

课程设计报告

**姓 名：** 甄泽宇

**学 号：** 19851974

**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**

# 自己的标题

……

## 1 Weka网站介绍

数据挖掘、机器学习这些字眼，在一些人看来，是门槛很高的东西。诚然，如果做算法实现甚至算法优化，确实需要很多背景知识。但事实是，绝大多数数据挖掘工程师，不需要去做算法层面的东西。他们的精力，集中在特征提取，算法选择和参数调优上。那么，一个可以方便地提供这些功能的工具，便是十分必要的了。而Weka，便是数据挖掘工具中的佼佼者。

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，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。

### 1.1 Weka发展历史

1992年末，新西兰怀卡托大学计算机科学系Ian Written博士申请基金。

1993年获新西兰政府资助，并于同年开发出接口和基础架构。

1994年发布了第一个Weka的内部版本。

1996年，第一个公开版本Weka2.1发布。Weka的早期版本主要采用C语言编写。

1997年，团队决定使用Java重新改写。Java允许用一个统一的接口来学习方案和方法的预处理和后处理，决定使用Java来代替C++或其他面向对象的语言，是因为Java编写的程序可以运行在绝大部分计算机上，而无须重新编译，更不需要修改源代码。Weka已经测试过的平台包括Linux、Windows和Macintosh操作系统，甚至PDA。最后的可执行程序复制过来即可运行，完全绿色，不需求复杂安装。当然，Java也有其缺点，最大的问题是它在速度上有缺陷，执行一个Java程序比相应的C语言程序马上好几倍。

1999年，发布Java的Weka 3版本。

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17年的发展历史）。Weka的每月下载次数已超过万次。

2013年，Weka的最新版本是3.7.8。

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。

### 1.2 Weka基本内容

Weka平台提供一个统一界面，汇集了当今经典的机器学习算法及数据预处理工具。作为知识获取的完整系统,包括了数据输入、预处理、知识获取、模式评估等环节，以及对数据及学习结果的可视化操作。并且可以通过对不同的学习方法所得出的结果进行比较，找出解决当前问题的最佳算法。

Weka平台提供一个统一界面，汇集了当今经典的机器学习算法及数据预处理工具。作为知识获取的完整系统,包括了数据输入、预处理、知识获取、模式评估等环节，以及对数据及学习结果的可视化操作。并且可以通过对不同的学习方法所得出的结果进行比较，找出解决当前问题的最佳算法。

Weka主要提供了数据挖掘的框架结构，并给出了一系列的数据挖掘算法，在这些数据挖掘算法生成模型时，用户可以自定义算法的各个相关参数，并且可以实现对测试集的验证, 和对算法的比较。

Weka提供了可编程的接口，使我们可以通过拓展算法来实现新的功能，同时Weka也允许我们通过调用它的公用包，在自己的项目中利用包中的算法实现相应的功能。

Weka系统汇集了最前沿的机器学习算法和数据预处理工具，以便用户能够快速灵活地将已有的数据处理方法应用于新的数据集。它为数据挖掘的整个过程提供全面的支持，包括准备输入数据、统计评估学习方案、输入数据和学习效果的可视化。Weka除了提供大量学习算法之外，还提供了适应范围很广的预处理工具，用户通过一个统一界面操作各种组件，比较不同的学习算法，找出能够解决问题的最有效的方法。

Weka系统包括处理标准数据挖掘问题的所有方法：回归、分类、聚类、关联规则以及属性选择。分析要进行处理的数据是一个重要的环节，Weka提供了很多用于数据可视化和预处理的工具。输入数据可以有两种形式，第一种是以ARFF格式为代表的文件；另一种是直接读取数据库表。

使用Weka的方式主要有三种：第一种是将学习方案应用于某个数据集，然后分析其输出，从而更多地了解这些数据；第二种是使用已经学习到的模型对新实例进行预测。第三种是使用多种学习器，然后根据其性能表现选择其中的一种来进行预测。用户使用交互式界面菜单中选择一种学习方法，大部分学习方案都带有可调节的参数，用户可以通过属性列表或对象编辑器修改参数，然后通过同一个评估模块对学习方案的性能进行评估。

## 2 Weka的基本结构和功能

### 2.1 Weka软件系统功能

Weka主要包括4个核心部分，也是4大功能环境，分别是：

#### 2.1.1 SimpleCLI

提供了一个简单的命令行界面，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

#### 2.2.2 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，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。

#### 2.2.3 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。

Experimenter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，而仅计算还没有被实验的那些。

#### 2.2.4 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＂（KnowledgeFlow）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。

### 2.2 **机器学习的分类**

机器学习分为两种主要类型。

#### 2.2.1 有监督学习（预测学习）

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，实例的个数称为训练样本数。（从分类角度看，即训练集中各组数据的类别已知）可以对所观察到的值与预测值进行比较，得到明确的误差值。

#### 2.2.2 无监督学习（描述学习）

目标是在给定一系列输入实例构成的数据集的条件下，发现数据中的有趣模式。因为我们不知道需要寻找什么样的模式，也没有明显的误差度量可供使用。

### **2.3 ARFF格式**

ARFF格式是Weka的专用格式。

数据集是实例的集合，每个实例包含一定的属性，属性的数据类型包括以下几类:

（1）标称型（nominal），只能取预定义值列表中的一个，如性别（男，女），衣服颜色（红、黄、蓝）。

（2）数值型（numeric），只能表示整数或实数。

（3）字符串型（string），由双引号引用的任意长度的字符列表。

（4）日期型（date）

（5）关系型（relational）

ARFF文件就是实例类型的外部表示，其中包括一个标题头（header），以描述属性的类型，还包含一个用逗号分隔的列表所表示的数据（data）部分。

## 3 数据集合

### 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。

### 3.5 举例

#### 3.5.1 weather

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 |

#### 3.5.2 Z-Alizadeh sani

Z-Alizadeh Sani数据集包含303名患者的记录，每个患者有54个特征。这些特征分为四组：人口统计学特征、症状和检查特征、心电图特征、实验室特征和回声特征。

表1-2 Z-Alizadeh sani dataset数据集合

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | Age | Weight | Length | Sex | BMI | DM | Obesity | CRF |
| 1 | 53 | 90 | 175 | Male | 29.3877551 | 0 | Y | N |
| 2 | 67 | 70 | 157 | Fmale | 28.39872 | 0 | Y | N |
| 3 | 54 | 54 | 164 | Male | 20.07733 | 0 | N | N |
| 4 | 66 | 67 | 158 | Fmale | 26.83865 | 0 | Y | N |
| 5 | 50 | 87 | 153 | Fmale | 37.16519 | 0 | Y | N |
| 6 | 50 | 75 | 175 | Male | 24.4898 | 0 | N | N |
| 7 | 55 | 80 | 165 | Male | 29.38476 | 0 | Y | N |
| 8 | 72 | 80 | 175 | Male | 26.12245 | 1 | Y | N |
| 9 | 58 | 84 | 163 | Fmale | 31.61579 | 0 | Y | N |
| 10 | 60 | 71 | 170 | Male | 24.56747 | 1 | N | N |
| 11 | 58 | 75 | 168 | Male | 26.57313 | 0 | Y | N |
| 12 | 80 | 67 | 153 | Fmale | 28.62147 | 0 | Y | N |
| 13 | 70 | 70 | 151 | Fmale | 30.70041 | 1 | Y | N |
| 14 | 67 | 63 | 154 | Fmale | 26.56434 | 1 | Y | N |
| 15 | 66 | 63 | 155 | Fmale | 26.22268 | 1 | Y | N |
| 16 | 59 | 81 | 167 | Male | 29.04371 | 1 | Y | N |
| 17 | 41 | 68 | 169 | Male | 23.80869 | 0 | N | N |
| 18 | 68 | 59 | 161 | Fmale | 22.76147 | 0 | N | N |
| 19 | 60 | 89 | 163 | Fmale | 33.49769 | 1 | Y | N |
| 20 | 65 | 72 | 150 | Fmale | 32 | 1 | Y | Y |
| 21 | 47 | 84 | 170 | Fmale | 29.06574 | 0 | Y | N |
| 22 | 66 | 89 | 151 | Fmale | 39.03338 | 0 | Y | N |
| 23 | 66 | 75 | 170 | Male | 25.95156 | 1 | Y | N |
| 24 | 72 | 66 | 161 | Fmale | 25.46198 | 1 | Y | N |
| 25 | 50 | 66 | 164 | Fmale | 24.53896 | 1 | N | N |

# 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1905甄泽宇19851074。

（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] 代码交流, URL: https://www.daimajiaoliu.com/daima/4ef21d19f900401. [2022-6-16]

[3] 代码交流,URL: https://www.daimajiaoliu.com/daima/61048af2b19ec00. [2022-6-16].