

课程设计报告

**姓 名：** 陈栩生

**学 号：** 19851261

**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**

# 你的标题

## 1 数据分析网站(Weka)

数据分析网站Weka由以下几部分组成：

### 1.1 项目（Project）

该部分介绍了Weka网站的项目目标和软件简介：

#### 1.1.1 项目目标

它们的项目目标是：

（1）使机器学习技术普遍可用;

（2）将它们应用于与新西兰工业相关的实际问题;

（3）开发新的机器学习算法并将其提供给世界;

（4）为该领域的理论框架做出贡献。

#### 1.1.2 软件

我们的团队已将多种标准ML技术整合到名为Weka的软件“工作台”中，用于怀卡托知识分析环境。有了它，特定领域的专家能够使用ML从太大而无法手动分析的数据库中获取有用的知识。Weka的用户是ML研究人员和工业科学家，但它也广泛用于教学。最近，我们的团队还研究了MOA，这是一种挖掘数据流的环境。

### 1.2 软件

该部分比较详细的介绍了Weka软件信息。

#### 1.2.1 Weka 3：Java中的机器学习软件

Weka是用于数据挖掘任务的机器学习算法的集合。它包含用于数据准备、分类、回归、聚类、关联规则挖掘和可视化的工具。

Weka只在新西兰的岛屿上发现，是一种不会飞的鸟，具有好奇的天性。名字的发音是这样的，鸟的声音是这样的。

Weka是根据GNU通用公共许可证发布的开源软件。

Weka支持深度学习!

#### 1.2.2 入门

提供了以下的链接：

（1）要求；

（2）下载；

（3）文档；

（4）FAQ;

（5）获取帮助

#### 1.2.3 更多信息

提供了以下的链接：

（1）引用Weka；

（2）数据集；

（3）相关项目；

（4）杂项代码;

（5）其他文献

#### 1.2.4 开发者

提供了以下的链接：

（1）发展；

（2）历史；

（3）颠覆；

（4）贡献者;

（5）商业许可证

### 1.3 书

该部分介绍了他们为机器学习提供了一系列令人兴奋的技术，其中包括用于分析数据和进行预测的实用工具，同时也为人工智能的最新进展提供动力。他们写了一本书，提供了对该领域的高度易懂的介绍，同时也迎合了想要深入研究现代概率建模和深度学习方法中可用的更多数学技术的读者。Chris Pal已加入lan Witten与Eibe Frank和Mark Hall的第四版，他在概率模型和深度学习方面的专业知识极大地扩展了本书的覆盖范围。为了给新材料腾出空间，我们现在提供了一个在线附录在Weka软件上。它是本书附录中对Weka的简要描述的扩展版本。本书继续提供对其描述的算法的Weka实现的参考。Weka MOOCs提供类似于第三版教程练习的活动。我们现在还提供有关其他软件的信息：自2011年我们编写第三版以来，机器学习的计算生态系统已经大幅增长。

### 1.4 培训班

该部分介绍了机器学习课程，他们整理了几门免费的在线课程，使用Weka教授机器学习和数据挖掘。课程视频可在Youtube上获得。这些课程托管在Future Learn平台上。

### 1.5 出版物

该部分介绍了他们出版过的所有书籍和课程。

### 1.6 人们

该部分介绍了参与该项目的学术人员、研究员、名誉研究员、博士生和硕士生。

### 1.7 有关的

该部分提供了其他网站的链接以及相关材料。

## 2 Weka软件系统功能

### 2.1 Weka介绍

Weka（Waikato Environment for Knowledge Analysis）是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、关联规则挖掘以及在新的交互式界面上的可视化的工具。它是由新西兰的The University of Waikato研发，是一款基于Java环境下开源的机器学习以及数据挖掘的软件套件。Weka是用Java写成的，它可以运行于几乎所有的操作平台,包括Linux，Windows等操作系统。用户可以使用Weka提供的可视化界面，也可以使用Weka提供的Api开发满足特定需求的模型和算法。同时Weka也支持加载开发者自己开发的扩展包，所以在Weka下几乎能找到了所有现在比较流行的数据挖掘模型和算法的实现。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。

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。

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。[1]

一般常用的有四个应用，分别是：

### 2.2 探索者（Explorer）

#### 2.2.1 探索者（Explorer）简单介绍

系统提供的最容易使用的图像用户接口。通过选择菜单和填写表单，可以调用Weka的所有功能。这是用来进行数据实验、挖掘的环境，它提供了分类，聚类，关联规则，特征选择，数据可视化等等功能。根据不同的功能把这个界面分成8个区域。

#### 2.2.2 探索者（Explorer）界面介绍

区域1的几个选项卡是用来切换不同的挖掘任务面板。主界面最左上角（标题栏下方）的是标签栏，分为五个部分，功能依次是：

（1）数据预处理(Preprocess)：选择和修改要处理的数据；

（2）分类(Classify)：训练和测试关于分类或回归的学习方案；

（3）聚类(Cluster)：从数据中学习聚类；

（4）关联(Associate)：从数据中学习关联规则；

（5）属性选择(Select attributes)：选择数据中最相关的属性；

（6）可视化(Visualize)：查看数据的交互式二维图像。

区域2是一些常用按钮。包括打开数据，保存及编辑功能。我们在这里把“bank-data.csv”另存为“bank-data.arff”。

在区域3中“Choose”某个“Filter”，可以实现筛选数据或者对数据进行某种变换。数据预处理主要就利用它来实现。

区域4展示了数据集的一些基本情况。

区域5中列出了数据集的所有属性。勾选一些属性并“Remove”就可以删除它们，删除后还可以利用区域2的“Undo”按钮找回。区域5上方的一排按钮是用来实现快速勾选的。

在区域5中选中某个属性，则区域6中有关于这个属性的摘要。注意对于数值属性和分类属性，摘要的方式是不一样的。图中显示的是对数值属性“income”的摘要。

区域7是区域5中选中属性的直方图。若数据集的最后一个属性（我们说过这是分类或回归任务的默认目标变量）是分类变量（这里的“pep”正好是），直方图中的每个长方形就会按照该变量的比例分成不同颜色的段。要想换个分段的依据，在区域7上方的下拉框中选个不同的分类属性就可以了。下拉框里选上“No Class”或者一个数值属性会变成黑白的直方图。

区域8是状态栏，可以查看Log以判断是否有错。右边的Weka鸟在动的话说明Weka正在执行挖掘任务。右键点击状态栏还可以执行Java内存的垃圾回收。[3]

### 2.3 实验者（Experimenter）

#### 2.3.1 实验者（Experimenter）简单介绍

用于帮助用户解答实际应用分类和回归技术中遇到的一个基本问题——对于一个已知问题，哪种方法及参数值能够取得最佳效果？通过Weka提供的实验者工作环境，用户可以比较不同的学习方案。尽管探索者界面也能通过交互完成这样的功能，但通过实验者界面，用户可以让处理过程实现自动化。实验者界面更加容易使用不同参数去设置分类器和过滤器，使之运行在一组数据集中，收集性能统计数据，实现重要的测试实验。

Experimenter完成的功能有如下几个：

（1）支持多个算法对多个数据集进行运算;

（2）支持分布式运算。

Explorer应用虽然也支持多个算法同时运行，但是那都是单个算法对单个数据集（除非人工合并成一个），而且不支持分布式运算。

Experimenter的分布式运算并不是算法层次的分布式运算，而是将不同数据集分配到不同的服务器进行的，所以有以下几点需要注意：

（1）服务器数量不能超过数据集的数量；

（2）每个服务器运行的试验都是相互独立互不干扰的，只有最终结果才会汇总到一个服务器上;

（3）服务器之间的远程调用时通过RMI进行的。[4]

以下是实验者（Experimenter）的功能界面介绍：

#### 2.3.2 simple模式的界面介绍

区域1：菜单部分，setup是设置运行参数，Run是设置完后开始运行，analyze。

区域2：功能选择，重点看一下画圈部分，这里分为两个功能模块，一个是simple，一个是advanced，前者是低配版，后者是高配版（功能齐全）。

区域3：这个是结果输出文件，选择一个文件，运行结束会将结果打印到该输出文件中，先选择左边的文件类型，然后点击Browse可以选择输出文件。

区域4：试验设置，重点看画圈部分，这里共有三种设置：

（1）Cross-validation：交叉验证，下面选择折数；

（2）Train/Test Percentage Split(data randomized)：切分训练数据集和测试数据集，切分的时候先打乱顺序，然后按照百分比进行选择数据；

（3）Train/Test Percentage Split(order preserved)：不会打乱顺序，按照指定百分比进行切分（可以事先合并好数据集，然后按照这个进行切分）。

区域5：迭代控制，设置算法迭代次数，重点看我画圈部分，数据集优先还是算法优先，因为数据集和算法是多对多进行的，如果选择数据集优先，会先使得每个算法轮流测试完所有数据集，然后再进行下一个算法，如果是算法优先，会针对一个数据集，先把所有算法轮流一遍，再换下一个数据集。

区域6：添加数据集，可以添加多个。

区域7：算法添加，可以添加多个，值得一提的是，下面有load和save一说，因为添加算法时要设置算法参数，如果实验次数多了，就要每次都要设置，很麻烦，这时就可以保存算法和算法参数，下次直接加载保存的文件即可。

需要注意几点：

（1）每次运行算法都需要点击区域2的new才能进行配置参数，new的意思是开始一次实验；

（2）在区域3中选择输出文本的时候，最好选择.csv格式，使用Excel打开，查看的时候比较容易看，列名与数据一一对应。

#### 2.3.3 advanced模式的界面介绍

区域1：这个是选择结果输出设置的，类似于simple版本的区域2，这里要先点击choose，选择一个模式，然后点击空白处设置输出参数，一共有是三个模式，分别对应.csv文件、数据库输出以及.ARFF输出。

区域2：这个是结果生成器，和simple模式中的区域4一致，即确定结果输出方式以及数据集的切分——训练集和测试集如何生成，但是功能更加细致，点击choose，就可以看到生成器的种类：

（1）Averaging Result Producer：平均化结果生成器，这个结果生成器取了一组试验(通常是cross-validation试验)运行的平均值；

（2）Cross Validation Result Producer：交叉验证结果生成器；

（3）Cross Validation Split Result Producer：交叉验证切分结果生成器；

（4）Database Result Producer：数据库结果生成器，将实验结果反馈到数据库中；

（5）Explicit Test Result Producer：加载外部测试集结果生成器，测试集使用外部加载的；

（6）Learning Rate Result Producer：学习率结果生成器，用来生成学习曲线的，一般要与交叉验证或者平均化结果生成器结合使用（其内部参数可以设置）；

（7）Random Split Result Producer：随机划分训练集和测试集的结果生成器。

区域3：这个是运行代数。

区域4：这个是布置远程试验用的，也就是将计算负载分布于多台计算机上。

区域5：选择添加算法，添加算法之前要先将我画圈的部分更改为enable，这样才能进行添加配置。

#### 2.3.4 analyze的界面介绍

区域1：该区域是source区，也可以称作是数据加载区，加载方式分为三种：

（1）本地文件加载；

（2）读取数据库里面的数据；

（3）加载算法运行结果。

需要注意的是，第三种加载模式的前提是需要先进行试验（也就是先setup，然后Run成功之后）。

数据加载是一切的基础，只有加载了数据，才能进行分析。

区域2：Action区，也就是操作区，能对数据进行三种操作：

（1）Perform test：进行性能测试；

（2）保存输出；

（3）打开Explorer界面，对数据进行操作。

区域3：这个是配置区，进行各种配置。

区域4：每进行一次执行操作，结果列表就会出现一行内容，通过点击内容可以在右边的测试输出框显示出对应的内容。

区域5：测试输出框。[4]

### 2.4 知识流（Knowledge Flow）

使用增量（分批）方式的算法来处理大型数据集，用户可以定制处理数据流的方式和顺序。按照一定顺序将代表数据源、预处理工具、学习算法、评估手段和可视化模块的各组件组合在一起，形成数据流。

### 2.5 简单命令行（Simple CLI）

该界面用于和用户进行交互，可以直接执行Weka命令。[2]

## 3 数据集合的定义

Weka的数据格式跟很多电子表格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。

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）。表1-1中一共有14个实例，5个属性，关系名称为“weather”。

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安装目录的“data”子目录下可以找到。需要注意的是，在Windows记事本打开这个文件时，可能会因为回车符定义不一致而导致分行不正常。推荐使用Ultra Edit这样的字符编辑软件察看ARFF文件的内容。

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 |

如何将其他文件转换为arff文件？

data文件需要先将后缀改成csv文件，打开csv文件后在首行插入字段名后保存，字段名可以在官网或names文件中查找。最后将csv文件在Weka中打开并另存为arff文件。例子：先在https://archive-beta.ics.uci.edu/ml/datasets/toxicity-2下载数据集Toxicity-13F.csv。将Toxicity-13F.csv在Weka中打开另存为Toxicity-13F.arff。

识别ARFF文件的重要依据是分行，因此不能在这种文件里随意的断行。空行（或全是空格的行）将被忽略。以“%”开始的行是注释，Weka将忽略这些行。如果你看到的“weather.arff”文件多了或少了些“%”开始的行，是没有影响的。

除去注释后，整个ARFF文件可以分为两个部分。第一部分给出了头信息（Head information），包括了对关系的声明和对属性的声明。第二部分给出了数据信息（Data information），即数据集中给出的数据。从“@data”标记开始，后面的就是数据信息了。

关系声明：关系名称在ARFF文件的第一个有效行来定义，格式为“@relation”是一个字符串。如果这个字符串包含空格，它必须加上引号（指英文标点的单引号或双引号）。

属性声明：属性声明用一列以“@attribute”开头的语句表示。数据集中的每一个属性都有它对应的“@attribute”语句，来定义它的属性名称和数据类型。

这些声明语句的顺序很重要。首先它表明了该项属性在数据部分的位置。例如，“humidity”是第三个被声明的属性，这说明数据部分那些被逗号 分开的列中，第三列数据 85 90 86 96 ... 是相应的“humidity”值。其次，最后一个声明的属性被称作class属性，在分类或回归任务中，它是默认的目标变量。

Weka支持的有四种，分别是数值型（numeric）、分类型（nominal）、字符串型（string）、日期和时间型（date []）还可以使用两个类型“integer”和“real”，但是Weka把它们都当作“numeric”看待。注意 “integer”，“real”，“numeric”，“date”，“string”这些关键字是区分大小写的，而 “relation”“attribute ”和“data”则不区分。

数值型属性可以是整数或者实数，但Weka把它们都当作实数看待；分类属性由列出一系列可能的类别名称并放在花括号中：{, , , ...} 。数据集中该属性的值只能是其中一种类别。字符串属性中可以包含任意的文本。这种类型的属性在文本挖掘中非常有用。日期和时间属性统一用“date”类型表示，它的格式是“@attribute date []”其中是这个属性的名称，是一个字符串，来规定该怎样解析和显示日期或时间的格式，默认的字符串是ISO-8601所给的日期时间组合格式“yyyy-MM-ddTHH:mm:ss”。数据信息部分表达日期的字符串必须符合声明中规定的格式要求。

## 参考文献

[1] 百度百科, URL: https://baike.baidu.com/item/weka/10701215?fr=aladdin. [2022-6-16].

[2] CSDN博主, URL: https://blog.csdn.net/sinat\_25873421/article/details/82724903. [2022-6-16].

[3] CSDN博主, URL: http://blog.csdn.net/u011067360/article/details/20844443. [2022-6-16].

[4] CSDN博主, URL: https://blog.csdn.net/yuangan1529/article/details/115236767. [2022-6-17].