**北京交通大学海滨学院**

**实习报告**

**学生姓名： 丁芃懿**

**学 号： 19851205**

**学院(系)：计算机与信息技术学院**

**班 级： 计科1902班**

**指导教师： 王志海**

**北京交通大学海滨学院教务处**

# 《信息系统集成与开发》课程设计报告

1 数据挖掘Weka官网

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.sogou.com/lemma/ShowInnerLink.htm?lemmaId=138836896&ss_c=ssc.citiao.link)，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。

如果想自己实现[数据挖掘算法](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=68975778&ss_c=ssc.citiao.link)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]](https://baike.sogou.com/v10579652.htm?fromTitle=weka&ch=frombaikevr#quote1)

2014年3月起，新西兰怀卡托大学将推出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。通过以Arff格式为代表的文件进行输入

2 系统平台的基本功能

Weka系统汇集了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成熟处理方法应用于新的数据集。包括处理标准数据挖掘问题的所有方法：回归、分类、聚类、关联规则和属性选择。

**2.1 Explorer(探索者)**

图形用户界面，通过选择菜单和填写表单，可以调用Weka的所有功能。

**2.2 KnowledgeFlow(知识流)**

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组合在一起，形成数据流。

**2.3 Experimenter(实验者)**

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。

**2.4 Simple CLI(简单命令行)**

该界面用于和用户进行交互，可以直接执行Weka命令.

3 数据准备

跟很多电子表格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。

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 | TRUE | no |
| 6 | rainy | cool | normal | TRUE | no |
| 7 | overcast | cool | normal | TRUE | no |
| 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 |

**3.1 数据文件的格式**

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。

**3.2 CSV格式文件**

CSV格式文件是一种逗号分隔值格式的文件，是一种用来存储数据的纯文本格式文件。CSV文件由任意数目的记录组成，记录间以某种换行符分隔；每条记录由字段组成，字段间的分隔符是其它字符或字符串。