数据挖掘实验报告

1、工具介绍

Weka 的全名是怀卡托智能分析环境（ Waikato Environment for Knowledge Analysis ），是一款免费的，非商业化的，基于JAVA 环境下开源的机器学习以及数据挖掘软件。它和它的源代码可在其官方网站下载。有趣的是，该软

件的缩写WEKA 也是New Zealand 独有的一种鸟名， 而Weka 的主要开发者同时恰好来自新西兰的the University of Waikato。Weka 提供的功能有数据处理，特征选择、分类、回归、聚类、关联规则、可视化等。本文将对Weka 的使用做一个简单的介绍，并通过简单的示例，使大家了解使用weka 的流程。本文将仅对图形界面的操作做介绍，不涉及命令行和代码层面的东西。

2、安装

Weka 的官方地址是http://www.cs.waikato.ac.nz/ml/weka/ 。点开左侧download 栏，可以进入下载页面，里面有windows ，mac os， linux 等平台下的版本，我们以windows 系统作为示例。目前稳定的版本是3.6。

如果本机没有安装java，可以选择带有jre 的版本。下载后是一个exe的可执行文件，双击进行安装即可。安装完毕，打开启动weka 的快捷方式，如果可以看到下面的界面，那么恭喜，安装成功了。



窗口右侧共有4 个应用，分别是

1） Explorer

用来进行数据实验、挖掘的环境， 它提供了分类， 聚类，关联规则， 特征选择， 数据可视化的功能。（ An environment

for exploring data with WEKA ）

2） Experimentor

用来进行实验，对不同学习方案进行数据测试的环境。（ An environment for performing experiments and

conducting statistical tests between learning schemes. ）

3）KnowledgeFlow

功能和Explorer 差不多，不过提供的接口不同，用户可以使用拖拽的方式去建立实验方案。另外，它支持增量

学习。（ This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One

advantage is that it supports incremental learning. ）

4）SimpleCLI

简单的命令行界面。（ Provides a simple command-line interface that allows direct execution of WEKA commands for

operating systems that do not provide their own command line interface. ）

3、数据格式

Weka 支持很多种文件格式，包括arff、xrff 、csv，甚至有libsvm 的格式。其中， arff 是最常用的格式，我们在这里仅介绍这一种。例子来自于weka 安装目录data 文件下的labor.arff 文件， 来源于加拿大劳资谈判的案例，它根据工人的个人信息，来预测劳资谈判的最终结果。

attribute-name 是特征的名称，后面是数据类型，常用数据类型有以下几种

1） numeric，数字类型，包括integer（整数）和real（实数）

2） nominal ，可以认为是枚举类型，即特征值是有限的集合，可以是字符串或数字。

3） string ，字符串类型，值可以是任意的字符串。

从“ @data”开始，是实际的数据部分。每一行代表一个实例，可以认为是一个特征向量。各个特征的顺序与头信息中的attribute 逐个对应，特征值之间用逗号分割。在有监督分类中，最后一列是标注的结果。

某些特征的数值如果是缺失的，可以用“？”代替。

4、数据挖掘流程

使用weka进行数据挖掘的流程如下

特征抽取

数据源

Weka格式的文件

数据预处理

训练

No

验证

满足要求?

模型

Yes

其中，在weka 内进行的是数据预处理，训练，验证这三个步骤。

1）数据预处理

数据预处理包括特征选择，特征值处理（比如归一化） ，样本选择等操作。

2）训练

训练包括算法选择，参数调整，模型训练。

3）验证

对模型结果进行验证。

本文剩余部分将以这个流程为主线，以分类为示例，介绍使用weka 进行数据挖掘的步骤。

5、数据预处理

**1）case 1**：特征值归一化

该项功能与类别无关，且是针对attribute 的，我们选择unsupervised -> attribute 下 面的Normalize。点开Normalize 所在的区域，将看到如下界面。左边的窗口，有几个参数可以选择。点击more，将出现右边的窗口，该窗口详细介绍了此功能。使用默认参数，点击ok，回到主窗口。在区域 4选好将要归一化的特征，可以是一个或多个，然后点击apply 。在可视化区域中，我们可以看到特征值从1 到 3 被归一到了0 到 1 之间。

**2）case 2:** 分类器特征筛选

该功能与类别相关， 选 择 supervised -> attribute 下 面的AttributeSelection 。该界面有两个选项， evaluator 是评价特征集合有效性的方法，search 是特征集合搜索的方法。在这里，我们使用InformationGainAttributeEval作为evaluator，使用 Ranker 作为 search，表示我们将根据特征的信息增益值对特征做排序。Ranker 中可以设置阈值，低于这个阈值的特征将被扔掉。点击apply，可以看到区域四里特征被重新排列，低于阈值的已被删掉。

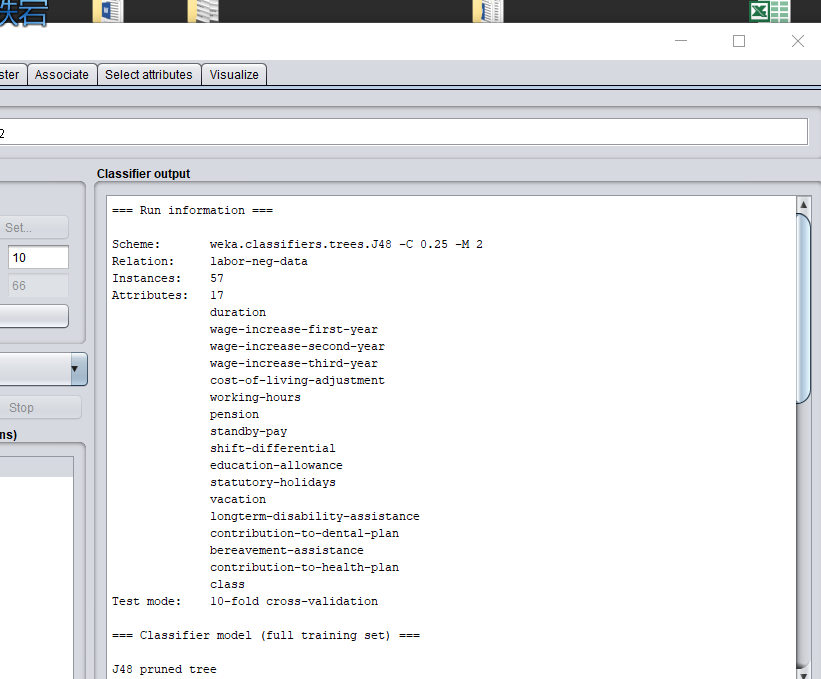
**3）case 3**：选择分类器错分的样本

选择 unsupervised -> instance 下面的RemoveMisclassified，可以看到6 个参数，classIndex 用来设置类别标签， classifier用来选择分类器，这里我们选择J48 决策树，invert我们选择true，这样保留的是错分样本，numFolds 用来设置交叉验证的参数。设置好参数之后，点击apply，可以看到样本的数量从57 减少到了7。

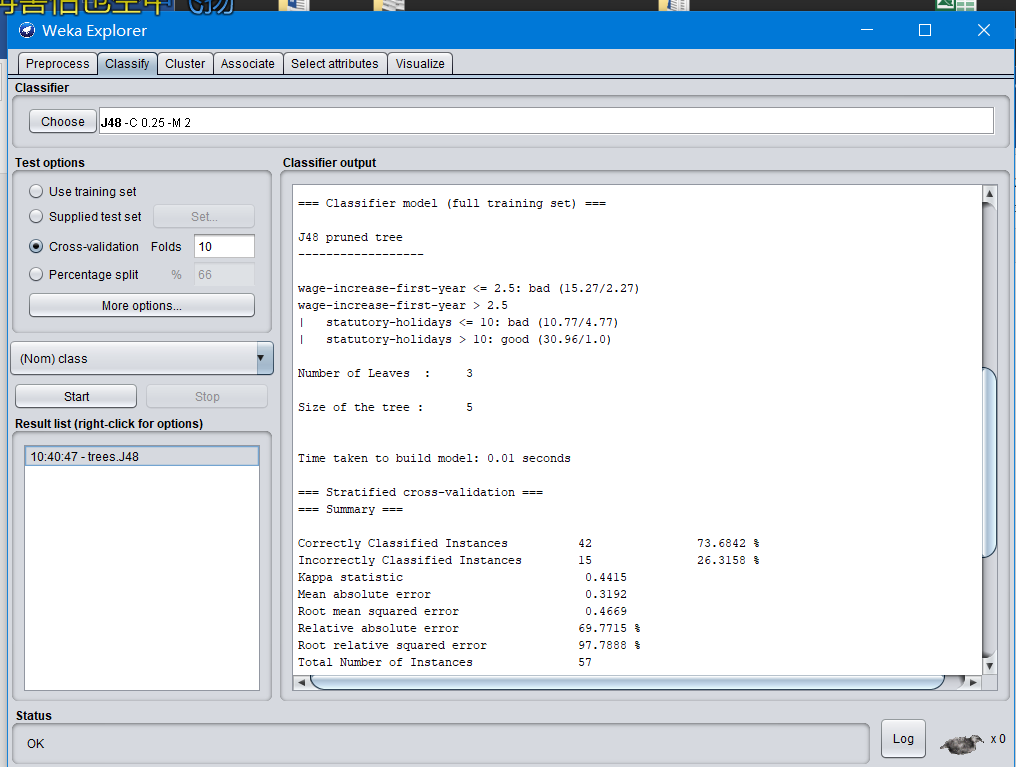
6、分类

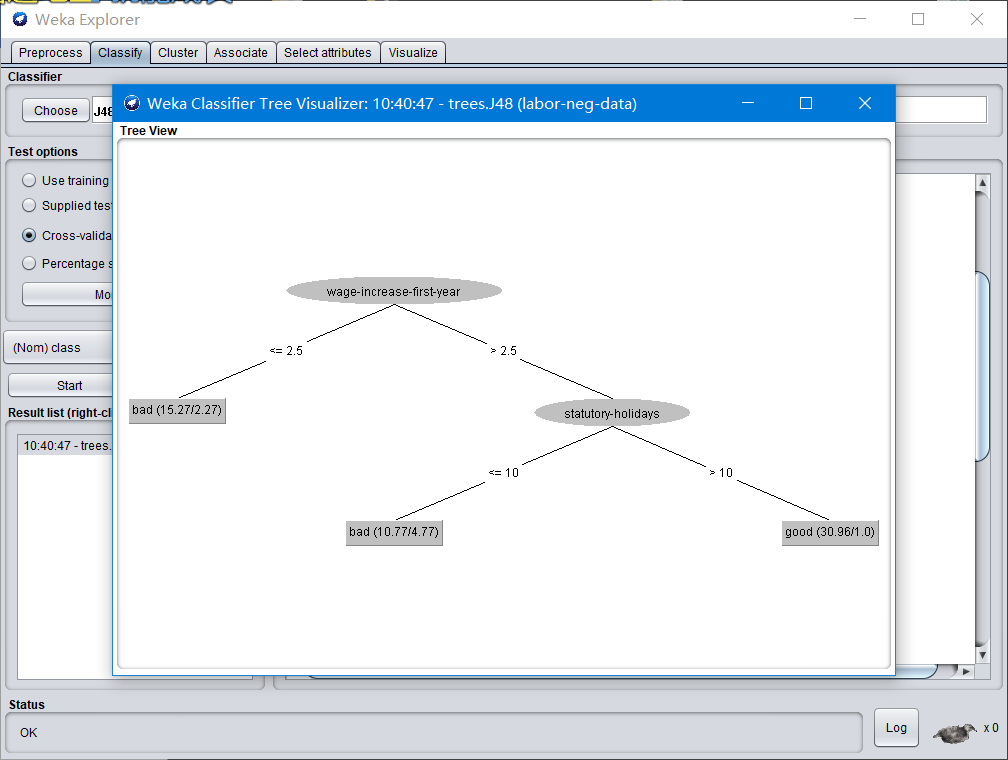
使用后J48对labor文件做分类

1. 打开 labor.arff 文件，切换到classify 面板
2. 选择 trees->J48 分类器，使用默认参数。
3. Test options 选择默认的十折交叉验证，点开More options ，勾选 Output predictions 。
4. 点击 start 按钮，启动实验。
5. 在右侧的Classifier output 里面，我们看到了实验的结果。



上图给出了实验用的分类器以及具体参数，实验名称，样本数量，特征数量以及所用特征，测试模式。

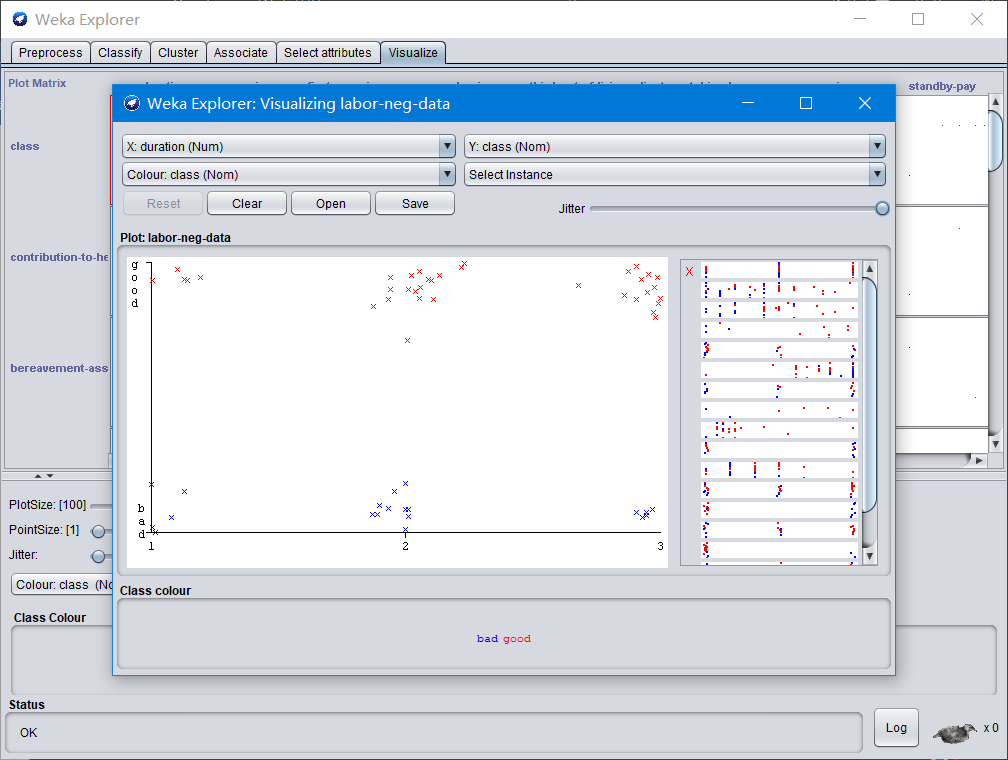


上图给出了生成的决策树，以及叶子节点数、树的节点数、 模型训练时间。 如果觉得这样不直观，可以在Result list 里面右键点击刚刚进行的实验，点击Visualize Tree，可以看到图形界面的决策树，十分直观。

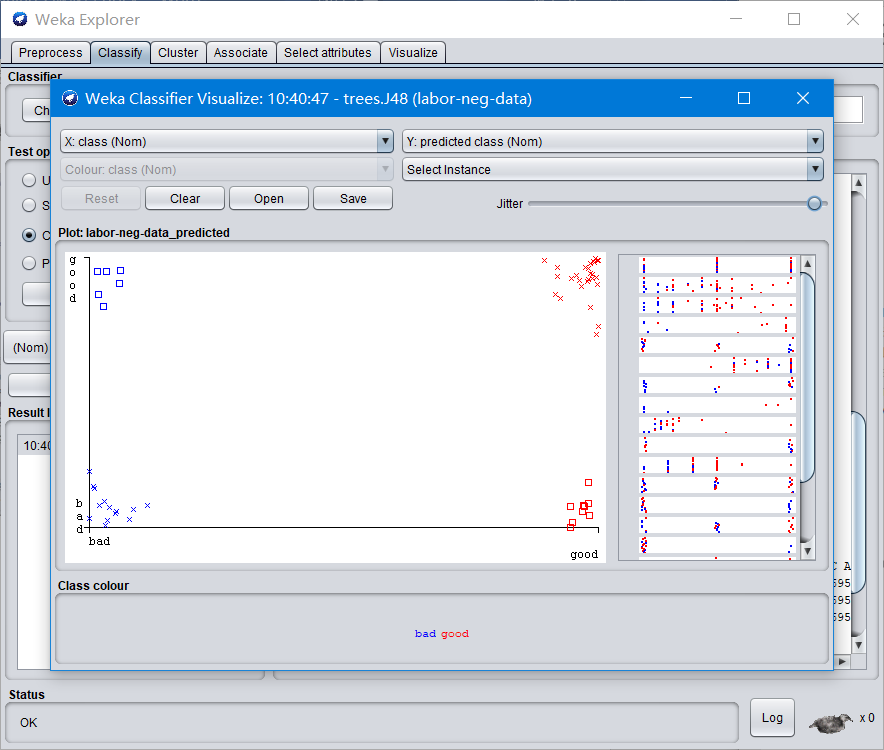
最下面是验证结果， 整体的 accuracy 是 73.68% ，bad 类准确率是60.9%，召回率70.0% ，good 类准确率是82.4% ， 召回率 75.7% 。

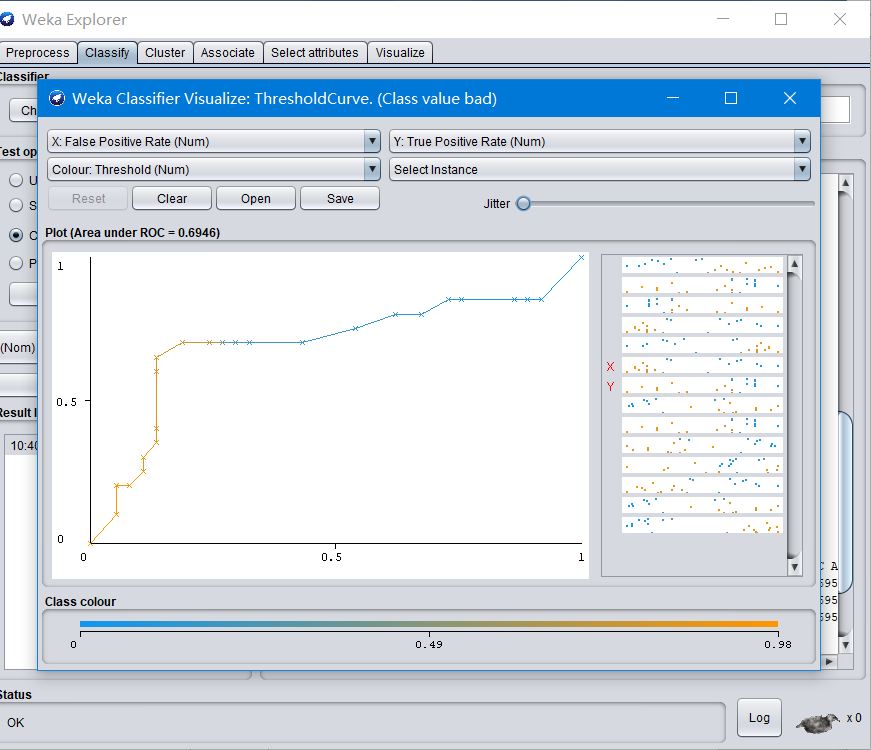
# 7、可视化

打开 Explorer 的 Visualize 面板， 可以看到最上面是一个二维的图形矩阵，该矩阵的行和列均为所有的特征（包括类别标签） ，第 i 行第 j 列表示特征i 和特征 j 在二维平面上的分布情况。图形上的每个点表示一个样本，不同的类别使用不同的颜色标识。下面有几个选项，PlotSize 可以调整图形的大小，PointSize 可以调整样本点的大小，Jitter 可以调整点之间的距离，有些时候点过于集中，可以通过调整Jitter 将它们分散开。



上图是duration 和class 两个特征的图形，可以看出， duration 并不是一个好特征，在各个特征值区间， good和bad 的分布差不多。单击某个区域的图形，会弹出另外一个窗口，这个窗口给出的也是某两个特征之间分布的图形，不同的是，在这里，通过点击样本点，可以弹出样本的详细信息。可视化还可以用来查看误分的样本，这是非常实用的一个功能。分类结束后，在Result list 里右键点击分类的记录，选择Visualize classify errors ，会弹出如下窗口。



这个窗口里面,十字表示分类正确的样本， 方块表示分类错误的样本， X 轴为实际类别， Y 轴为预测类别， 蓝色为实际的bad，红色为实际的good。这样，蓝色方块就表示实际为bad，但为误分为good 的样本，红色方块表示实际为good，被误分为bad 的样本。单击这些点，便可以看到该样本的各个特征值，分析为什么这个样本被误分了。再介绍一个比较实用的功能，右键点击Result list 里的记录，选择Visualize threshold curve ，然后选好类别，可以看到如下图形

该图给出的是分类置信度在不同阈值下， 分类效果评价标准的对比情况。上图给出的是假阳性比率和真阳性比

率在不同阈值下的对比，其实给出的就是ROC 曲线。我们可以通过选择颜色，方便地观察不同评价标准的分布情

况。如果X 轴和Y 轴选择的是准确率和召回率，那我们可以通过这个图，在这两个值之间做trade-off ，选择一个合

适的阈值。

8、小结

本次实验基本上都是安装网上的教程来一步步操作的，基本上是傻瓜式模仿，但是通过本次实验还是知道了一些数据挖掘数据分析的一些方法和工具，至少在提及数据挖掘数值分析的时候不会完全不熟悉，也知道了weka工具的一些功能和使用weka进行数据挖掘的基本流程。虽然并没有弄懂其中的原理，但是还是有收获的。