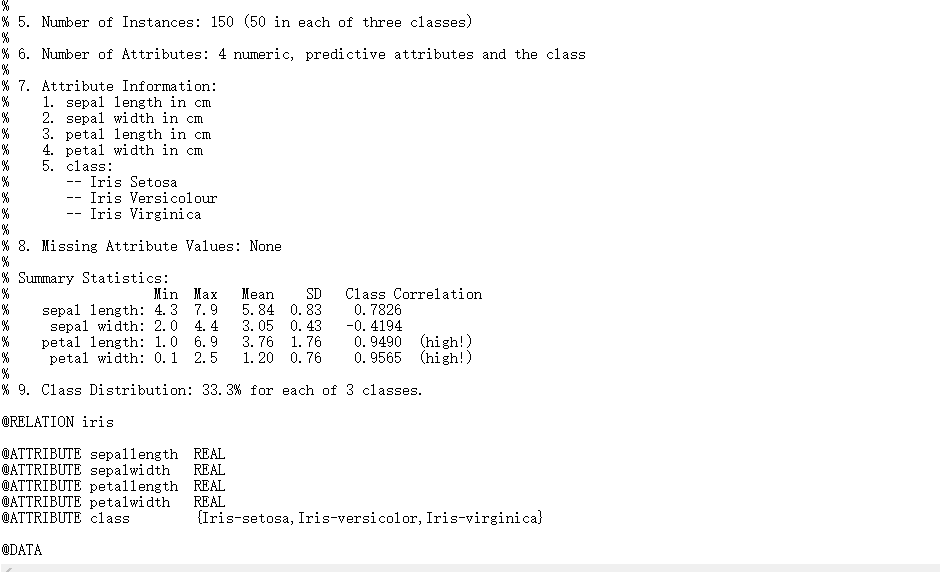
**Iris数据集的分类**

**2017Z8009061078 李中欢**

1. **数据集**

Iris数据集以鸢尾花的特征作为数据来源，包含150个数据，分为3类，每类50个数据，每个数据包含4个属性，三类分别为:setosa, versicolor, virginica，数据包含4个独立的属性,这些属性变量测量植物的花朵,比如萼片长度, 萼片宽度,花瓣长度, 花瓣宽度.



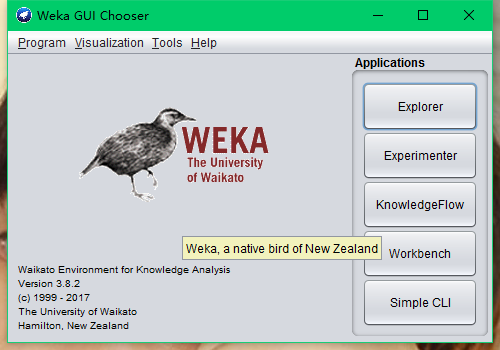
1. **实验目标和工具**

实验目标：1）熟悉机器学习一般算法。2）学会使用weka进行数据分析。

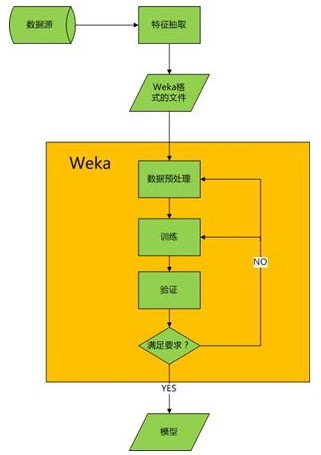
实验工具：1）使用Weka软件，2）python语言，3） 使用Tensorflow深度学习框架搭建简单神经网络。

1. **使用Weka进行数据分类**

由于数据提供支持weka软件的.arff数据格式，所以我们使用该格式利用软件进行分析。打开软件，我们会看到如下图所示的界面。



使用weka进行数据挖掘的流程如下：

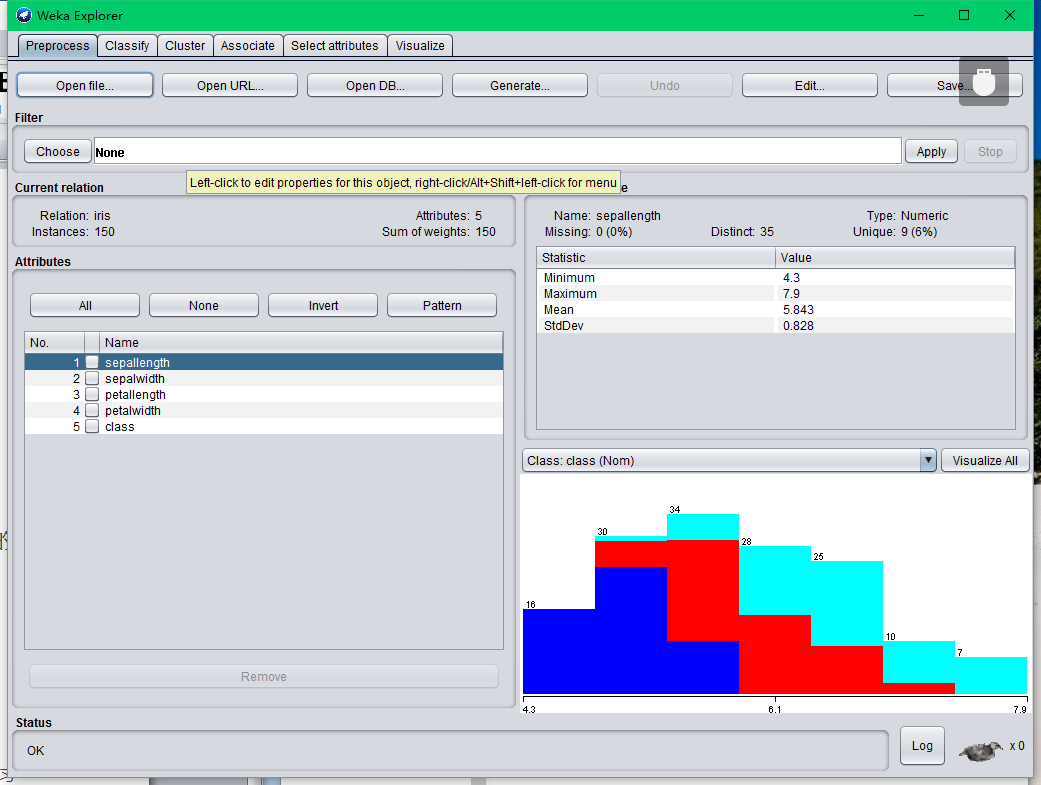


其中，黄色部分是我们利用软件进行的步骤，需要进行数据预处理，训练，验证这三个步骤。

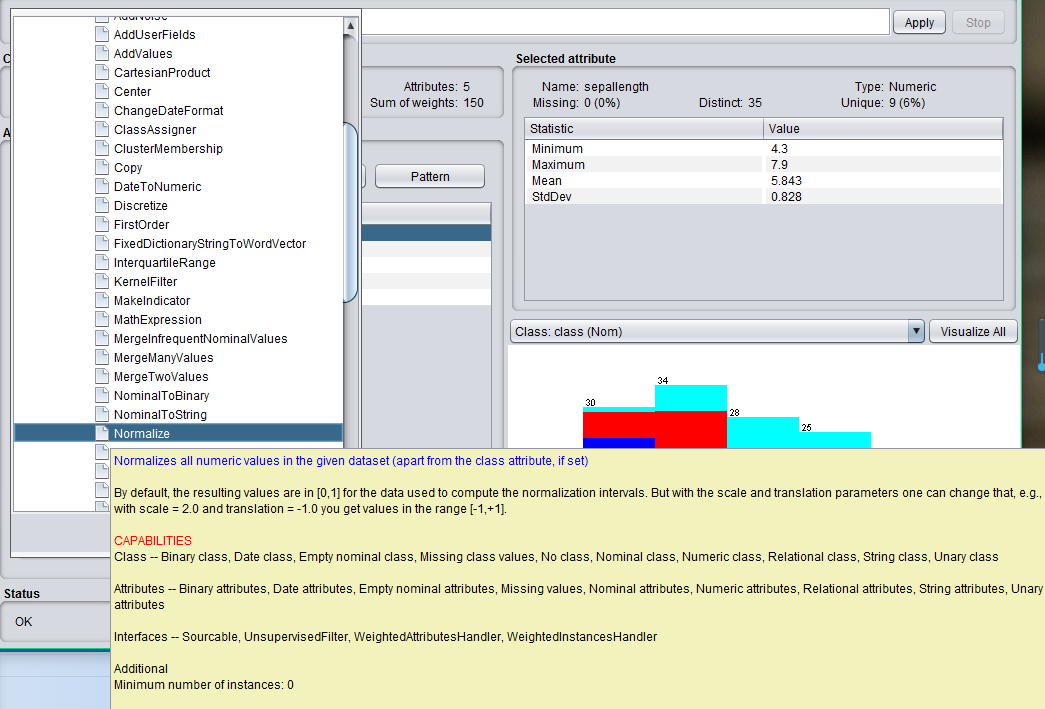
1）数据预处理  
数据预处理包括特征选择，特征值处理（比如归一化），样本选择等操作。  
 2）训练  
训练包括算法选择，参数调整，模型训练。  
 3）验证  
对模型结果进行验证。

1. **数据预处理**

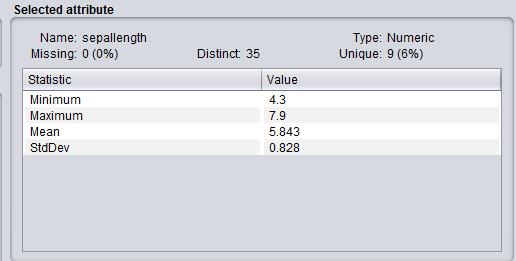
打开Explorer界面，点“open file”，选择data数据中的“iris.arff”文件，将会看到如下界面：

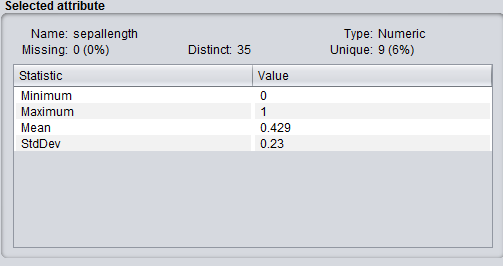


1.特征值归一化

点击Filter下的choose按钮，选择unsupervised->attribute->Normalize，

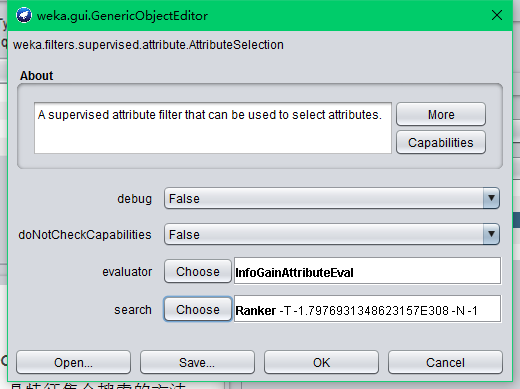
在Attributes中选择要归一化的特征，然后点击apply，在可视化区域中，可以看到特征被归一到了0到1之间，对比如下：



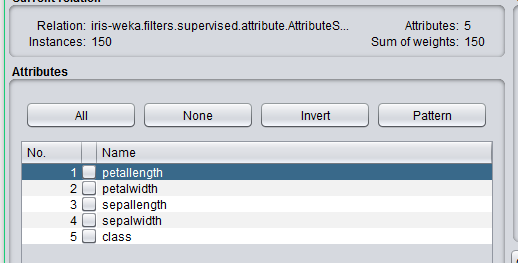


2.分类器特征筛选

选择supervised->attribute->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。

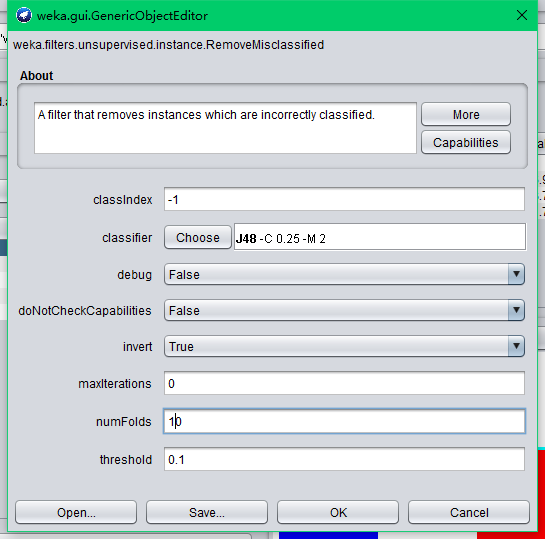


点击apply后，特征值被重新排序；



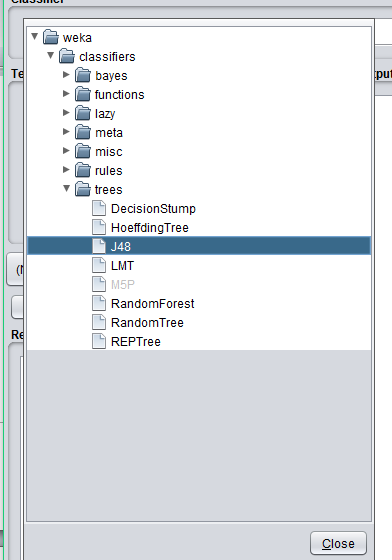
3.选择分类器错分的样本

选择unsupervised->instance->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。



**五、分类**

使用j48对文件做分类，打开iris.arff文件，切换到classify面板，选择tree->J48分类器，使用默认参数；



Test options选择默认的十字交叉验证，点开more options，勾选Output predictions，点击start，启动实验；右侧Classifier output，可得实验结果：

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: iris-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.InfoGainAttributeEval-Sweka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1-weka.filters.unsupervised.instance.RemoveMisclassified-Wweka.classifiers.trees.J48 -C 0.25 -M 2-C-1-F10-T0.1-I0-V

Instances: 17

Attributes: 5

petallength

petalwidth

sepallength

sepalwidth

class

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

------------------

petalwidth <= 0.6: Iris-setosa (2.0)

petalwidth > 0.6

| sepalwidth <= 3: Iris-virginica (11.0/4.0)

| sepalwidth > 3: Iris-versicolor (4.0)

Number of Leaves : 3

Size of the tree : 5

Time taken to build model: 0 seconds

=== Predictions on test data ===

inst# actual predicted error prediction

1 2:Iris-versicolor 3:Iris-virginica + 0.667

2 3:Iris-virginica 3:Iris-virginica 0.667

1 2:Iris-versicolor 2:Iris-versicolor 1

2 3:Iris-virginica 3:Iris-virginica 0.6

1 2:Iris-versicolor 3:Iris-virginica + 1

2 3:Iris-virginica 2:Iris-versicolor + 0.857

1 2:Iris-versicolor 3:Iris-virginica + 1

2 3:Iris-virginica 3:Iris-virginica 1

1 2:Iris-versicolor 2:Iris-versicolor 1

2 3:Iris-virginica 3:Iris-virginica 0.6

1 2:Iris-versicolor 3:Iris-virginica + 1

2 3:Iris-virginica 3:Iris-virginica 0.75

1 2:Iris-versicolor 2:Iris-versicolor 1

2 3:Iris-virginica 3:Iris-virginica 0.6

1 2:Iris-versicolor 2:Iris-versicolor 1

1 1:Iris-setosa 2:Iris-versicolor + 0.8

1 1:Iris-setosa 2:Iris-versicolor + 0.8

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 10 58.8235 %

Incorrectly Classified Instances 7 41.1765 %

Kappa statistic 0.2699

Mean absolute error 0.3101

Root mean squared error 0.4896

Relative absolute error 75.0553 %

Root relative squared error 107.6648 %

Total Number of Instances 17

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.000 0.000 ? 0.000 ? ? 1.000 1.000 Iris-setosa

0.500 0.333 0.571 0.500 0.533 0.169 0.556 0.729 Iris-versicolor

0.857 0.400 0.600 0.857 0.706 0.457 0.657 0.502 Iris-virginica

Weighted Avg. 0.588 0.322 ? 0.588 ? ? 0.650 0.667

=== Confusion Matrix ===

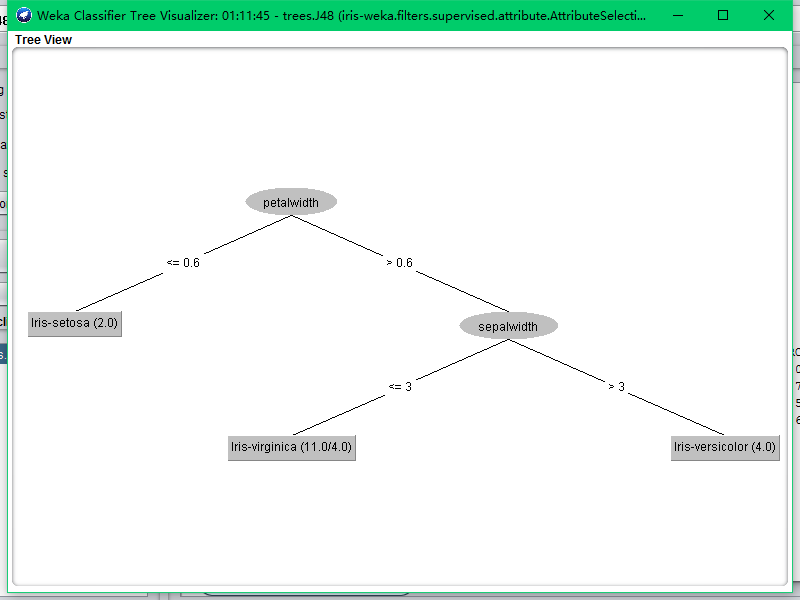
a b c <-- classified as

0 2 0 | a = Iris-setosa

0 4 4 | b = Iris-versicolor

0 1 6 | c = Iris-virginica

结果给出了实验用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。另外还有生成的决策树信息，叶子节点数、树的节点数、模型训练时间，图形界面的决策树如下：



最后给出了预测结果，可以看到每个样本的实际分类，预测分类，是否错分，预测概率等信息；验证结果显示，整体准确率是58.8235%，三种类Iris-setosa、Iris-versicolor、Iris-virginica准确率分别是无、57.1%、60%，召回率0、50%、85.7%。

**六、总结**

经过这次机器学习算法在iris数据集上分类任务，熟悉了普遍机器学习算法，本文仅仅针对weka的Explorer界面的某些功能做了介绍，Explorer其它的功能，比如聚类、关联规则、特征选择。