

课程设计报告

**姓 名：** 雷云凤

**学 号：** 19851047

**专 业：**计算机科学与技术

**系 别：**计算机科学与技术

**指导教师：** 王志海

**2022年6月**

# 《信息系统集成与开发》课程设计报告

Weka是用于数据挖掘认为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。这个名字的发音是这样的，鸟儿听起来像这样。Weka是根据GNU通用公共许可证发布的开源软件。

## 1 Weka网站简介

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（与之对应的是SPSS公司商业数据挖掘产品--Clementine ）的，基于Java环境下开源的机器学习（machine learning）以及数据挖掘（data mining）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也是新西兰独有的一种鸟名（[新西兰秧鸡](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=574284&ss_c=ssc.citiao.link" \t "https://baike.sogou.com/_blank)），而Weka的主要开发者同时恰好来自新西兰的[怀卡托大学](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=485150&ss_c=ssc.citiao.link" \t "https://baike.sogou.com/_blank)（The University of Waikato）[1]。

Weka的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），同时 Weka也是新西兰的一种鸟名，而 Weka的主要开发者来自新西兰。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](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=138836896&ss_c=ssc.citiao.link)，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。

如果想自己实现[数据挖掘算法](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=68975778&ss_c=ssc.citiao.link)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。

2014年3月起，新西兰怀卡托大学将推出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[2]。

Weka 是怀卡托智能分析环境(Waikato Envi ronment for Knowledge Analysis)的英文字首缩写，Weka的发音类似新西兰本土一种不会飞的鸟，因此Weka系统使用该鸟作为其徽标。

## 2 Weka系统功能

Weka的主页面窗口有四个模块：

### 2.1 Explorer

Explorer(探索者)界面，是Weka的主要图形化用户界面，其全部功能都可通过菜单选择或表单填写进行访问,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，口语进行不同的实验对比不同算法的结果。Experimentor用来进行实验，对不同学习方案进行数据测试的环境。Knowledge Flow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。Simple CLI简单的命令行界面。

### 2.2 Experimenter

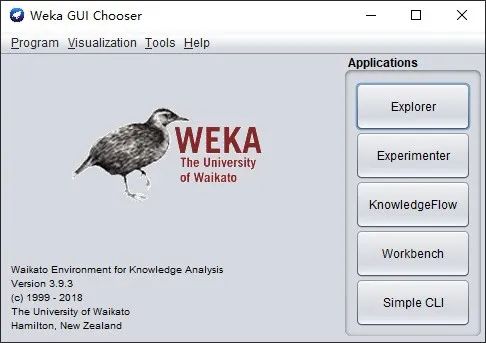
Experimentor(实验者)用来进行实验，对不同学习方案进行数据测试的环境。帮助用户解答实际应用分类和回归技术中遇到的一个基本问题。

### 2.3 KnowledegFlow

KnowledgeFlow(知识流)通过拖拽的方式建立实验方案，与Explorer相似。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结合在一起，形成数据流。

### 2.4 Simple CLI

Simple CLI（简单命令行）是为不提供自己的命令行界面的操作系统提供的，该界面用于和用户进行交互，可以直接执行Weka命令。



## 3 数据集合格式

Weka支持很多文件格式，包括arff、xrff、csv，甚至有libsvm的格式。其中arff是最常见的格式，arff是最常见的格式，我们在这里仅介绍这一种。Arff全称是Attribute-Relation File Format，以下是一个arff格式的文件的例子。

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。

文件中，”%”开头的是注释。剩余的可以分为两大部分，头信息(header information)的数据信息(data information)。

头信息中，”@relation”开头的行代表关系名称，在整个文件的第一行(除去注释)。格式是@relation;”@attribute”开头的代表特征，格式是@attribute；”@data”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。

Weka所处理的数据集是一个二维的表格。

### 3.1 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)。

示例表格一共有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。

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。

### 3.2 声明

#### 3.2.1 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 <relation-name>

<relation-name>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](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=10799706&ss_c=ssc.citiao.link" \t "https://baike.sogou.com/_blank)或[双引号](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=342060&ss_c=ssc.citiao.link)）。

#### 3.2.2 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](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=74633135&ss_c=ssc.citiao.link)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。

属性声明的格式为

@attribute <attribute-name> <datatype>

其中<attribute-name>是必须以字母开头的字符串。和关系名称一样，如果这个字符串包含空格，它必须加上引号。

WEKA支持的<datatype>有四种，分别是

numeric-------------------------数值型

<nominal-specification>-----分类（nominal）型

string----------------------------字符串型

date [<date-format>]--------日期和时间型

其中<nominal-specification> 和<date-format> 将在下面说明。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。

### 3.3 属性

#### 3.3.1 数值属性

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。

#### 3.3.2 分类属性

分类属性由<nominal-specification>列出一系列可能的类别名称并放在花括号中：{<nominal- name1>, <nominal-name2>, <nominal-name3>, ...} 。数据集中该属性的值只能是其中一种类别。

例如如下的属性声明说明“outlook”属性有三种类别：“sunny”，“ overcast”和“rainy”。而数据集中每个实例对应的“outlook”值必是这三者之一。

@attribute outlook {sunny, overcast, rainy}

如果类别名称带有空格，仍需要将之放入引号中。

#### 3.3.3 字符串属性

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。

示例：

@ATTRIBUTE LCC string

#### 3.3.4 日期和时间属性

日期和时间属性统一用“date”类型表示，它的格式是

@attribute <name> date [<date-format>]

其中<name>是这个属性的名称，<date-format>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。

#### 3.3.5 Relational型属性

在WEKA 3.5版中增加了一种属性类型叫做Relational，有了这种类型我们可以像[关系型数据库](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=277136&ss_c=ssc.citiao.link" \t "https://baike.sogou.com/_blank)那样处理多个维度了。但是这种类型目前还不见广泛应用，暂不作介绍。

### 3.4 数据

#### 3.4.1 数据信息

数据信息中“@data”标记独占一行，剩下的是各个实例的数据。

每个实例占一行。实例的各属性值用逗号“,”隔开。如果某个属性的值是缺失值（missing value），用问号“?”表示，且这个问号不能省略。例如：

@data

sunny,85,85,FALSE,no

?,78,90,?,yes

字符串属性和分类属性的值是区分大小写的。若值中含有空格，必须被引号括起来。例如：

@relation LCCvsLCSH

@attribute LCC string

@attribute LCSH string

@data

AG5, 'Encyclopedias and dictionaries.;Twentieth century.'

AS262, 'Science -- Soviet Union -- History.'

日期属性的值必须与属性声明中给定的相一致。例如：

@RELATION Timestamps

@ATTRIBUTE timestamp DATE "yyyy-MM-dd HH:mm:ss"

@DATA

"2001-04-03 12:12:12"

"2001-05-03 12:59:55"

#### 3.4.2 稀疏数据

有的时候数据集中含有大量的0值（比如购物篮分析），这个时候用稀疏格式的数据存贮更加省空间。

稀疏格式是针对数据信息中某个实例的表示而言，不需要修改ARFF文件的其它部分。看如下的数据：

@data

0, X, 0, Y, "class A"

0, 0, W, 0, "class B"

用稀疏格式表达的话就是

@data

{1 X, 3 Y, 4 "class A"}

{2 W, 4 "class B"}

每个实例用花括号括起来。实例中每一个非0的属性值用<index> <空格> <value>表示。<index>是属性的序号，从0开始计；<value>是属性值。属性值之间仍用逗号隔开。 这里每个实例的数值必须按属性的顺序来写，如 {1 X, 3 Y, 4 "class A"}，不能写成{3 Y, 1 X, 4 "class A"}。

注意在稀疏格式中没有注明的属性值不是缺失值，而是0值。若要表示缺失值必须显式的用问号表示出来。

### 3.5 csv转换为arrf格式

我们接触到的数据大多是csv格式的，如果想要将数据更改为arrf格式，我们可以借助Weka工具将csv格式转换为arrf格式。

#### 3.5.1 csv格式

逗号分隔值（Comma-Separated Values，CSV，有时也称为字符分隔值，因为分隔字符也可以不是逗号），其文件以纯文本形式存储表格数据（数字和文本）。纯文本意味着该文件是一个字符序列，不含必须像二进制数字那样被解读的数据。CSV文件由任意数目的记录组成，记录间以某种换行符分隔；每条记录由字段组成，字段间的分隔符是其它字符或字符串，最常见的是逗号或制表符。通常，所有记录都有完全相同的字段序列。通常都是纯文本文件。建议使用Word或是记事本来开启，再则先另存新档后用Excel开启，也是方法之一[4]。

#### 3.5.2 arff格式

arff是一种Weka专用的文件格式，由Andrew Donkin创立，有传言说arff代表Andrew's Ridiculous File Format(安德鲁的荒唐文件格式)，但在Weka的正式文档中明确说明arff代表Attribute-Relation File Format(属性——关系文件格式)。该文件是ASCII文本文件，描述共享一组属性结构的实例列表，由独立且无序的实例组成，是Weka表示数据集的标准方法，arff不涉及实例之间的关系。

#### 3.5.3 获取csv文件

本次案例我们将使用UCI网站中的AIR Quality数据进行演示，首先我们从网站中把我们需要的数据下载下来，进入网站之后选择Data Folder进入下载页面，之后我们选择AirQualityUCI.zip压缩包进行下载，下载之后将压缩包里的内容解压出来，解压出来之后会发现压缩包了有两个文件夹，一个是csv格式的一个是xlsx格式的。该文件属性有15个，实例有9358条，为了方便展示，我们选择6条属性和10条实例作为示范。

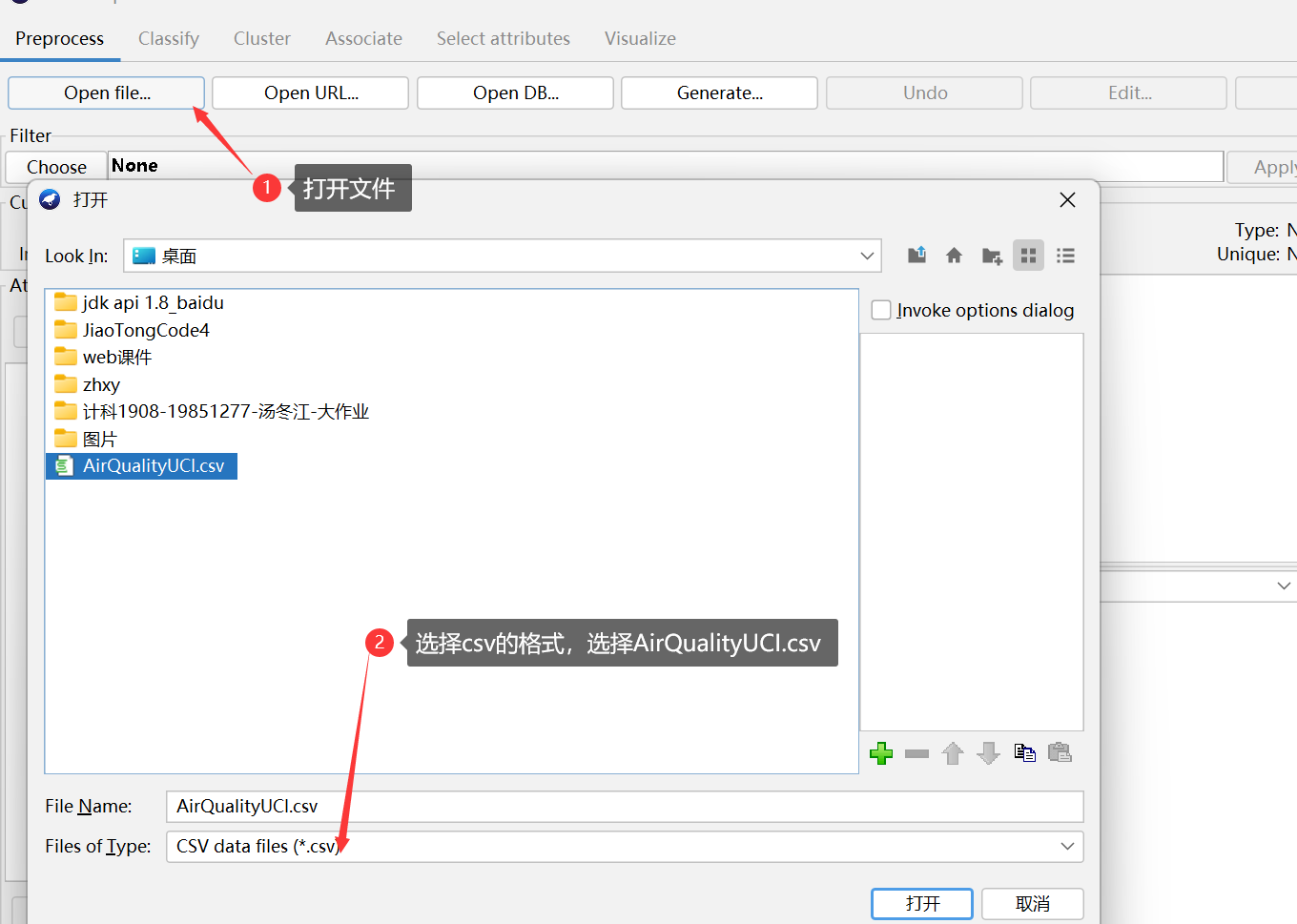
表3-1 AirQuality数据集合

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 序号 | Data | Time | CO(GT) | PT08.S1(CO) | NMHC(GT) | C6H6(GT) |
| 1 | 2004/3/10 | 18:00:00 | 2.6 | 1360 | 150 | 11.9 |
| 2 | 2004/3/10 | 19:00:00 | 2 | 1292 | 112 | 9.4 |
| 3 | 2004/3/10 | 20:00:00 | 2.2 | 1402 | 88 | 9.0 |
| 4 | 2004/3/10 | 21:00:00 | 2.2 | 1376 | 80 | 9.2 |
| 5 | 2004/3/10 | 22:00:00 | 1.6 | 1272 | 51 | 6.5 |
| 6 | 2004/3/10 | 23:00:00 | 1.2 | 1197 | 38 | 4.7 |
| 7 | 2004/3/11 | 0:00:00 | 1.2 | 1185 | 31 | 3.6 |
| 8 | 2004/3/11 | 1:00:00 | 1 | 1136 | 31 | 3.3 |
| 9 | 2004/3/11 | 2:00:00 | 0.9 | 1094 | 24 | 2.3 |
| 10 | 2004/3/11 | 3:00:00 | 0.6 | 1010 | 19 | 1.7 |

#### 3.5.3 转换csv格式为arff格式

1. 打开Weka，选择explorer模式

2. 这时会打开新的explorer窗口，我们选择Open File选项，可以对我们的数据进行选择。



1. 打开文件后，我们点击save选择文件类型为arff，选择文件名称以及存放位置后，点击保存，这样我们的数据就保存为了arff格式，这样我们就将csv文件转换为了arff格式。

## 4、美国金县二手房价的分析与预测

影响房价的因素，考虑到短期的货币政策、中期的土地供应和长期的人口统计影响商品房价格的主要宏观因素是：国家宏观政策、区域经济发展水平和监管政策市之间的房价。与宏观环境因素相比，中观和微观环境因素，这些因素对房价有更直接的影响。中观因素，例如与地铁站的距离，与公园的距离，如果房子位于学区。微观因素，如房子的年龄，房子是否有电梯，房子的布局等，都会影响房子的价格。土地的价值在于其位置。例如，有生活资源、商业资源和景观资源。生物资源是一个基本的前提，围绕应该至少有一个吃饭的地方，一个超市，一个餐馆和一辆地铁巴士。商业资源比生物资源高一个层次，景观资源是最高层次。然而，如果豪华别墅位于远离城市的地方，也是没有价值,近几年来计算机网络技术高速发展，网络信息量呈指数级增长。二手房价的分析和预测可以形成一种个性化网络信息采集与处理系统。将该系统应用微博用户兴趣群体发现信息，能有效提二手房价的分析和预测数据可视化服务系统的满意程度，实为一举两得。

4.1 数据理解

4.1.1 数据来源

美国金县二手房价的数据集选自datafountain网站，一共包含21614个二手房价的信息，这些都是美国金县2014年5月至2015年5月美国真实的二手房屋销售价格以及房屋的基本信息。这个站点的宗旨是为学术界和有关行业的爱好者提供一个组织机器学习竞赛，数据库编写和代码分享的平台。在平台上对问题进行说明，并为对应的问题寻找有竞争力的解决办法。参与者可以在平台上下载相关的资料，利用机器学习、深度学习等相关知识，构建相应的数学模型，并对问题进行求解，并给出结论。

数据集下载网址为<https://www.datafountain.cn/datasets/67>。

4.2 数据属性

数据主要包括2014年5月至2015年5月美国金县的房屋销售价格以及房屋的基本信息。数据总共包括21416条记录，

# 课程要求与注意事项

（1）改名：你的腾讯会议系统用户名必须包括学号，建议改为：班级+姓名+学号。例如，计科1904刘紫琦19851078。

（2）确信自己加入我们课程的微信群（任课教师按照不同授课班，共有3个不同的群，不要加错群）。

（3）课程设计报告要求严格遵循格式规定（这是评定成绩的依据之一）。

（4）建立人工智能应用系统：机器学习应用系统与大数据分析。

（5）2022年6月24日13:00PM提交课程报告给各班班长。

（6）阅读Weka系统网站（或者其他中文网站），并撰写网站简介。注意。必须有参考文献著录与引用[2]。

（7）安装Weka系统，并尝试运行。

（9）主要考核依据：一个课程设计报告

1）格式正确；

2）字数：25-35页：

3）

（10）次要考核手段：

1）课堂活跃程度（腾讯会议聊天）；

2）成员登记表（腾讯会议系统导出）；

3）7

（11）补充作业：写一个程序，从一个纯文本文件顺序读取，并按照学号进行记数。

为保障校外实践教学工作顺利进行，现将学生校外实践过程中应承担的安全责任明确如下：

## 参考文献

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

[2] Weka官网, URL: Weka 3 - Data Mining with Open Source Machine Learning Software in Java (waikato.ac.nz). [2022-6-16]

[3] 简书,URL: https://www.jianshu.com/p/73c6fce1dbe9. [2022-6-16].

[4] Weka官方文档, URL: https://waikato.github.io/weka-wiki/documentation/. [2022-6-16].