

课程设计报告

**姓 名：**周芳

**学 号：**19851298

**专 业：**计算机科学与技术

**系 别：**计算机信息技术

**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是根据GNU通用公共许可证发布的开源软件。

## 1 WeKa网站

Weka的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），是一款基于Java环境下开源的机器学习（machine learning）以及数据挖掘（data mining）软件。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.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0%E7%AE%97%E6%B3%95/18635836" \t "https://baike.baidu.com/item/weka/_blank)，包括对数据进行预处理，分类，回归、聚类、[关联规则](https://baike.baidu.com/item/%E5%85%B3%E8%81%94%E8%A7%84%E5%88%99" \t "https://baike.baidu.com/item/weka/_blank)以及在新的交互式界面上的可视化[1]。

Weka附带内置帮助，并包括一个全面的手册。有关Weka中实现的机器学习技术以及软件本身的介绍，请查看《数据挖掘：实用机器学习工具和技术》一书及其在Weka工作台上免费提供的在线附录，其中概述了该软件。与本书密切相关的是，在Weka中也有关于使用机器学习技术进行数据挖掘的免费在线课程。

Weka可以从其他几个软件系统中用于数据科学，并且在科学计算生态系统中有一组关于WEKA的幻灯片，涵盖Octave / Matlab，R，Python和Hadoop。一个页面，其中包含有关Weka支持导入PMML模型的新闻和文档。关于使用JDBC驱动程序将Weka连接到MongoDB的简短教程。

当下载特定版本的Weka源代码时，它包含在发行版中，包含在一个名为 的.jar文件（.zip文件的一种形式）中。但是，也可以直接从 Weka 的 [Subversion 源代码存储库读取源代码](https://waikato.github.io/weka-wiki/subversion/)，并且还可以[基于 Web 访问该存储库](https://svn.cms.waikato.ac.nz/svn/weka/)。

如果使用 Git 而不是 Subversion 进行软件开发，那么 [Weka 3.8 的 Subversion 存储库分支](https://github.com/Waikato/weka-3.8)有一个 Git 镜像，[Weka 的主开发主干](https://github.com/Waikato/weka-trunk)也有相应的镜像。请注意，这些镜像是只读的，请使用 Subversion 将更改提交到软件，而不是 Git。

核心Weka发行版包括来自[MTJ项目](https://github.com/fommil/matrix-toolkits-java)的第三方库代码，用于Java中的快速矩阵代数，用于生成解析器的[Java CUP项目](http://www2.cs.tum.edu/projects/cup/index.php)，[Bounce项目中](http://www.edankert.com/bounce/index.html)的身份验证对话框以及[Apache Commons Compress](http://commons.apache.org/proper/commons-compress/)库。有关详细信息，请参阅源代码存储库的 lib 文件夹。

Weka的包管理器提供了对大量可选库的访问，其中许多是由其他机构的开发人员贡献的。有关这些包的作者以及这些 Weka 包中使用的第三方库的信息，请查阅 Javadoc 以获取相关包和相应的包 lib 文件夹[2]。

## 2 weka软件系统功能

Weka是怀卡托智能分析环境(Waikato Environment for Knowledge Analysis)的英文字首缩写，新西兰怀卡托大学用Java开发的数据挖掘著名开源软件。功能有数据处理、特征选择、分类、回归、可视化等，支持多种数据文件格式，如arff、xrff、csv等，主流的数据格式是csv和arff。

Weka的主页面窗口有四个模块：

1. Explorer：进行数据的特征选择、分类、回归、聚类、关联规则、[数据可视化](https://cloud.tencent.com/product/yuntu?from=10680" \t "https://cloud.tencent.com/developer/article/_blank)等功能，口语进行不同的实验对比不同算法的结果；
2. Experimenter：使用增量式的算法处理大型数据集，对不同学习方案进行数据测试；
3. KnowledgeFlow：通过拖拽的方式建立实验方案，与Explorer相似；
4. Simple CLI：命令行界面，用于和用户进行交互。

### 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

Experimenter(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]。

## 3 数据集合的定义

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。

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）。

### 3.2 声明

声明分为关系声明和属性声明

#### 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或双引号）。

#### 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" \t "https://baike.sogou.com/_blank)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

属性大致分为五类

#### 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Weka3.5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

一般的数据分为数据信息和稀疏数据

#### 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arff格式

我们通常接触到的数据为csv格式较多，若想要将数据更改为Weka通用的arff格式，我们可以利用Weka工具进行格式的转换。

#### 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。

#### 3.5.2 获取csv文件

本次案例我们使用UCI网站中的Soybean(Large)数据进行演示，我们先从网站中下载下来我们所需要的数据，进入网站后选择Download按钮进入下载页面，我们选择soybean-large.data数据进行下载下来发现是一个纯文本文件，用记事本打开，发现此文件大约有300多条实例，属性多达36个，为了方便用表展示，我们只截取了前7个，实例也仅选取了前10个以作示范。

将网站主页中的属性名（features）依次复制到该data文件中，属性之间用逗号隔开，这样我们就构造出来了csv格式的数据文件。保存文件，将该文件的后缀改为csv，我们就得到这个数据集的csv格式文件。

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 | FALSE | yes |
| 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/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] 博客, URL: http://t.zoukankan.com/chamie-p-4551514.html [2022-6-16].

[4] 博客, URL: https://baike.baidu.com/item/CSV/10739 [2022-6-18].