大数据基础与应用
BIG DATA FOUNDATION AND APPLICATION

数据分析算法

北京理工大学计算机学院　孙新

2019年1月

用决策树归纳分类

- 决策树的生成由两个阶段组成
  - 判定树构建
    - 开始时，所有的训练样本都在根节点
    - 递归的通过选定的属性，来划分样本 （必须是离散值）
  - 树剪枝
    - 许多分枝反映的是训练数据中的噪声和孤立点，树剪枝试图检测和剪去这种分枝
- 决策树的使用：对未知样本进行分类
  - 通过将样本的属性值与判定树相比较

决策树算法

- 大多数决策树学习算法是一种核心算法的变体，采用自顶向下的贪婪搜索遍历可能的决策树空间
- ID3算法是用来构造决策树的常用算法，该方法使用信息增益选择测试属性。
- C4.5由J.Ross Quinlan 在ID3的基础上提出的。用信息增益率来选择属性，克服了用信息增益来选择属性时偏向选择值多的属性的不足，同时可以处理连续数值型属性

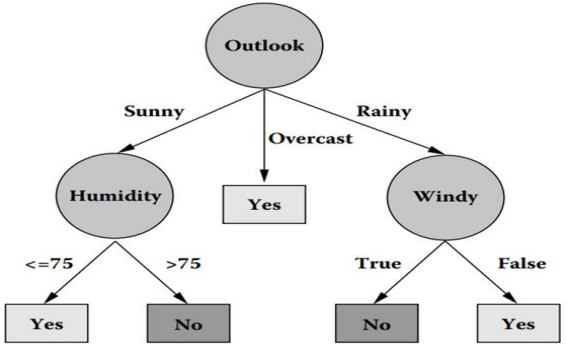
- ID3算法通过自顶向下构造决策树来进行学习
- 构造过程：
  - 选择根节点 - 使用统计测试确定每一个实例属性单独分类训练样例的能力，分类能力最好的属性被选作树的根节点
  - 为根节点属性的每个可能值产生一个分支，并把训练样例排列到适当的分支
  - 重复上面的过程，用每个分支节点关联的训练样例来选取在该点被测试的最佳属性，直到满足以下两个条件中的任一个：
    1) 所有的属性已经被这条路径包括；
    2)与这个节点关联的所有训练样例具有相同的目标属性值

ID3算法的核心问题是选取在树的每个节点要测试的属性。

决策树算法的关键问题一：分裂属性选择

- 信息增益：选择具有最高信息增益的属性来作为节点的分裂属性。
-
- D中信息量的定义

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$



- 按照属性A划分D中的元组，且属性A将D划分成v个不同的类。在该划分之后，信息量的定义

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

决策树算法的关键问题二：剪枝。

- 剪枝主要分为两种方法：先剪枝和后剪枝（C4.5采用后剪枝方法）。
  - 先剪枝方法：通过提前停止树的构造（比如决定在某个节点不再分裂或划分训练元组的子集）而对树剪枝。
  - 后剪枝方法：由完全成长的树剪去子树而形成，通过删除节点的分枝并用树叶来替换它，而树叶一般用子树中最频繁的类别来标记。
  - 最常用的终止条件包括

（1）决策树达到一定的高度;
（2）到达某节点的实体个数小于某个阈值;
（3）每次扩展对系统性能的增益小于某个阈值。

决策树算法

- 算法小结：决策树是一种类似二叉树或多叉树的树结构。
  - 树中的每个非叶结点（包括根结点）对应于训练样本集总一个非类属性的测试，
  - 非叶结点的每一个分支对应属性的一个测试结果，
  - 每个叶结点代表一个类或类分布。
  - 从根结点到叶子结点的一条路径形成一条分类规则。
- 决策树可以很方便地转化为分类规则，一种非常直观的分类模型的表示形式。
- 决策树是有监督的学习方法，属于一种归纳学习算法

归纳学习（Inductive Learning）是在从大量经验数据中归纳抽取一般的判定规则和模式，是机器学习中最核心、最成熟的分支。

4.2 决策树算法

决策树的优点：
（1）模型直观清晰，计算量相对来说不是很大；
（2）可以生成可理解的规则，分类规则易于解释；
（3）可以处理连续和离散字段；
（4）决策树可以清晰的显示哪些字段比较重要，提供了将学习结果决策树到等价规则集的转换功能

决策树的不足：
（1）对连续性的字段比较难预测
（2）当类别太多时，错误可能会增加的比较快
（3）一般的算法分类的时候，只是根据一个属性来分类。
（4）不是全局最优

采用Weka分析软件学习分类算法

- Weka简介 WEKA的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis）。
- WEKA作为一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理，分类，回归、聚类、关联规则以及在新的交互式界面上的可视化。

❖ **Weka下载网址：**

❖ http://www.cs.waikato.ac.nz/ml/weka/downloading.html

UCI数据集网址：
http://archive.ics.uci.edu/ml/

采用Weka分析软件学习分类算法

- 鸢尾花卉数据集（Iris数据集）是常用的分类实验数据集。
- 数据集包含150个数据集，分为3类，每类50个数据。每个数据包含4个属性，可通过花萼长度，花萼宽度，花瓣长度，花瓣宽度4个属性预测鸢尾所属分类（Setosa，Versicolour，Virginica）

sepal length in cm （花萼长度）	Sepal width in cm （花萼宽度）	petal length in cm （花瓣长度）	petal width in cm（花瓣宽度）	class
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
6	2.2	4	1	versicolor
6.9	3.2	5.7	2.3	virginica
……				

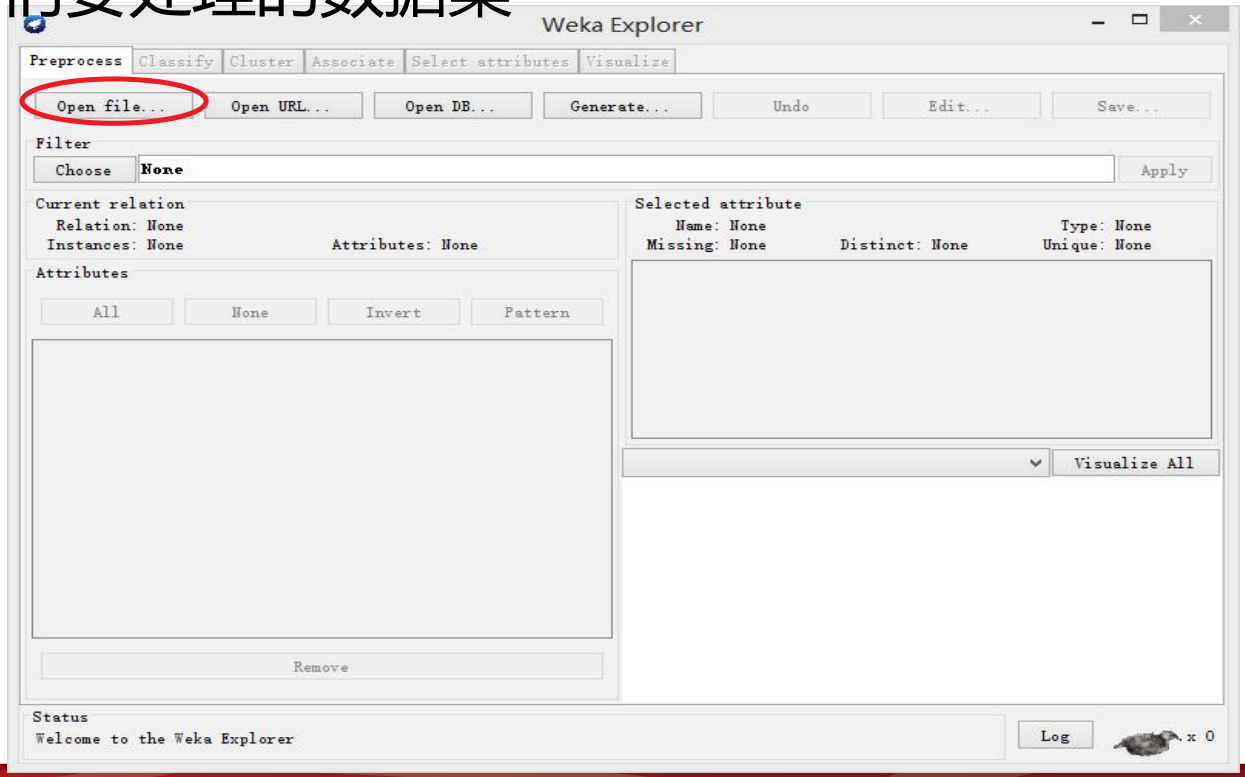
http://archive.ics.uci.edu/ml/

采用Weka分析软件学习分类算法

- Weka数据挖掘软件的主界面

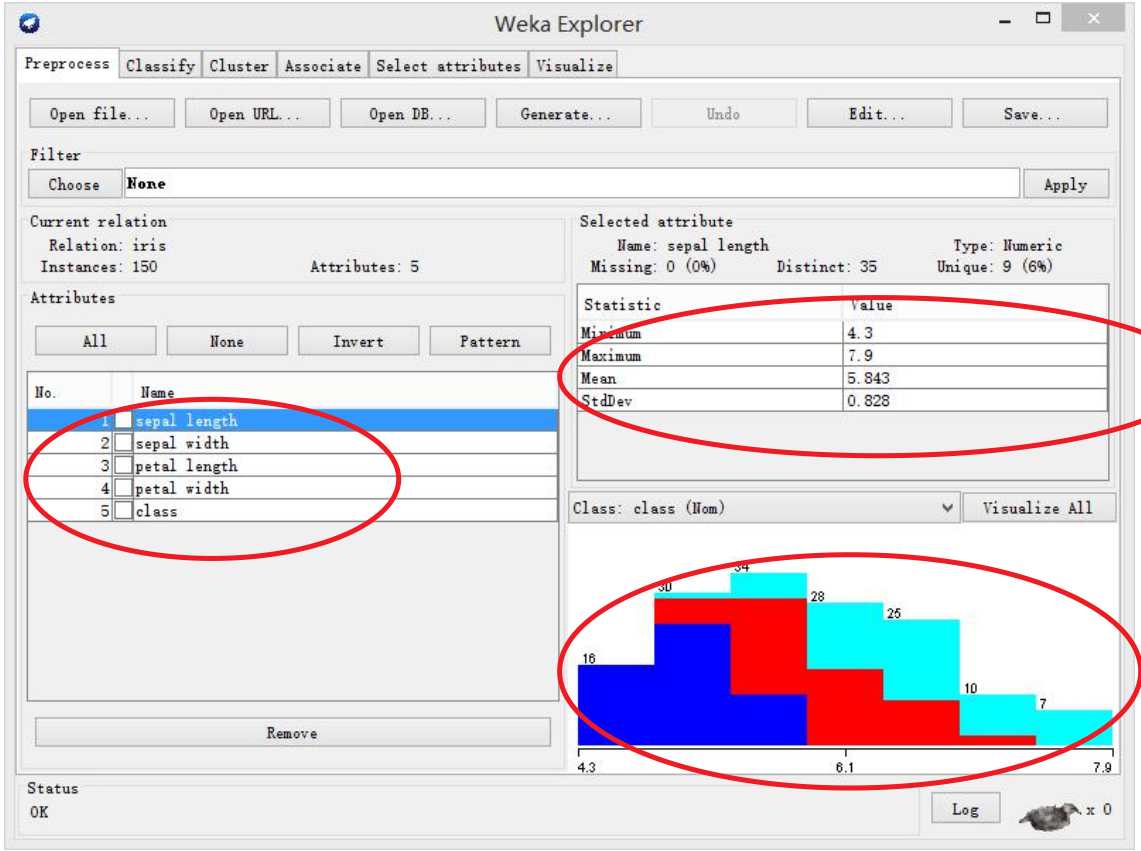
采用Weka分析软件学习分类算法

- Weka 导入Iris.arff文件：进行分类算法的第一步就是要导入我们要处理的数据集



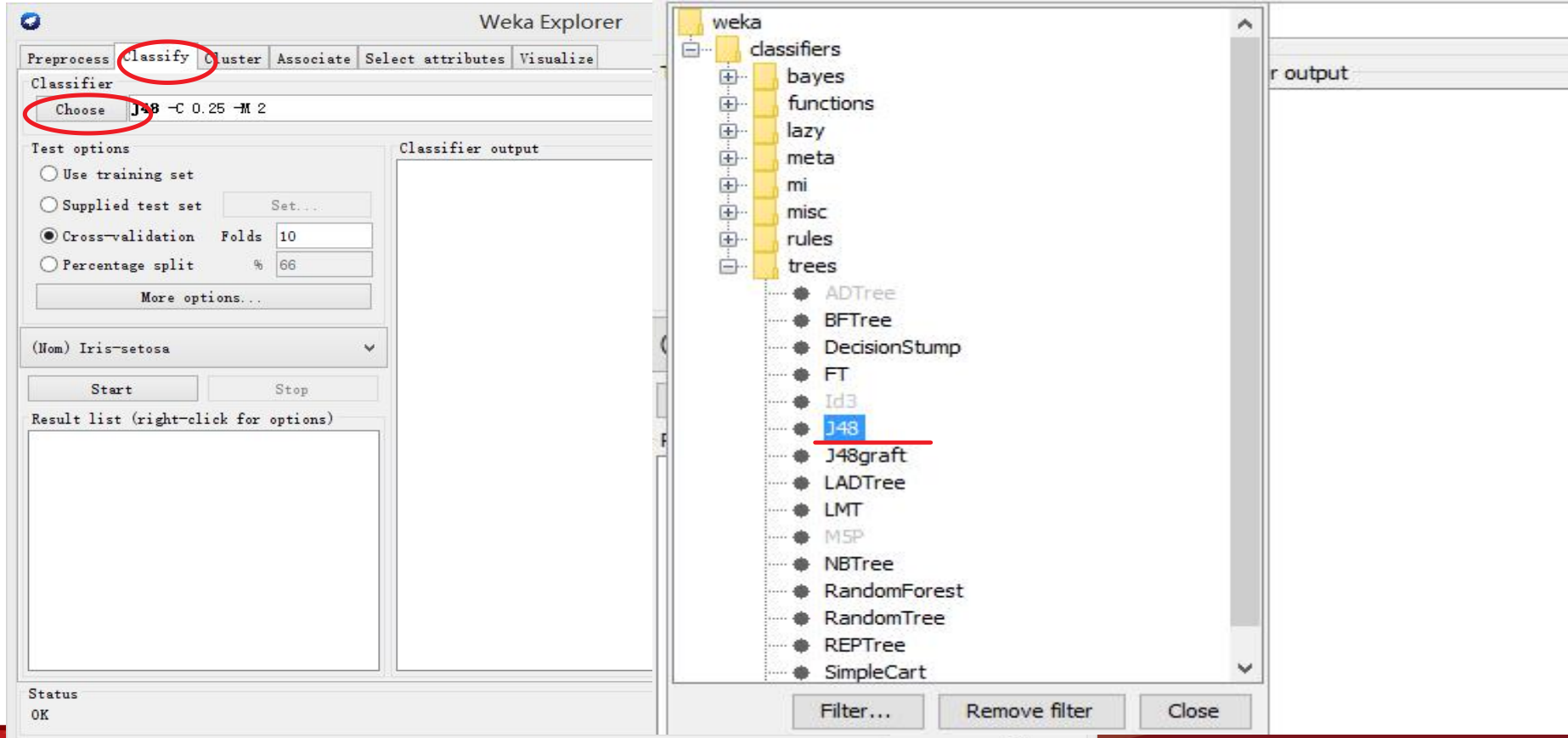
采用Weka分析软件学习分类算法

- 数据集特征显示

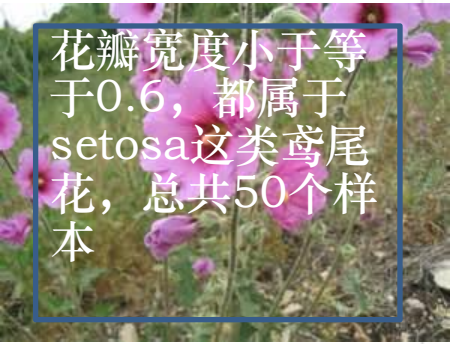


采用Weka分析软件学习分类算法

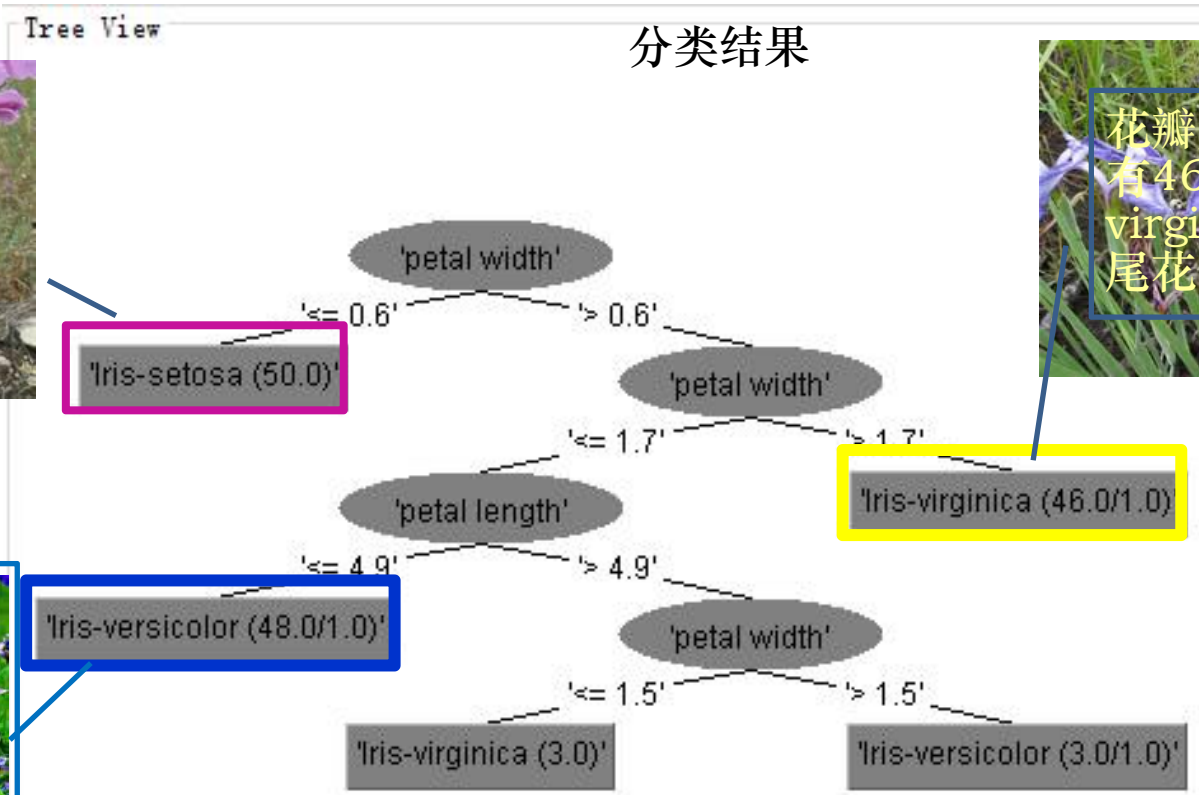
- Weka 分类算法

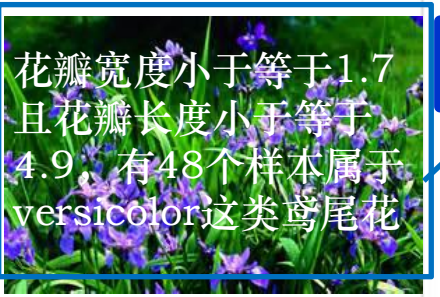


采用Weka分析软件学习分类算法



分类结果





谢 谢

Thank you for your attention!