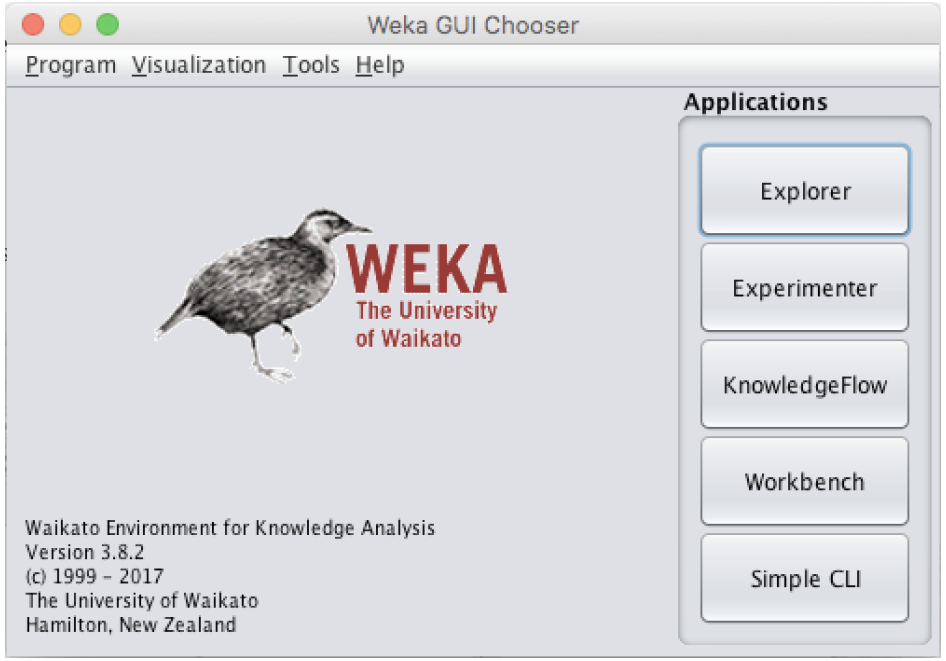
Introduction to Weka

by Prof.Weijia Jia (賈維嘉) E11-4007
Email: jiawj@umac.mo

Introduction

- WEKA的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），它的源代码:
- http://www.cs.waikato.ac.nz/ml/weka
- 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。



Introduction

- 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。
- 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。

Data Format

- WEKA所处理的数据集是图中那样的一个二维的表格
- 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
- 这样一个表格（数据集），在WEKA看来，呈现了属性之间的一种关系(Relation)

weather.arff

Relation: weather

No.	1.outlook Nominal	2.temperature Numeric	3.humidity Numeric	4.windy Nominal	5.play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

Data Format

- 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
- 图中所示的二维表格存储在ARFF文件中

```
% ARFF file for the weather data with some numric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
%
% 14 instances
%
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Data Format

- 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>

```
% ARFF file for the weather data with some numric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
%
% 14 instances
%
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Data Format

- 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。

```
% ARFF file for the weather data with some numric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
%
% 14 instances
%
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Data Format

- 属性声明的格式为 @attribute <attribute-name> <datatype>
- WEKA支持的<datatype>有四种:
  - numeric(real/integer)-数值型
  - <nominal-specification>-分类（nominal）型
  - string-字符串型
  - date [<date-format>]-日期和时间型

```
% ARFF file for the weather data with some numric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
%
% 14 instances
%
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Data Format

- 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。

```
% ARFF file for the weather data with some numric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
%
% 14 instances
%
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Kmeans-city

- 数据介绍
- 现有1999年全国31个省份城镇居民家庭平均每人全年消费性支出的八个主 要变量数据，这八个变量分别是:食品、衣着、家庭设备用品及服务、医疗 保健、交通和通讯、娱乐教育文化服务、居住以及杂项商品和服务。利用已 有数据，对31个省份进行聚类。
- 实验目的
- 通过聚类，了解1999年各个省份的消费水平在国内的情况。

Kmeans-City

城市	食品	衣着	家庭设备用品及服务	医疗保健	交通和通讯	娱乐教育文化服务	居住	杂项商品和服务
北京	2959	730.79	749.41	513.34	467.87	1141.82	478.42	457.64
天津	2460	495.47	697.33	302.87	284.19	735.97	570.84	305.08
河北	1496	515.9	362.37	285.32	272.95	540.58	364.91	188.63
山西	1406	477.77	290.15	208.57	201.5	414.72	281.84	212.1
内蒙古	1304	524.29	254.83	192.17	249.81	463.09	287.87	192.96
辽宁	1731	553.9	246.91	279.81	239.18	445.2	330.24	163.86
吉林	1562	492.42	200.49	218.36	220.69	459.62	360.48	147.76
黑龙江	1410	510.71	211.88	277.11	224.65	376.82	317.61	152.85
上海	3712	550.74	893.37	346.93	527	1034.98	720.33	462.03
江苏	2208	449.37	572.4	211.92	302.09	585.23	429.77	252.54
浙江	2629	557.32	689.73	435.69	514.66	795.87	575.76	323.36
安徽	1845	430.29	271.28	126.33	250.56	513.18	314	151.39
福建	2709	428.11	334.12	160.77	405.14	461.67	535.13	232.29
江西	1564	303.65	233.81	107.9	209.7	393.99	509.39	160.12
山东	1676	613.32	550.71	219.79	272.59	599.43	371.62	211.84
河南	1428	431.79	288.55	208.14	217	337.76	421.31	165.32
湖南	1942	512.27	401.39	206.06	321.29	697.22	492.6	226.45
湖北	1783	511.88	282.84	201.01	237.6	617.74	523.52	182.52
广东	3055	353.23	564.56	356.27	811.88	873.06	1082.82	420.81
广西	2034	300.82	338.65	157.78	329.06	621.74	587.02	218.27
海南	2058	186.44	202.72	171.79	329.65	477.17	312.93	279.19
重庆	2303	589.99	516.21	236.55	403.92	730.05	438.41	225.8
四川	1974	507.76	344.79	203.21	240.24	575.1	430.36	223.46
贵州	1674	437.75	461.61	153.32	254.66	445.59	346.11	191.48
云南	2194	537.01	369.07	249.54	290.84	561.91	407.7	330.95
西藏	2647	839.7	204.44	209.11	379.3	371.04	269.59	389.33
陕西	1473	390.89	447.95	259.51	230.61	490.9	469.1	191.34
甘肃	1526	472.98	328.9	219.86	206.65	449.69	249.66	228.19
青海	1655	437.77	258.78	303	244.93	479.53	288.56	236.51
宁夏	1375	480.89	273.84	317.32	251.08	424.75	228.73	195.93
新疆	1609	536.05	432.46	235.82	250.28	541.3	344.85	214.4

Kmeans-city

```
@relation city

@attribute city {Beijing,Tianjin,Hebei,Shanxi,Neimenggu,Liaoning,Jilin,
@attribute food numeric
@attribute clothes numeric
@attribute family-devices-services numeric
@attribute health-care numeric
@attribute transport-communication numeric
@attribute entertainment-education-cultural-services numeric
@attribute settlement numeric
@attribute miscellaneous-commodities-services numeric

@data
Beijing,2959.19,730.79,749.41,513.34,467.87,1141.82,478.42,457.64
Tianjin,2459.77,495.47,697.33,302.87,284.19,735.97,570.84,305.08
Hebei,1495.63,515.9,362.37,285.32,272.95,540.58,364.91,188.63
Shanxi,1406.33,477.77,290.15,208.57,201.5,414.72,281.84,212.1
Neimenggu,1303.97,524.29,254.83,192.17,249.81,463.09,287.87,192.96
```

DBSCAN-Internet

- 数据介绍
- 现有大学校园网的日志数据，290条大学生的校园网使用情况数据，数据包 括用户ID，设备的MAC地址，IP地址，开始上网时间，停止上网时间，上 网时长，校园网套餐等。利用已有数据，分析学生上网的模式。
- 实验目的
- 通过DBSCAN聚类，分析学生上网时间和上网时长的模式。

DBSCAN-Internet

学生上网日志（单条数据格式）	
记录编号	2c929293466b97a6014754607e457d68
学生编号	U201215025
MAC地址	A417314EEA7B
IP地址	10.12.49.26
开始上网时间	2014-07-20 22:44:18.540000000
停止上网时间	2014-07-20 23:10:16.540000000
上网时长	1558

DBSCAN-Internet

```
@relation online_time_students_add_start_minute

@attribute Record-Number {2c929293466b97a6014754607e457d68,2c929293466b97a60147546099a57d81,2c929293466b
@attribute student-number {U201215025,U201116197,M201373803,M201370611,U201112081,U201114462,M201370069,
@attribute mac-address {A417314EEA7B,F0DEF1C78366,88539523E88D,F0DEF167324F,B888E3813D3C,E4D53D56E6DD,E0
@attribute ip-address {10.12.49.26,222.20.71.38,10.12.59.230,218.197.241.94,218.197.229.5,10.12.67.39,22
@attribute start-time {'2014-07-20 22:44:18.540000000','2014-07-20 12:14:21.380000000','2014-07-20 22:56
@attribute stop-time {'2014-07-20 23:10:16.540000000','2014-07-20 23:25:22.380000000','2014-07-20 23:25:
@attribute online-time numeric
@attribute year numeric
@attribute education-background {'undergraduate dynamic IP template','graduate dynamic IP template','net
@attribute internet-charge {'100 yuan per half year','20 yuan per month','counting days','20 yuan monthl
@attribute type {internet,hust}
@attribute start-minute numeric

@data
2c929293466b97a6014754607e457d68,U201215025,A417314EEA7B,10.12.49.26,'2014-07-20 22:44:18.540000000','20
2c929293466b97a60147546099a57d81,U201116197,F0DEF1C78366,222.20.71.38,'2014-07-20 12:14:21.380000000','2
2c929293466b97a60147546099fa7d86,M201373803,88539523E88D,10.12.59.230,'2014-07-20 22:56:41.593000000','2
2c929293466b97a6014754609a137d88,M201370611,F0DEF167324F,218.197.241.94,'2014-07-20 23:19:30.930000000',
2c929293466b97a601475460ab577d99,U201112081,B888E3813D3C,218.197.229.5,'2014-07-20 16:51:56.657000000','
2c929293466b97a601475460ae397d9e,U201114462,E4D53D56E6DD,10.12.67.39,'2014-07-20 23:06:05.413000000','20
2c929293466b97a601475460b4047da6,M201370069,E0DB55ACB504,222.20.118.5,'2014-07-20 22:26:21.753000000','2
2c929293466b97a601475460b4397da8,U201211086,208984D919A3,222.20.49.33,'2014-07-20 23:14:48.283000000','2
2c929293466b97a601475460b5037dac,D201377912,E89A8F3D7456,218.197.252.167,'2014-07-20 20:32:24.337000000'
```