南京大学120周年校庆
120th ANNIVERSARY
NANJING UNIVERSITY
1902 - 2022

Weka实验

22秋《大数据分析》

南京大学120周年校庆
120th ANNIVERSARY
NANJING UNIVERSITY
1902 - 2022

CONTENT

目录

南京大学120周年校庆
120th ANNIVERSARY
NANJING UNIVERSITY
1902 - 2022

01 Weka与实验介绍

基本任务

使用weka工具包进行降维、分类、聚类3项任务的练习作业；实验代码用Java实现

Weka官网

https://www.cs.waikato.ac.nz/ml/weka/

Weka算法包介绍

package	description
bayes	a set of classification algorithms that use Bayes Theorem such as Naive Bayes, Naive Bayes Multinominal.
trees	Contains decision trees algorithms, such as Decision Stump and Random Forest.
function	a set of regression functions, such as Linear and Logistic Regression.
lazy	lazy learning algorithms, such as Locally Weighted Learning (LWL) and k-Nearest Neighbors.
meta	a set of ensemble methods and dimensionality reduction algorithms such as AdaBoost and Bagging to reduce variance.
misc	such as SerializedClassifier that can be used to load a pre-trained model to make predictions.
rules	Rules-based algorithms such as ZeroR.

南京大学120周年校庆
120th ANNIVERSARY
NANJING UNIVERSITY
1902 - 2022

02 数据预处理方法

数据预处理方法

1.缺失值处理 unsupervised.attribute.ReplaceMissingValues
使用均值和众数填充缺失值，默认跳过标签列；

2.标准化 Standardize
标准化给定数据集中所有数值属性的值到一个 0 均值和单位方差的正态分布。

3.规范化 unsupervised.attribute.Normalize
将所有数据通过数据变换，转换到指定范围内，参数为scale,translation,
转换后的范围为[translation,scale+translation],如scale=2.0,translation=-1,
则范围为[-1,1].

4. 离散化处理
supervised.attribute.Discretize 或 unsupervised.attribute.Discretize
将连续属性变为离散属性，例如将某个固定范围的浮点值，转变为3类确定值。

南京大学120周年校庆
120th ANNIVERSARY
NANJING UNIVERSITY
1902 - 2022

03 数据集及实验示例

数据集及实验示例

任务	方法	对应的Weka实现
降维 reduce demension	PCA和LDA，AdaBoost和Bagging	PrincipleComponent，LDA，AdaBoost 和 Bagging
分类 classification	决策树C4.5和朴素贝叶斯	J48 和 NaiveByes
聚类 clustering	K-means 和 EM	SimpleKMeans 和 EM

降维

降维：即用一组个数为 d 的向量 Zi 来代表个数为 D 的向量 Xi 所包含的有用信息，其中 d<D，通俗来讲，就是将高维度下降至低维度；将高维数据下降为低维数据。

降维的经典算法：PCA(principal component analysis)，LDA，SOM(self organized maps),AdaBoost and Bagging等。

降维

降维方法又分为线性降维和非线性降维，非线性降维又分为基于核函数和基于特征值的方法。

无监督降维算法不使用样本标签值，因此是一种无监督学习算法，其典型代表是PCA (主成分分析法, principle components analysis)；

有监督的降维算法则使用了样本标签值，因此是一种有监督学习算法，其典型代表是LDA (线性判别分析法, linear discriminant analysis)；

降维数据集

任务	数据集名称	样本大小	来源
降维	cpu.arff	209，属性个数7	weka
降维	titanic.arff	768，属性个数6	http://www.sc.ehu.es/ccwbayes/master/selected-dbs/supervised-classification/titanic.arff
降维	diabetes.arff	2200，属性个数6	weka

降维数据集介绍

eigenvalue	proportion	cumulative	
3.35674	0.55946	0.55946	-0.469MMAX-0.435CHMIN-0.429CACH-0.427MMIN-0.374CHMAX...
0.82936	0.13823	0.69768	0.682MYCT+0.559CHMAX-0.333MMIN+0.275CHMIN+0.152CACH...
0.73923	0.1232	0.82089	0.669MYCT+0.548MMIN-0.426CHMAX+0.264MMAX-0.03CHMIN...
0.49632	0.08272	0.90361	0.714CACH-0.477MMAX-0.436CHMAX+0.255CHMIN-0.088MMIN...
0.40442	0.0674	0.97101	0.812CHMIN-0.519CACH-0.226CHMAX-0.135MMAX-0.045MYCT...

数据集 cpu.arff 包含了CPU的不同属性，可以用PCA方式找到合适的属性维度V1-V5，同时保留的维度个数是可以通过Rank.setNumToSelect调整的。上图是cpu数据执行PCA的结果，其V1-V5数据可以作为后续任务的输入。

此外，diabetes.arff 和 titanic.arff 数据集的降维任务也是同理。

分类

分类是一个认识、区分和理解概念和对象的过程。常用的分类方法有贝叶斯，决策树，逻辑回归，最近邻，SVM等。

关于分类任务，Weka 有 J48(决策树 C4.5的实现版) ，NaiveBayes 等等这样的分类算法的实现方法。需要注意的是分类任务在Weka执行中，需要指定预测的标签类。

分类数据集

任务	数据集名称	样本大小	来源
分类	car_data.arff	1728	https://archive.ics.uci.edu/ml/datasets/Car+Evaluation
分类	adult_income_uk.arff	48842	https://archive.ics.uci.edu/ml/datasets/Adult
分类	spambase.arff spambase_test.arff	4601	http://archive.ics.uci.edu/ml/datasets/Spambase

分类数据集

(1) Car data(denote 1997)

目的：预测某类车的评价是好/坏

Class Distribution (number of instances per class)

class	N	N[%]	
unacc	1210	(70.023 %)	不可接受
acc	384	(22.222 %)	可接受
good	69	(3.993 %)	好
v-good	65	(3.762 %)	很好

有1728条数据。

决策树在此数据集上的正确率可以达到96.3%

属性为CAR(car acceptability)，
PRICE(overall price)，
buying(buying price)，
maint(price of the maintenance)，
TECH (technical characteristics)，
COMFORT(comfort)，
doors(number of doors)，
persons(capacity in terms of persons to carry)，
lug_boot (the size of luggage boot)，
safety(estimated safety of the car)，
标签类为CAR.

分类数据集

（2）数据集：adult_income_uk.arff

- 样本大小 ：48842（ 80% train 20% test ）
- Task: Prediction task is to determine whether a person makes over 50K a year.
- Description of attributes: https://archive.ics.uci.edu/ml/datasets/Adult

- 目标：观察测试数据集上的预测正确率

分类数据集

（3）数据集：spambase.arff

- 目标：采用朴素贝叶斯训练出模型，测试test数据集数据是否为垃圾邮件。
- 补充说明：给出的arff文件是由官网csv文件预处理过的，直接使用即可。

聚类

把相似的东西分到一组，采用计算相似度的算法即可进行，clustering 通常并不需要使用训练数据进行学习，也常被称为无监督学习。

The most popular clustering algorithms that WEKA offers are SimpleKMeans, HierarchicalClusterer, and EM.

K-means 方法

- The WEKA SimpleKMeans algorithm uses Euclidean distance measure to compute distances between instances and clusters.
- K-Means分群演算最大的问题在于如何决定分群的K。Weka中有改良的XMeans方法，可以在指定范围内找出合适的分群群数K，例如给定分群范围[5,10]，XMeans演算法可能决定的最佳分群群数为6。Weka 3.8没有自带XMeans，需要从Tools > package manager 安装XMeans。
- 最终产生的分群的每个群，群的质心的数值是连续数值的平均值，离散值的众数。

聚类数据集

任务	数据集名称	样本大小	来源
聚类	bmw-browsers.csv	100	http://learnersdesk.weebly.com/uploads/7/4/1/9/7419971/bmw-browsers.arff
聚类	iris.arff	150	https://archive.ics.uci.edu/ml/datasets/Car+Evaluation
聚类	bank.arff	600	https://learnersdesk.weebly.com/weka-tutorials.html

聚类数据集

（1）数据集：bmw-browsers.arff

- 属性：Dealership: 是否去过汽车经销店，0/1 表示 没有/有
  Showroom: 顾客是否有看展示间，0/1 表示 没有/有
  ComputerSearch：顾客是否使用电脑搜索，0/1 表示 没有/有
  M5: 顾客是否有看 BMW M5系车，0/1 表示 没有/有
  3 Series：顾客是否有看 BMW 3系车，0/1 表示 没有/有
  Z4: 顾客是否有看BMW Z4 敞篷跑车，0/1 表示 没有/有
  Financing: 顾客是否符合贷款资格，0/1 表示 没有/有
  Purchase: 顾客是否真实购车，0/1 表示 没有/有

- 下面设置k-means的群数为5，可以将数据集中的用户分为5类人群。

聚类数据集

Final cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(100.0)	(26.0)	(27.0)	(5.0)	(14.0)	(28.0)
Dealership	0.6	0.9615	0.6667	1	0.8571	0
Showroom	0.72	0.6923	0.6667	0	0.5714	1
ComputerSearch	0.43	0.6538	0	1	0.8571	0.3214
M5	0.53	0.4615	0.963	1	0.7143	0
3Series	0.55	0.3846	0.4444	0.8	0.0714	1
Z4	0.45	0.5385	0	0.8	0.5714	0.6786
Financing	0.61	0.4615	0.6296	0.8	1	0.5
Purchase	0.39	0	0.5185	0.4	1	0.3214

Cluster 0 可以称为“宝马dreamer”，他们有看车的行为，但是什么都不买；

Cluster 1：可以称为“M5 爱好者”，他们更关注M5，很少看3series和Z4，但是购买率只有52%。可以通过增加M5的车型和推销人员来提高销量。

Cluster 2：样本个数较少，可以忽略该群；

Cluster 3：该群体可以认为是BMW的拥趸，他们在门店，搜索对于各个系车都有查看，同时购买率极高，由于他们对M5和Z4的车查看率较高，可以增加M5和Z4在搜索网站的排名，来提高购买率；

Cluster 4：该类人群可以认为是BMW的入门客户，他们更倾向于看3系列，而不是更贵的M5，同时该类人群只有50%的人有贷款资格，可以考虑降低贷款标准或者降低3系价格，来提高购买率。

聚类数据集

		Cluster#		
Attribute	Full Data	0	1	2
	(150.0)	(50.0)	(50.0)	(50.0)
sepallength	5.8433	5.936	5.006	6.588
sepalwidth	3.054	2.77	3.418	2.974
petallength	3.7587	4.26	1.464	5.552
petalwidth	1.1987	1.326	0.244	2.026
class	Iris-setosa	Iris-versicolor	Iris-setosa	Iris-virginica

（2）数据集：iris.arff

设置分群数为3类，观察到KMeans方法可以正好分为3种类别的鸢尾属植物。
分别是 versicolor 杂色， setosa 刚毛，virginia类的鸢尾属植物。
可以观察到不同类别的一些属性特征。(sepal 萼片, petal花瓣)

当然除了K-Means，也可以用EM方法。

聚类数据集

		Cluster#					
Attribute	Full Data	0	1	2	3	4	5
	(600.0)	(77.0)	(76.0)	(77.0)	(147.0)	(106.0)	(117.0)
age	42.395	37.1299	44.2763	48.3117	39.1156	39.3019	47.6667
sex	FEMALE	FEMALE	FEMALE	FEMALE	FEMALE	MALE	MALE
region	INNER_CITY	INNER_CITY	RURAL	INNER_CITY	TOWN	INNER_CITY	TOWN
income	27524.0312	23377.7604	27772.3746	27668.4396	24047.3865	26359.8	35419.2842
married	YES	NO	YES	YES	YES	YES	NO
children	0	3	2	1	0	0	2
car	NO	NO	NO	NO	NO	YES	YES
save_act	YES	YES	YES	NO	YES	NO	YES
current_act	YES	YES	YES	YES	YES	YES	YES
mortgage	NO	NO	NO	NO	NO	YES	NO
pep	NO	NO	NO	YES	NO	YES	YES

（3）数据集： bank.arff

设置群数为6，从分群结果可以分析出不同人群的特点。

南京大学120周年校庆
120th ANNIVERSARY
NANJING UNIVERSITY
1902 - 2022

04 实验补充说明

Weka环境

1. JDK8，Weka 3.8

2.导入weka jar，需要有maven；没有maven需要手动导入jar。

```
<dependency>
    <groupId>nz.ac.waikato.cms.weka</groupId>
    <artifactId>weka-stable</artifactId>
    <version>3.8.6</version>
</dependency>
```

3. weka官方文档

https://waikato.github.io/weka-wiki/documentation/

Weka数据集文件I/O

1. 实验中Weka数据集文件为*.arff，其具体的文件规则可见官网
https://www.cs.waikato.ac.nz/ml/weka/arff.html
支持属性为 numeric, <nominal-specification>,string, date [<date-format>]

2. 读取数据集
.arff文件 -> instances -> for each instance

```
//方法一：使用DataSource类的read方法来加载arff文件
Instances data1 = DataSource.read( location: "data/weather.nominal.arff");

//方法二：使用直接指定加载器的方法来加载arff文件
ArffLoader arffLoader = new ArffLoader();//创建ArffLoader实例
arffLoader.setSource(new File( pathname: "data/weather.nominal.arff"));
Instances data2 = arffLoader.getDataSet();
```

arff文件格式介绍

arff文件是weka特有的一种数据格式，官网上的文件格式介绍如下。
https://www.cs.waikato.ac.nz/ml/weka/arff.html

1.数值类型属性
@attribute <attr_name> [numeric|integer|real]

2.时间类型属性
@attribute timestamp DATE "yyyy-MM-dd HH:mm:ss"

3.集合类型属性
@attribute sex {FEMALE,MALE} //{}内属性字母区分大小写，属性名和属性类型不区分大小写
@attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
@attribute married {NO,YES}

4.部分数据集中可能用?表示缺省值
@data
4.4,?,1.5,?,Iris-setosa