

课程设计报告

**姓 名：** 付金伟

**学 号：** 19851182

**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**

# 机器学习应用系统与大数据分析

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" \t "_blank)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" \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。它和它的[源代码](https://baike.baidu.com/item/%E6%BA%90%E4%BB%A3%E7%A0%81/3969" \t "_blank)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" \t "_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.baidu.com/item/%E6%80%80%E5%8D%A1%E6%89%98%E5%A4%A7%E5%AD%A6/7082322" \t "_blank)（The University of Waikato）[1]。

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。

我们整理了几门免费的在线课程，教授使用Weka进行机器学习和数据挖掘。这些课程的视频可在Youtube上找到。

Weka支持深度学习[2]。

Weka提供的功能有数据处理，特征选择、分类、回归、聚类、关联规则、可视化等。

### 2.1 数据处理

Weka的窗口右侧共有四个应用，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。Experimentor用来进行实验，对不同学习方案进行数据测试的环境。KnowledgeFlow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。SimpleCLI简单的命令行界面。

### 2.2 特征选择

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。在Weka内进行的是数据预处理，训练，验证这三个步骤。数据预处理包括特征选择，特征值处理（比如归一化），样本选择等操作。训练包括算法选择，参数调整，模型训练。验证是对模型结果进行验证。数据预处理，打开Explorer界面，点”open file”，在Weka安装目录下，选择data目录里的”labor.arff”文件，会看到主界面，将整个区域分为七个部分。在Explorer中，打开classifer选项卡，整个界面被分成几个区域。分别是Classifier，点击choose按钮，可以选择Weka提供的分类器。Test options评价模型效果的方法，有四个选项。可视化，打开Explorer的Visualize面板，可以看到最上面是一个二维的图形矩阵，该矩阵的行和列均为所有的特征（包括类别标签），第i行第j列表示特征i和特征j在二维平面上的分布情况。图形上的每个点表示一个样本，不同的类别使用不同的颜色标识[3]。

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。

### 2.3 分类与回归

Weka是新西兰怀卡托大学用Java开发的数据挖掘著名开源软件，该系统自1993年开始由新西兰政府资助，至今已经历了20年的发展，其功能已经十分强大和成熟。Weka集合了大量的机器学习和相关技术，受领域发展和用户需求所推动，代表了当今数据挖掘和机器学习领域的最高水平。

Weka是一种使用Java语言编写的数据挖掘机器学习软件，是GNU协议下分发的开源软件。Weka主要用于科研、教育和应用领域。

#### 2.3.1 分类

国内外很多著名大学都采用Weka作为数据挖掘和机器学习课程的实践工具。Weka还有另外一个名字叫作Pentaho Data Mining Community Edition(Pentaho数据挖掘社区版)，此外，Pentaho的网站([http://Weka.pentaho.com/](http://weka.pentaho.com/))还维护一个称为Pentaho Data Mining Enterprise Edition(Pentaho数据挖掘企业版)的版本，它主要提供技术支持和管理升级。另一个用Java编写的著名数据挖掘工具RapidMiner通过WekaExtension(Weka扩展)支持Weka，以充分利用Weka的“约100个额外的建模方案，其中包括额外的决策树、规则学习器和回归估计器”。

Weka通常使用ARFF文件格式的文件。也可以直接使用CSV文件格式的文件，但与传统CSV文件不同，Weka能识别的CSV文件要求第一行给各列的定义。因为CSV文件比较容易获得，excel表格文件可以直接另存为csv文件。推荐使用csv文件。

#### 2.3.2 回归

遵循Linux发布模型，其中版本号的偶数第二位表示“稳定”版本，奇数秒数字表示“开发”版本（例如，3.0.x是稳定版本，3.1.x是开发版本）。如果使用的是开发版本，则可能会有新功能，但这些功能完全可能是暂时的和/或不稳定的，并且不能保证API和/或模型的向后兼容性。如果需要在应用程序中进行教学或部署的稳定性，最好使用稳定版本的Weka。

源代码存储库，当您下载特定版本的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/)。

如果喜欢使用 Git 而不是 Subversion 进行软件开发，那么 [Weka3.8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，我们继续使用 Subversion 将更改提交到软件，而不是 Git。

### 2.4 聚类

### 关联规则

### 可视化等。

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，包括Weka3的早期非Java前身，是在[新西兰](https://en.wikipedia.org/wiki/New_Zealand)[汉密尔顿](https://en.wikipedia.org/wiki/Hamilton,_New_Zealand)[的怀卡托大学](https://www.waikato.ac.nz/)[计算机科学系](https://www.cs.waikato.ac.nz/)开发的。Weka3的大部分内容是由Eibe Frank，Mark Hall，Peter Reutemann和Len Trigg编写的，但许多其他人做出了重大贡献，特别是Remco Bouckaert，Richard Kirkby，Ashraf Kibriya，Xin Xu和Malcolm Ware。有关贡献者的完整信息，请查看从Weka源代码中提取的Javadoc，这是可用[文档](https://waikato.github.io/weka-wiki/documentation/)的一部分。

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

类路径修改，环境变量告诉 Java 在何处查找类。由于Java以“先到先得”的方式进行搜索，因此您必须注意在CLASSPATH中放置的位置和内容。就我个人而言，我从不使用环境变量，因为我经常并行处理不同版本的项目。CLASSPATH只会把事情搞砸，如果你不小心（或者只是忘记删除一个条目）。[ANT](https://waikato.github.io/weka-wiki/ant/)为构建（并分离源代码和类文件）Java项目提供了一种很好的方式。但是，如果您只处理完全独立的项目，那么使用环境变量可能是最简单的。

可视化选择，可从弹出式菜单中访问*分类器*、*聚类*和*属性选择*面板中的可视化在“结果”列表中的条目上单击鼠标右键以显示菜单。您将看到用于查看或保存文本输出的选项，以及用于可视化错误，聚类，树等的更多选项，具体取决于方案。

Weka下载问题，下载Weka时，请确保生成的文件大小与Weka网页相同。否则，事情将无法正常工作。显然，一些网络浏览器在下载Weka时遇到问题。另请注意，Weka主页仅链接到[sourceforge.net 托管的文件](http://sourceforge.net/)。这通常涉及重定向到镜像，您将从中下载实际文件。

走出记忆异常，大多数 Java 虚拟机仅分配一定的最大内存量来运行Java程序。通常，这比计算机中的RAM量要少得多。[此处](https://docs.oracle.com/javase/8/docs/technotes/guides/vm/gc-ergonomics.html)提供了一些有关适用于Java 8的 Oracle Java 虚拟机中默认堆大小的信息。但是，您可以通过设置适当的选项来扩展虚拟机的可用内存。例如，使用Oracle的JDK，您可以将最大Java堆大小设置为2GB。

设置Oracle Java虚拟机的最大堆大小（并覆盖启动脚本等中可能提供的任何其他设置）的可靠方法是使用\_JAVA\_OPTIONS 环境变量来指定该选项。[这里有](http://stackoverflow.com/questions/28327620/difference-between-java-options-java-tool-options-and-java-opts)更多信息。

得到错误消息训练和测试集不兼容，将最大Java堆大小设置为2GB。

设置Oracle Java虚拟机的最大堆大小（并覆盖启动脚本等中可能提供的任何其他设置）的可靠方法是使用\_JAVA\_OPTIONS环境变量来指定该选项。[这里有](http://stackoverflow.com/questions/28327620/difference-between-java-options-java-tool-options-and-java-opts)更多信息。

无法从数据库中读取为止数据类型，由于有太多不同的数据库，每个数据库都有自己的数据类型，因此不可能事先定义所有这些数据库。因此，Weka附带了针对不同数据库的设置，允许您在不进行任何额外调整的情况下运行实验。但是，如果要从其他数据源读取数据库，则可能需要告诉Weka如何导入这些数据类型。

尝试在类路径中添加JDBC驱动程序错误，Weka的数据库默认设置试图在启动时找到一些常见的[JDBC](http://en.wikipedia.org/wiki/jdbc)驱动程序类（“JDBC”是连接到数据库的Java方式，如MySQL，HSQLDB等）。只需将这些 JDBC 驱动程序添加到 [CLASSPATH](https://waikato.github.io/weka-wiki/classpath/) 中，Weka就可以自动连接到这些数据库。如果您没有尝试访问数据库，请忘记这些消息。否则，请查看[数据库](https://waikato.github.io/weka-wiki/databases/)一文以获取更多信息（默认情况下，您尝试连接到的数据库类型可能不会列出）。

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。 [1]

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" \t "_blank)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WekaMOOC。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。

这里我们要介绍一下Weka中的术语。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](https://baike.baidu.com/item/%E4%BA%8C%E7%BB%B4%E8%A1%A8%E6%A0%BC" \t "_blank)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。

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），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。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New Zealand独有的一种鸟名，而Weka的主要开发者同时恰好来自新西兰的the University of Waikato。（本段摘自百度百科）。

Weka提供的功能有数据处理，特征选择、分类、回归、聚类、关联规则、可视化等。本文将对Weka的使用做一个简单的介绍，并通过简单的示例，使大家了解使用Weka的流程。本文将仅对图形界面的操作做介绍，不涉及命令行和代码层面的东西。

Weka是基于java，用于数据挖掘和知识分析一个平台。来自世界各地的java爱好者们都可以把自己的算法放在这个平台上，然后从海量数据中发掘其背后隐藏的种种 关系；开发并投入使用两年左右的时间，但是已经是很多人受益，特别是摆脱了繁重的海量数据。

安装的是Weka-3-4-10jre，虽然有3.5的版本，但是一般校园使用已经足够。安装方法很简单，并且继承了java 运行环境，不用担心任何配置。版本据说提供了function接口，并且支持反编译。用户可用性以及扩展性很强。

安装完成，运行Weka图标（那只火鸟，也是Weka名称的由来），出现一个小型的GUI，提供了四种操作环境：CLI,Explorer,Experimenter,knowledgeflow。

其实四种操作基本原理都大同小异，只是提供的environment不一样。看自己平时熟悉使用什么，比如喜欢直接用代码的朋友一定热衷于：CLI，喜欢图标的倾情于:Knowledgeflow，通常功能形控制的:explorer,experimenter感觉和explorer类似，致使支持格式不一样。

打开后，我们可以看到顶层的六个标签，分别是：Preprocess、Classify、Cluster、Associate、Select attributes、Visualize。在preprocess里面有个open控件，是用来打开sourcefile的，Weka支持的文档格式为.arff，其实是一个文本数据集。也支持URL或者DB打开方式，并且支持数据转换。

打开数据文件后，可以使用Filter进行一下过滤，相当于“预处理的预处理”。Filter提供了许多算法来过滤数据，比如 filters/unsupervised/instance/normalize应该是一个标准化的算法。当然，也可以编写你自己的算法!

这时窗体上已经给出这个数据集的一些基本特征了，比如有多少属性，各属性的一些简单统计量，右下方还给出一些可视化效果比如柱状图。通过这些可以初步了解这个数据集了。

接下来的两个标签页是classify(分类)和cluster(聚类)，接触数据挖掘的人对它们一定不会陌生。同样Weka有许多分类和聚类算法可供选择，在这里面称为clasifier和clusterer。不过Weka提供的classify功能似乎还不够灵活，只能定长度和定频率地分类。但这个关系不大，现在很多数据处理软件都可以做到这个，比如excel。Cluster功能强大，提供了许多巧妙的聚类算法，选定一个算法，给出你所需要生成的聚类数目，就可以自动完成。

这是一个用于发掘Associate Rules(关联规则)的模块。

将前面导入的数据使用Associator进行发掘，就可以发现其中无数隐藏的关系。Weka-3 -4提供了Apriori、PredictiveApriori、Tertius三种关联规则发掘算法，不过我感觉这已经够用了。选定一个算法，进行一些必要的设置，包括支持度上界、下界，每次运算的支持度递减值，等等。另外一个重要的参数：所需要生成的关联规则个数。太不可思议了，以前我们能从海量数据中发现一个关联规则就已经沾沾自喜，现在Weka居然问你想生成多少关联规则!

参数设置完成，点Start，之后10条关联规则已经生成，可以提交了。当然，你还可以分析一下哪些规则比较有用，哪一条有潜在收益，这就需要business sense了。

另外两个标签页还没怎么看。Select attributes大概是针对单属性的分析？Visualize则提供了许多可视化效果，需要拿出去演示时很方便。也有可能是我用错了。

Weka实在是一个伟大的工具。基于java，却没有运行其它java程序那种慢吞吞的感觉。也不排除刚加了一个内存的关系。有了Weka，Data Mining也可以轻轻松松了[4]!

有关表格的规定

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 |

# 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].

[2] 百度百科, URL: https???????. [2022-6-16].

。

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］．