

课程设计报告

**姓 名： 娄建鹏**

**学 号： 19851002**

**专 业： 计算机科学与技术**

**系 别： 计算机与信息技术**

**指导教师： 王志海**

**2022年6月**

# 1 数据分析网站

WeKa全名是怀卡托智能分析环境，非商业化的，基于Java环境下开源的[机器学习](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0/217599" \t "_blank)，是收集数据采集任务的机器学习算法。

1.1 什么是WeKa

WeKa全名是怀卡托智能分析环境，非商业化的，基于Java环境下开源的[机器学习](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0/217599)，是收集数据采集任务的机器学习算法。它含有用于数据准备、分类、回归、聚类、关联规则采集和可视化的工具。如果想自己实现数据挖掘算法的话，可以参考WeKa的接口文档。在WeKa中集成自己的算法甚至借鉴它的方法自己实现可视化工具并不是件很困难的事情。

WeKa仅被发现于新西兰的岛屿上，WeKa是不会飞的鸟类，他具有好奇的特性。这个名字的发音像这样，并且鸟儿的叫声听起来也像这样。

WeKa是在GNU通用公共许可证的发布下的开源软件。并且，我们放了一些免费课程在网上，教授使用WeKa用来机器学习和数据采集，这些视频课程可以在Youtube上找到。而且，WeKa支持深度学习。

2005年8月，在第11届ACM SIGKDD国际会议上，怀卡托大学的WeKa小组荣获了数据挖掘和知识探索领域的最高服务奖，WeKa系统得到了广泛的认可，被誉为数据挖掘和机器学习历史上的里程碑，是现今最完备的数据挖掘工具之一（已有11年的发展历史）。WeKa的每月下载次数已超过万次。

2014年3月起，[新西兰怀卡托大学](https://baike.baidu.com/item/%E6%96%B0%E8%A5%BF%E5%85%B0%E6%80%80%E5%8D%A1%E6%89%98%E5%A4%A7%E5%AD%A6/417707)将推出WeKa免费网课，课程分为初级和高级两个部分，每个部分时长5周。初级课程将于2014年3月3日开课，高级课程于2014年4月下旬开课。课程具体内容参见怀卡托大学网站WeKa MOOC。课程在优酷网站也有专辑。

WeKa的主要开发者同时恰好来自新西兰的[怀卡托大学](https://baike.baidu.com/item/%E6%80%80%E5%8D%A1%E6%89%98%E5%A4%A7%E5%AD%A6/7082322)。

1.2 WeKa包含的工具

**（1）数据格式**

跟很多[电子表格](https://baike.baidu.com/item/%E7%94%B5%E5%AD%90%E8%A1%A8%E6%A0%BC)或数据分析软件一样，WeKa所处理的数据集是一个二维的表格。这里我们要介绍一下WeKa中的术语。表格里的一个横行称作一个实例（Instance），相当于统计学中的一个样本，或者数据库中的一条记录。竖行称作一个属性（Attribute），相当于统计学中的一个变量，或者数据库中的一个字段。这样一个表格，或者叫数据集，在WeKa看来，呈现了属性之间的一种关系(Relation)。

Weka存储数据的格式是ARFF（Attribute-Relation File Format）文件，这是一种ASCII文本文件。[二维表格](https://baike.baidu.com/item/%E4%BA%8C%E7%BB%B4%E8%A1%A8%E6%A0%BC)存储在如下的ARFF文件中。这也就是WeKa自带的“weather.arff”文件，在WeKa[安装目录](https://baike.baidu.com/item/%E5%AE%89%E8%A3%85%E7%9B%AE%E5%BD%95)的“data”子目录下可以找到。

**（2）关联规则**

关联规则又称购物栏分析。目前，WeKa的[关联规则](https://baike.baidu.com/item/%E5%85%B3%E8%81%94%E8%A7%84%E5%88%99)分析功能仅能用来作示范，不适合用来挖掘大型数据集。

默认关联规则分析是用Apriori算法，我们就用这个算法，但是点“Choose”右边的文本框修改默认的参数，弹出的窗口中点“More”可以看到各参数的说明。

**（3）分类与回归**

Weka把分类(Classification)和回归(Regression)都放在“Classify”选项卡中，这是有原因的。

在这两个任务中，都有一个目标属性（输出变量）。我们希望根据一个样本(WeKa中称作实例)的一组特征（输入变量），对目标进行预测。为了实现这一目的，我们需要有一个训练数据集，这个数据集中每个实例的输入和输出都是已知的。观察训练集中的实例，可以建立起预测的模型。有了这个模型，我们就可以新的输出未知的实例进行预测了。衡量模型的好坏就在于预测的准确程度。

在WeKa中，待预测的目标（输出）被称作Class属性，这应该是来自分类任务的“类”。一般的，若Class属性是分类型时我们的任务才叫分类，Class属性是数值型时我们的任务叫回归。

**（4）聚类**

聚类分析中的“类”（cluster）和前面分类的“类”（class）是不同的，对cluster更加准确的翻译应该是“簇”。聚类的任务是把所有的实例分配到若干的簇，使得同一个簇的实例聚集在一个簇中心的周围，它们之间距离的比较近；而不同簇实例之间的距离比较远。对于由数值型属性刻画的实例来说，这个距离通常指欧氏距离。聚类分析，通常使用最常见的K均值（[K-means](https://baike.baidu.com/item/K-means)）算法。

**（5）数据准备**

使用WeKa作数据挖掘，面临的第一个问题往往是我们的数据不是ARFF格式的。幸好，WeKa还提供了对CSV文件的支持，而这种格式是被很多其他软件所支持的。此外，WeKa还提供了通过JDBC访问数据库的功能。

1.3 WeKa提供的4个功能环境

**（1）SimpleCLI**

提供了一个简单的[命令行](https://so.csdn.net/so/search?q=%E5%91%BD%E4%BB%A4%E8%A1%8C&spm=1001.2101.3001.7020)界面，从而可以在没有自带命令行的操作系统中直接执行WeKa命令。  
使用命令行有两个好处：一个是可以把模型保存下来，这样有新的待预测数据出现时，不用每次重新建模，直接应用保存好的模型即可。另一个是对预测结果给出了置信度，我们可以有选择的采纳预测结果，例如，只考虑那些置信度在85%以上的结果。  
**（2）Explorer**

使用WeKa探索数据的环境。在这个环境中，WeKa提供了数据的预处理，数据格式的转化（从CSV格式到ARFF格式的转化，详见第4部分），各种数据挖掘算法（包括分类与回归算法，聚类算法，关联规则等），并提供了结果的可视化工具。

对于一个数据集，通过简单的数据的预处理，并对数据挖掘算法进行选择（在WeKa3.5版本之后，加入了算法的过滤功能，可以过滤掉那些不适合当前数据集类型的算法），接着通过窗口界面对算法的参数进行配置，最后点击“Start”按钮就可以运行了。

可视化工具分为对数据集的可视化和对部分结果的可视化，并且我们可以通过属性选择工具(Select Attribute)，通过搜索数据集中所有属性的可能组合，找出预测效果最好的那一组属性。  
**（3）Experimenter**

运行算法试验、管理算法方案之间的统计检验的环境。Experiment环境可以让用户创建，运行，修改和分析算法试验，这也许比单独的分析各个算法更加方便。例如，用户可创建一次试验，在一系列数据集上运行多个算法（schemes），然后分析结果以判断是否某个算法比其他算法（在统计意义下）更好。

Explorermenter主要包括简单模式，复杂模式和远程模式。复杂模式是对简单模式的基本功能的扩充，而远程模式允许我们通过分布式的方法进行实验。  
就功能模块而言，分为设置模块，运行模块和分析模块。在设置模块中我们可以自定义实验，加入多个算法和多方的源数据（支持ARFF文件，CSV文件和数据库），在运行模块中我们可以运行我们的实验，而在分析模块中，我们可以分析各种算法的准确性，并提供了各种统计方法对结果进行检验比较。

值得一提的是，我们可以把实验的各种参数，包括算法，数据集等，保存以方便下一次相同实验的进行；也可以把各种算法保存，方便应用在不同的数据集上；如果数据集来源于数据库的话，实验在过程中可以中止并继续（原因可以是被中止或者是扩展了实验），这样就不用重新运行那些已实验过的算法/数据集祝贺，而仅计算还没有被实验的那些。  
**（4）KnowledgeFlow**

这个环境本质上和Explorer所支持的功能是一样的，但是它有一个可以拖放的界面。它有一个优势，就是支持增量学习（incremental learning）。KnowledgeFlow为WeKa提供了一个＂数据流＂形式的界面．用户可以从一个工具栏中选择组件，把它们放置在面板上并按一定的顺序连接起来，这样组成一个＂知识流＂（knowledge flow）来处理和分析数据．目前，所有的WeKa分类器（classifier）、筛选器（filter）、聚类器（clusterer）、载入器（loader）、保存器（saver），以及一些其他的功能可以在KnowledgeFlow中使用。

KnowledgeFlow可以使用增量模式（incrementally）或者批量模式（inbatches）来处理数据（Explorer只能使用批量模式）。当然对数据进行增量学习要求分类器能够根据各实例逐个逐个的更新．现在WeKa中有五个分类器能够增量地处理数据：NaiveBayesUpdateable，IB1，IBk，LWR（局部加权回归）。还有一个meta分类器RacedIncrementalLogitBoost可以使用任意基于回归的学习器来增量地学习离散的分类任务。

# 2 软件系统功能



# 3 数据集合格式

3.1数据格式表格

表1.1 Weather数据集合

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | outlook | temperature | humidity | windy | play |
| 1 | sunny | hot | high | FALSE | no |
| 2 | sunny | hot | high | TRUE | no |
| 3 | overcast | hot | high | FALSE | yes |
| 4 | rainy | mild | high | FALSE | yes |
| 5 | rainy | cool | normal | FALSE | yes |
| 6 | rainy | cool | normal | TRUE | no |
| 7 | overcast | cool | normal | TRUE | yes |
| 8 | sunny | mild | high | FALSE | no |
| 9 | sunny | cool | normal | FALSE | yes |
| 10 | rainy | mild | normal | FALSE | yes |
| 11 | sunny | mild | normal | TRUE | yes |
| 12 | overcast | mild | high | TRUE | yes |
| 13 | overcast | hot | normal | FALSE | yes |
| 14 | rainy | mild | high | TRUE | no |



**参考文献**

[1] 百度百科, URL: https://baike.baidu.com/item/weka/1070121. [2022-6-16].

[2] 博客, URL: https://blog.csdn.net/shaoz/article/details/6841498?. [2022-6-16].

。

参考文献以文献在整个论文中出现的次序用［1］、［2］、［3］……形式统一排序、依次列出。

参考文献的表示格式为：

期刊文献：［序号］主要责任者．文献题名［J］．刊名，年，卷（期）：引用部分起止页码．

专著、论文集、学位论文、报告： ［序号］ 主要责任者．文献题名［文献类型标识］．出版地：出版者，出版年. 引用部分起止页码．

论文集中的析出文献：［序号］主要责任者. 析出文献题名［A］. 原文献主要责任者(任选)．原文献题名［C］. 出版地：出版者，出版年. 析出文献起止页码.

专利：［序号］ 专利所有者. 专利题名［P］. 专利国别：专利号, 出版日期．

报纸文章：［序号］ 主要责任者．文献题名［N］． 报纸名，出版日期（版次）．

电子文献：［序号］ 主要责任者. 电子文献题名［电子文献及载体类型标识］.电子文献的出处或可获得地址，发表或更新日期/引用日期（任选）．

国际、国家标准：［序号］ 标准编号，标准名称［S］．