

课程设计报告

**姓 名：** 马雯

**学 号：** 19851048

**专 业：**计算机科学与技术

**系 别：**计算机与信息技术学院

**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介绍

## 1.1 weka 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](https://baike.baidu.com/item/%E5%85%8D%E8%B4%B9/131326)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](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0/217599)（machine learning）以及[数据挖掘](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98/216477)（data mining）软件。它和它的[源代码](https://baike.baidu.com/item/%E6%BA%90%E4%BB%A3%E7%A0%81/3969)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.baidu.com/item/%E6%96%B0%E8%A5%BF%E5%85%B0%E7%A7%A7%E9%B8%A1/5991395)），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.baidu.com/item/%E6%80%80%E5%8D%A1%E6%89%98%E5%A4%A7%E5%AD%A6/7082322)（The University of Waikato）。

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.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0%E7%AE%97%E6%B3%95/18635836)，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)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](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。

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。

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。

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。

我们打算对前面的“bank-data”数据作关联规则的分析。用“Explorer”打开“bank-data-final.arff”后，切 换到“Associate”选项卡。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。

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)，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)以及在新的交互式界面上的可视化。Weka是根据GNU通用公共许可证发布的开源软件[1]。

[Weka](https://so.csdn.net/so/search?q=weka&spm=1001.2101.3001.7020)主界面各个功能键概述:

Explorer（探索者）：最容易使用的图形用户界面，通过菜单和填写表单可以调用Weka的所有功能。

缺陷：要求将所需数据全部依次读进内存，一旦用户打开某个数据集，就会批量读取全部数据。因此，这种批量方式仅适合处理中小规模问题。

KnowledgeFlow（知识流）：可以使用增量方式的算法来处理大规模的数据集，无需一次性全部导入。

Experimenter（实验者）：对于一个已知问题，哪种方法及参数值能够取得最佳效果。即可以用于进行多种算法性能的比较。ARFF格式是Weka的专用格式[2]。

Weka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上找到。

Weka支持深度学习[3]!

# Weka系统功能

## 2.1 Weka功能简介

Weka的主页面窗口有四个模块：

### 2.1.1 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。

### 2.1.2 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。

### 2.1.3 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。

### 2.1.4 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。

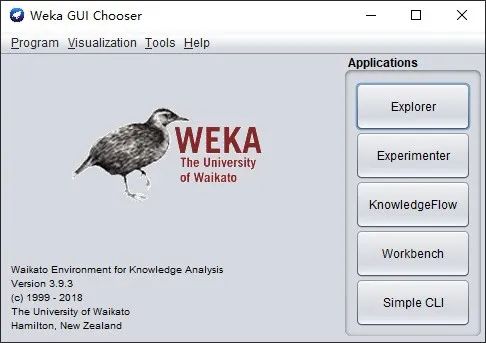


图1：Weka主界面

## Weka功能简介

### 2.2.1 数据挖掘和机器学习

机器学习方法构成数据挖掘的核心，绝大多数的数据挖掘技术来自于机器学习领域。数据挖掘是在数据中寻找模式的过程。机器学习定义为能够自动寻找数据中的模式的一套方法，使用所发现的模式进行预测或者决策。  
机器学习分为两种主要类型，第一种称为有监督的学习，或称为预测学习，其目标是在给定一系列输入输出实例所构成的数据集的条件下，学习输入X到输出y的映射关系。这里的数据集称为训练集。第二种为无监督的学习，或称描述学习，在给定一系列仅由输入实例构成的数据集的条件下，其目标是发现数据中的有趣模式。无监督学习有时也称知识发现。这类问题没有明确的定义，也没有明显的误差度量可供使用。

### 2.2.2 数据和数据集

[数据挖掘](https://so.csdn.net/so/search?q=%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98&spm=1001.2101.3001.7020)的对象可以是各种各样的书记处，这些书记处可以以各种形式存储，如数据库、数据仓库、数据文件、流数据、多媒体、网页，等等。即可以集中存储在数据存储库中，也可以分布在世界各地的网络服务器上。数据集就是待处理数据对象的集合。数据对象有多个别名，如记录、点、行、向量、案例、样本、观测等。属性用于刻画数据对象的基本特征。属性也有多个别名，如变量、特征、字段、维、列等。

 属性可以分为四种类型：标称（nominal）、序数（ordinal）、区间（interval）和比率（ratio），其中，标称属性仅提供区分对象的足够信息，如性别（男、女）等；序数属性的值可以提供确定对象的顺序的足够信息，如成绩等级（优、良、中、及格、不及格）等；区间属性的值之间的差是有意义的，即存在测量单位，如温度、日历日期等；比率属性的值之间的差和比值都是有意义的，如绝对温度、年龄，长度等。标称属性和序数属性统称为分类的（Categorical）或定性的（Qualitative）属性，它们的取值为集合，即使使用数值来表示，也不具备数的大部分性质，知识一个符号而已；区间属性和比率属性统称为定量的（Quantitative）或数值的（Numeric）属性，定量属性可以使用整数值或连续值来表示，具备数的大部分性质。

### 2.2.3 ARFF格式

 ARFF是Weka专用的数据格式，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力列表，由独立且无序的实例组成。

### 2.2.4 预处理

 数据挖掘所使用的数据往往不是专门为挖掘准备的，人的错误、测量设备的限制以及数据收集过程的漏洞都可能导致一些问题。数据挖掘只能通过两个方面设法避免数据质量问题：1.数据质量问题的检测与纠正；2.使用能容忍低质量数据的算法。第一种方式在数据挖掘前检测并纠正一些质量问题，这个过程称为数据的预处理；第二种方式需要提高算法的健壮性。

 数据预处理是数据挖掘的重要步骤，Weka专门提供若干过滤器进行预处理，还在Explorer界面中提供选择属性标签页专门处理属性的自动选择问题。数据预处理包括如下技术：

 1）聚类

聚类（Aggregation）就是将两个或多个对象合并为单个对象。一般定量数据通常通过求和或者求平均值的方式进行聚集，定性数据通常通过汇总进行聚集。聚集通过数据规约来减少数据量，所导致的较小[数据集](https://so.csdn.net/so/search?q=%E6%95%B0%E6%8D%AE%E9%9B%86&spm=1001.2101.3001.7020)只需要较少内存和处理时间的开销。但有可能导致有趣细节的丢失。

 2）抽样

如果处理全部数据的开销过大，数据预处理可以使用抽样，只选择数据对象的子集进行分析。使用抽样可以压缩数据量，但是要确保样本近似地具有原数据相同的性质。

3）维度归约

 维度是指数据集中属性的数目。维度归约（Dimension Reduction）是指创建新属性，通过数据编码或数据变换，将一些旧属性合并在一起以降低数据集的维度。

4）属性选择

 除维度归约外，降低维度的另一种方法是仅只使用属性的一个子集。表面看来似乎这种方法可能丢失信息，但很多情况下，数据集存在冗余和不相关的属性。如果将全部可能的属性子集作为输入，然后选取能产生最好结果的子集（理想算法），这样工作量巨大。这是考虑三种标准的属性选择方法：嵌入、过滤和包装。

嵌入方法（Embedded Approach）将属性选择作为数据挖掘算法的一部分。在挖掘算法运行期间，算法本身决定使用哪些属性以及忽略哪些属性。决策树算法通常使用这种方法。过滤方法（Filter Approach）在运行数据挖掘算法之间，使用独立于数据挖掘任务的方法进行属性选择。包装方法（Wrapper Approach）将学习算法的结果作为评价准则的一部分，使用类似前文介绍的理想算法，但通常无法枚举出全部可能的子集以找出最佳属性子集。

 5）属性创建

 属性创建就是通过对数据集中旧的属性进行处理。创建新的数据集，这样能更有效的获得重要的信息。由于通常新数据集的维度比原数据集少，可以获得维度归约带来的好处。属性创建有三种方式：属性提取、映射数据到新空间和属性构造。

 6）离散化和二元化

将连续属性转化为分类属性称为离散化（Discretization），将连续和离散属性转换为一个或多个二元属性称为二元化（Binarization）。

7）变量变换

变量变换也称属性变换（variable transform），是指用于变量的所有值的变换。变量变换包括简单函数变换和规范化。

### 2.2.5 分类与回归

### 2.2.6 聚类分析

聚类（clustering）就是将数据集划分为有若干相似实例阻生的簇（cluster）的过程，使得同一个簇中实例间的相似度最大化，不同簇的实例间的相似度最小化。

### 2.2.7 关联分析

关联分析（association analysis）方法就是用于发现隐藏在大型数据集中的有意义的联系，这种联系可以用关联规则（association rule）进行表示。最具代表性的例子就是商家通过关联分析挖掘商场销售数据，发现客户的购买习惯。

### 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。

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]

[2] 百度百科，URL：https://baike.baidu.com/item/Weka/10701215?fr=aladdin

。

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］．