

课程设计报告

|  |  |
| --- | --- |
| **姓名：** | **宋晓涵** |
| **学号：** | **19851122** |
| **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**

你自己的标题

Weka是一款免费的，非商业化的，基于JAVA环境下开源的机器学习以及数据挖掘软件。[Weka](https://so.csdn.net/so/search?q=Weka&spm=1001.2101.3001.7020)系统汇集了前沿的机器学习算法和数据预处理工具，提供数据处理，特征选择、分类、回归、聚类、关联规则、可视化等功能，以便用户能够快速灵活地将已有的成熟处理方法应用于新的数据集。

## 1 Weka网站介绍

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](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0/217599" \t "_blank)（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" \t "_blank)（data mining）软件。同时Weka也是[新西兰](https://www.zhihu.com/search?q=%E6%96%B0%E8%A5%BF%E5%85%B0&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22article%22%2C%22sourceId%22%3A%22450076241%22%7D" \t "_blank)的一种鸟名，而Weka的主要开发者同时恰好来自新西兰怀卡托大学 [1] 。

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://link.zhihu.com/?target=https%3A//baike.baidu.com/item/%25E6%259C%25BA%25E5%2599%25A8%25E5%25AD%25A6%25E4%25B9%25A0%25E7%25AE%2597%25E6%25B3%2595/18635836" \t "_blank)，包括对数据进行预处理，分类，回归、聚类、[关联规则](https://link.zhihu.com/?target=https%3A//baike.baidu.com/item/%25E5%2585%25B3%25E8%2581%2594%25E8%25A7%2584%25E5%2588%2599)以及在新的交互式界面上的可视化。

Weka平台提供一个统一界面，汇集了当今经典的[机器学习](http://lib.csdn.net/base/2" \t "_blank" \o "undefined)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 [2] 。

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。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)。其中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 [1] 。

## 2 Weka系统功能介绍

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处理方法应用于新的数据集。

Weka主界面为Weka GUI选择器，它通过四个按钮提供四种主要的应用程序供用户选择 [3] 。

### 2.1 **Explorer（探索者）**

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。

虽然探索者界面使用很方便，但它也存在一个缺陷，要求它将所需数据全部一次读进内存，一旦用户打开某个数据集，就会读取全部数据。因此，这种批量方式仅适合处理中小规模的问题。而**知识流**刚好能够弥补这一缺陷。

### 2.2 **KnowledgeFlow（知识流）**

使用增量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构件组合在一起，形成数据流。

### 2.3 **Experimenter（实验者）**

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？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。

### 2.4 **Simple CLI（简单命令行）**

这个界面是为不提供自己的命令行界面的操作系统提供的，该简单命令行界面用户和用户进行交互，可以直接执行Weka命令。

## 3 数据集合

### 3.1 数据格式arff

Weka的数据是.arff（Attribute-Relation File Format）格式的，是ASCII文本文件。以Weka自带的weather.nominal.arff文件为例：

@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}

@attribute temperature {hot, mild, cool}

@attribute humidity {high, normal}

@attribute windy {TRUE, FALSE}

@attribute play {yes, no}

@data

sunny,hot,high,FALSE,no

sunny,hot,high,TRUE,no

overcast,hot,high,FALSE,yes

rainy,mild,high,FALSE,yes

rainy,cool,normal,FALSE,yes

rainy,cool,normal,TRUE,no

overcast,cool,normal,TRUE,yes

sunny,mild,high,FALSE,no

sunny,cool,normal,FALSE,yes

rainy,mild,normal,FALSE,yes

sunny,mild,normal,TRUE,yes

overcast,mild,high,TRUE,yes

overcast,hot,normal,FALSE,yes

rainy,mild,high,TRUE,no

Weka所处理的数据集是一个二维的表格。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。

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 |

整个ARFF文件可以分为两个部分。第一部分给出了头信息，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，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 [4] 。

### 3.2 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）。

### 3.3 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。

声明语句的顺序很重要:它表明了该项属性在数据部分的位置;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。

### 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。

### 3.5 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。

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。

@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"}

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。

### 3.6 格式转换

在Weka中给用户提供了从数据库和Excel导出CSV文件转换成ARFF文件的方法。

举例：UCI数据集Challenger USA Space Shuttle O-Ring中，将o-ring-erosion-only.data文件转换成ARFF格式文件。

1. 格式强转，将.data后缀名改为.csv格式：





1. Weka打开CSV文件，将CSV格式文件保存为ARFF格式文件：

参考文献:

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

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

[3] 博客园， URL: https://www.cnblogs.com/chamie/p/4551514.html, [2022-6-17]

[4] CSDN, URL:https://blog.csdn.net/buaalei/article/details/7103055, [2022-6-18]