

课程设计报告

|  |  |
| --- | --- |
| **姓 名：** | **王盛** |
| **学 号：** | **19851188** |
| **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**

课程设计报告

**1 WeKa官网介绍**

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。[1]

**2 WeKa系统功能**

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，它通过右边的四个按钮提供四种主要的应用程序供用户选择：  
 探索者（Explore），系统提供的最容易使用的图像用户接口。通过选择菜单和填写表单，可以调用WeKa的所有功能。  
 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。  
知识流（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。  
实验者（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。简单命令行（Simple CLI），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)数据格式  
 WeKa的数据是.arff（Attribute-Relation File Format）格式的，是ASCII文本文件。相当于一张二维的表格，@attribute中第二列是表中各项的属性，第二列指定了表格该项的类型，一般有五种类型：{NORMINAL, NUMBERIC, STRING, DATE, RELATION}。  
其中@RELATION 我自己的理解是相当于给个名字，这二个字符串包含空格，必须加上引号，英文符号中的单引和双引都可；@ATTRIBUTE 中第二列是表中各项的属性，第二列指定了表格该项的类型，一般有五种类型：{NORMINAL, NUMBERIC, STRING, DATE, RELATION}；@DATA标记后面写的都是数据，每一行都是一组数据。  
 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，如果不嫌麻烦写个写文件的函数也一样。  
 需要注意的是，在Windows记事本打开文件时，可能会因为回车符定义不一致而导致分行不正常。推荐使用UltraEdit这样的字符编辑软件察看ARFF文件的内容。  
 (4)稀疏数据格式  
 稀疏数据格式Sparse ARFF和ARFF文件很像，是WeKa中提供的为了有大量零值数据格式。稀疏数据格式一般是<index><space><value>。  
 (5)文本数据格式  
 文本数据格式在数据格式中并没有，但是在WeKa的数据样例中有这样的例子。当你把文本用TextDirectoryLoader将文本集合转化成数据的时候就会形成这样的格式。  
 (6)关联规则  
 关联规则又称购物栏分析。目前，WeKa的关联规则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。  
 (7)数据准备  
 使用WeKa作数据挖掘，面临的第一个问题往往是我们的数据不是ARFF格式的。幸好，WeKa还提供了对CSV文件的支持，而这种格式是被很多其他软件所支持的。此外，WeKa还提供了通过JDBC访问数据库的功能。  
 (8)分类与回归  
 WeKa把分类(Classification)和回归(Regression)都放在"Classify"选项卡中，这是有原因的。  
 在这两个任务中，都有一个目标属性(输出变量)。我们希望根据一个样本(WeKa中称作实例)的一组特征(输入变量)，对目标进行预测。为了实现 这一目的，我们需要有一个训练数据集，这个数据集中每个实例的输入和输出都是已知的。观察训练集中的实例，可以建立起预测的模型。有了这个模型，我们就可 以新的输出未知的实例进行预测了。衡量模型的好坏就在于预测的准确程度。  
 在WeKa中，待预测的目标(输出)被称作Class属性，这应该是来自分类任务的"类"。一般的，若Class属性是分类型时我们的任务才叫分类，Class属性是数值型时我们的任务叫回归。  
 (9)聚类分析  
 聚类分析中的"类"(cluster)和前面分类的"类"(class)是不同的，对cluster更加准确的翻译应该是"簇"。聚类的任务是把 所有的实例分配到若干的簇，使得同一个簇的实例聚集在一个簇中心的周围，它们之间距离的比较近;而不同簇实例之间的距离比较远。对于由数值型属性刻画的实例来说，这个距离通常指欧氏距离。

**3** 数据集合格式

**3.1 数据集格式简介**

表3-1 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 |

将上述表格中数据制作成WeKa数据集如下：

@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数据以“%”开始的行是注释，WeKa将忽略这些行。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。[2]

**3.2 通过WeKa构建数据集**

首先通过UCI下载数据集，本文所选北京PM 2.5情况数据集，数据集地址：<https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data>，通过网站下载到CSV数据集。简单列举数据集情况如下：

表3-2 北京PM2.5数据集

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No | year | month | day | hour | pm2.5 | DEWP |
| 1 | 2010 | 1 | 1 | 0 | NA | -21 |
| 2 | 2010 | 1 | 1 | 1 | NA | -21 |
| 3 | 2010 | 1 | 1 | 2 | NA | -21 |
| 4 | 2010 | 1 | 1 | 3 | NA | -21 |
| 5 | 2010 | 1 | 1 | 4 | NA | -20 |
| 6 | 2010 | 1 | 1 | 5 | NA | -19 |
| 7 | 2010 | 1 | 1 | 6 | NA | -19 |
| 8 | 2010 | 1 | 1 | 7 | NA | -19 |
| 9 | 2010 | 1 | 1 | 8 | NA | -19 |
| 10 | 2010 | 1 | 1 | 9 | NA | -20 |

表3-3 北京PM2.5数据集

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No | TEMP | PRES | cbwd | Iws | Is | Ir |
| 1 | -11 | 1021 | NW | 1.79 | 0 | 0 |
| 2 | -12 | 1020 | NW | 4.92 | 0 | 0 |
| 3 | -11 | 1019 | NW | 6.71 | 0 | 0 |
| 4 | -14 | 1019 | NW | 9.84 | 0 | 0 |
| 5 | -12 | 1018 | NW | 12.97 | 0 | 0 |
| 6 | -10 | 1017 | NW | 16.1 | 0 | 0 |
| 7 | -9 | 1017 | NW | 19.23 | 0 | 0 |
| 8 | -9 | 1017 | NW | 21.02 | 0 | 0 |
| 9 | -9 | 1017 | NW | 24.15 | 0 | 0 |
| 10 | -8 | 1017 | NW | 27.28 | 0 | 0 |

表3-4 字段说明

|  |  |
| --- | --- |
| 字段名称 | 含义 |
| No | 行号 |
| year | 此行数据的年份 |
| month | 此行数据的月份 |
| day | 此行数据的日期 |
| hour | 此行数据的小时数 |
| pm2.5 | pm2.5浓度（微克/立方米） |
| DEWP | 露点 |
| TEMP | 温度 |
| PRES | 压力（hPa） |
| cbwd | 组合风向 |
| Iws | 累计风速（m/s） |
| Is | 累计降雪小时数 |
| Ir | 累计降雨小时数 |

了解了该数据集的构成，打开WeKa并进入Explorer界面，点击Open File，在弹出的界面中，选择文件拓展名为csv，打开我们下载好的csv数据集。并点击Save，将csv的后缀名去掉，即可保存为arff文件。打开arff文件部分显示如下：

@relation PRSA\_data\_2010.1.1-2014.12.31

@attribute No numeric

@attribute year numeric

@attribute month numeric

@attribute day numeric

@attribute hour numeric

@attribute pm2.5

@attribute DEWP numeric

@attribute TEMP numeric

@attribute PRES numeric

@attribute cbwd {NW,cv,NE,SE}

@attribute Iws numeric

@attribute Is numeric

@attribute Ir numeric

@data

1,2010,1,1,0,NA,-21,-11,1021,NW,1.79,0,0

2,2010,1,1,1,NA,-21,-12,1020,NW,4.92,0,0

3,2010,1,1,2,NA,-21,-11,1019,NW,6.71,0,0

4,2010,1,1,3,NA,-21,-14,1019,NW,9.84,0,0

5,2010,1,1,4,NA,-20,-12,1018,NW,12.97,0,0

6,2010,1,1,5,NA,-19,-10,1017,NW,16.1,0,0

7,2010,1,1,6,NA,-19,-9,1017,NW,19.23,0,0

8,2010,1,1,7,NA,-19,-9,1017,NW,21.02,0,0

通过上述操作，我们就将下载的csv数据集转换为了WeKa支持的arff文件。在WeKa中，选择打开刚转换好的arff文件，即可在右侧的面板中查看字段的相关信息，下方也会自动生成可视化图标展示数据的情况。

**参考文献**

1. 百度百科, URL: https://baike.baidu.com/item/weka/10701215?fr=aladdin#2. [2022-6-16].
2. CSDN, URL: https://blog.csdn.net/lijianggongyuan/article/details/44245131. [2022-6-16].