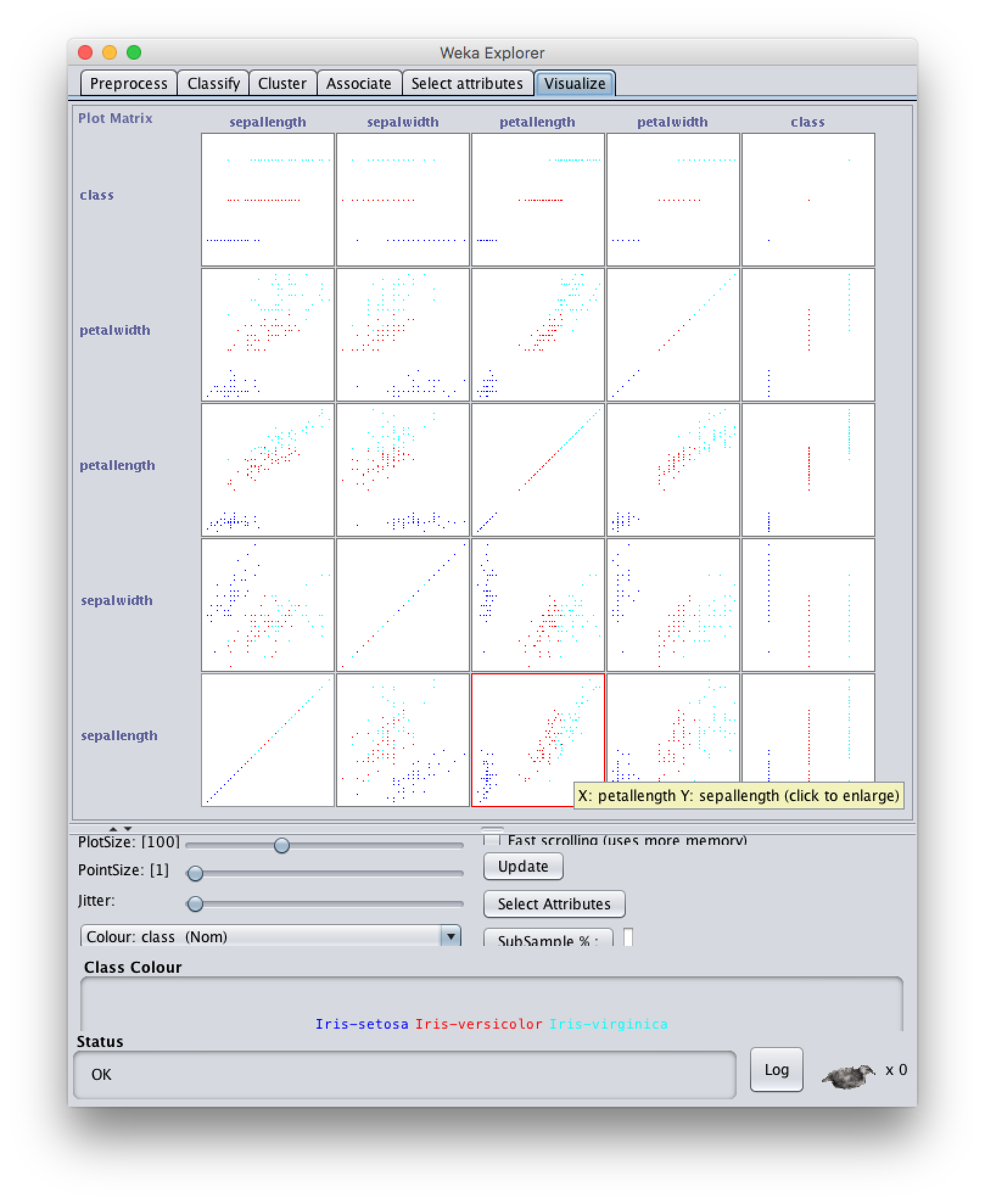
3.1 运行Weka：

* 运行Weka：“开始-程序-Weka”，将启动Weka图形界面选择程序（GUI chooser）。
* 点击 “Explorer”打开主界面。

3.2 装载数据集和数据集可视化：

* Preprocess面板中，点击<Open file…>，选择Weka安装目录中data文件夹下的iris.arff文件。
* 数据集可视化：切换到Visualize面板，将看到如下图所示的可视化结果。其中每个小方块代表一种属性对的可视化结果。比如最左上角的方块表示以sepallength为横轴，以class为纵轴的可视化，即对数据集中每条实例，以其（sepallength，class）为坐标进行画图。另外，不同类别的实例将用不同颜色画图，界面下方的Class Colour部分显示了类别的颜色取值。
* Plot Matrix结构：由于鸢尾花数据集有5个属性（包括类别属性），因此可视化结果由25个方块组成，构成了一个类似于对称矩阵的结构，称为Plot Matrix。也就是说，假如沿对角线进行折叠，左上方方块中的画图结果与右下方的画图结果能完全重叠。
* 通过可视化结果观察属性分布特征：点击第一行中的前四个方块，即以类别属性为纵轴，分别以四个属性为横轴的画图结果。点击方块将在新窗口中展示详细信息。比较四个属性的类别分布情况，**观察哪个属性对于区分Iris-setosa类的实例最有效？为什么？**

**Petallength属性对区分Iris-setosa类的实例最为有效，分布更为集中**

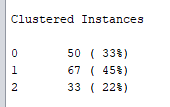
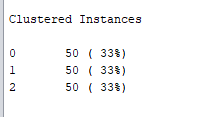


3.3 构建层级式聚类：

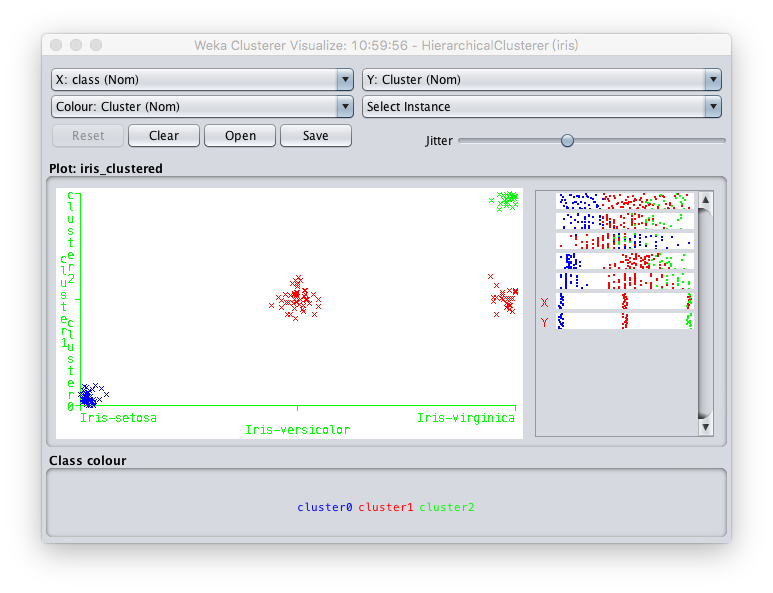
* 点击Cluster面板，在Clusterer部分点击Choose，将显示聚类算法的列表。接下来要构建的两种聚类算法分别是：HierarchicalClusterer（层级式聚类）和SimpleKMeans（k均值聚类）。
* 选择HierarchicalClusterer作为聚类器，打开其参数配置窗口。其中，numClusters设置聚类数目，设为3。linkType设置聚类间距离的计算方法，点击More按钮将显示每个选项的详细说明。**若要将所有聚类成员间距离的平均值作为类间距离，应该选择哪个选项？**

**应选择AVERAGE**

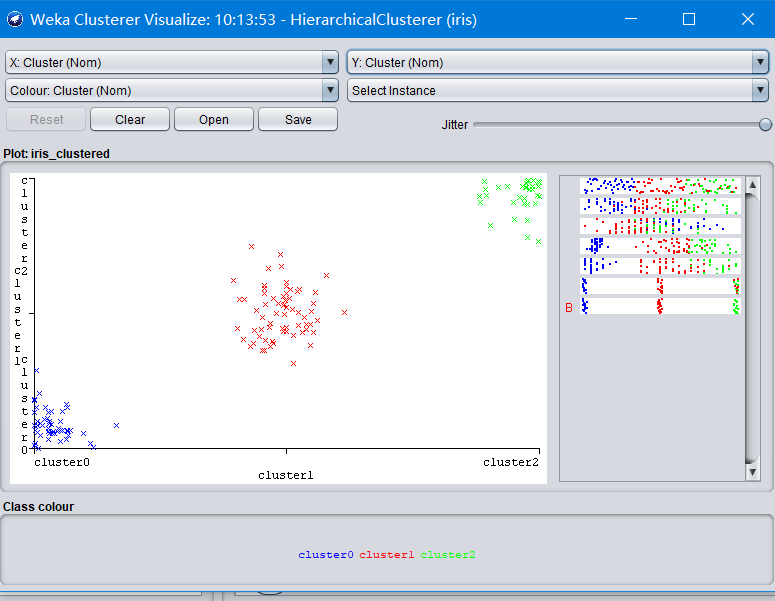
* 运行聚类算法：配置好参数之后，能不能直接点击Start运行聚类算法呢？此时得到的结果是怎样的？看上去很准确，有没有问题？回到Preprocess面板，鸢尾花数据集本身是带有类属性的（即class属性），而聚类过程应该**不能**看到该属性，即只根据花瓣和花萼的性质来进行聚类。因此，我们需要在聚类过程中排除类属性。点击Ignore attributes，选择class，再次运行聚类算法，观察此时得到的结果，并与之前得到的结果进行比较。

* 聚类结果评估：下面介绍两种评判聚类结果质量的方法。第一种是主观评价法。在Result list版块中右键点击上一次的聚类结果，选择Visualize cluster assignment，在打开的可视化面板中，属于不同聚类的实例将以不同颜色的点显示。由于鸢尾花数据集自身带有类别属性，我们可以将聚类所得到的分组与原始类别进行比较。在第一个下拉列表中选择X:class属性作为X轴，在第二个下拉列表中选择Y: Cluster作为Y轴。缺省状态下重叠的数据点将覆盖显示，因此图中只显示了几个点，将设置Jitter参数的拖动条向右滑动，即使重叠点不重叠显示，可以看到不同类别实例在聚类结果中的分布情况，如下图所示。可以看到，层级式聚类能将Iris-setosa类的实例与其他两类的实例区分开，Iris-virginica类的实例也能很好的聚集，而Iris-versicolor类的部分实例容易与Iris-virginica类的实例混淆。



* **尝试不同的参数配置，将你认为比较准确的聚类结果的可视化截图并粘贴到下面。**

****

* 第二种评估方法是利用数据集的类别属性，将聚类结果与类别属性进行对照，评估聚类算法对“还原”数据集原有分类的能力。在Cluster面板的Cluster mode板块中，选择Classes to clusters evaluation，并在下拉列表中选择(Nom) class属性作为类别属性。注意，当选择这种评估方式时，聚类过程会自动忽略所选择的类别属性，只在评估时用到类别信息。点击Start重新运行聚类算法。观察右侧Clusterer output中的输出的变化。**将你认为比较准确的聚类结果的可视化截图并粘贴到下面。**



* 聚类过程可视化。在Result list版块中右键点击某次聚类结果，选择Visualize tree，将显示代表聚类过程的树状结构（称为Dendrogram），每个分支代表一次合并，即将两个聚类合并成一个聚类。观察树状结构，思考：如果要以相同的算法设置得到5个聚类、7个聚类、10个聚类，即只改变聚类的数目，是否需要重新运行层级聚类算法？为什么？

不需要，只需在原有的基础上相应的合成目标聚类即可

3.4 构建k均值聚类器

* 选择SimpleKMeans，将numClusters设置为3，选择Classes to clusters evaluation。**所得到的三个聚类各包含多少实例？**

Cluster 0 <-- Iris-versicolor

Cluster 1 <-- Iris-setosa

Cluster 2 <-- Iris-virginica

* 再次将聚类结果可视化，以类属性为横轴，以聚类为纵轴，调整Jitter取值，**将可视化的聚类结果截图粘贴在下方。**

