weka降维

201250203 陈张熠

一. 使用方法理解

我使用的是PCA方法

PCA是是一种线性降维算法，也是数据预处理方法。

通过计算数据矩阵的协方差矩阵，然后得到协方差矩阵的特征值特征向量，选择特征值最大的k个特征值所对应的特征向量组成的矩阵，再将数据矩阵转换到新的空间当中，实现数据降维。

下图为协方差矩阵计算公式

$$Cov(X,Y,Z) = \begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

本实验中主要使用了weka中四种方法：

1. Instances对象创建方法Instances(java.io.Reader reader):通过使用FileReader读取arff文件，并分配对每个实例分配一个权重
2. PrincipalComponents中的setMaximumAttributeNames (int value):设置使用PCA时留下最高的属性个数
3. PrincipalComponents中的buildEvaluator(Instances data):初始化PCA对象，并且对数据进行分析

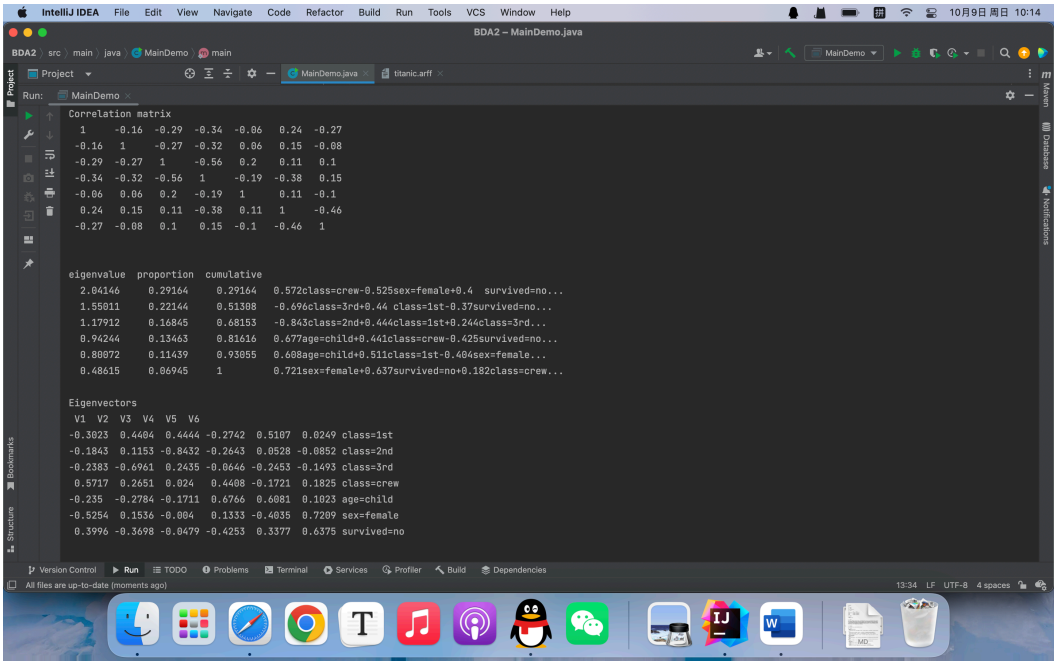
二. 数据集处理的思路

先通过创建FileReader对象，并让其读取对应的arff文件。

通过此Reader对象，创建Instances对象

最后通过通过PrincipalComponents中的setMaximumAttributeNames先设置留下属性的个数，再使用buildEvaluator进行分析

三. 实验结果


IntelliJ IDEA File Edit View Navigate Code Refactor Build Run Tools VCS Window Help
10月9日 周日 10:14
BDA2 – MainDemo.java
BDA2 src main java MainDemo main
MainDemo
Project
MainDemo.java
titanic.arff
Run: MainDemo
Correlation matrix
1 -0.16 -0.29 -0.34 -0.06 0.24 -0.27
-0.16 1 -0.27 -0.32 0.06 0.15 -0.08
-0.29 -0.27 1 -0.56 0.2 0.11 0.1
-0.34 -0.32 -0.56 1 -0.19 -0.38 0.15
-0.06 0.06 0.2 -0.19 1 0.11 -0.1
0.24 0.15 0.11 -0.38 0.11 1 -0.46
-0.27 -0.08 0.1 0.15 -0.1 -0.46 1
eigenvalue proportion cumulative
2.04146 0.29164 0.29164 0.572class=crew-0.525sex=female+0.4 survived=no...
1.55011 0.22144 0.51308 -0.696class=3rd+0.44 class=1st-0.37survived=no...
1.17912 0.16845 0.68153 -0.843class=2nd+0.444class=1st+0.244class=3rd...
0.94244 0.13463 0.81616 0.677age=child+0.441class=crew-0.425survived=no...
0.80072 0.11439 0.93055 0.608age=child+0.511class=1st-0.404sex=female...
0.48615 0.06945 1 0.721sex=female+0.637survived=no+0.182class=crew...
Eigenvectors
V1 V2 V3 V4 V5 V6
-0.3023 0.4404 0.4444 -0.2742 0.5107 0.0249 class=1st
-0.1843 0.1153 -0.8432 -0.2643 0.0528 -0.0852 class=2nd
-0.2383 -0.6961 0.2435 -0.0646 -0.2453 -0.1493 class=3rd
0.5717 0.2651 0.024 0.4408 -0.1721 0.1825 class=crew
-0.235 -0.2784 -0.1711 0.6766 0.6081 0.1023 age=child
-0.5254 0.1536 -0.004 0.1333 -0.4035 0.7209 sex=female
0.3996 -0.3698 -0.0479 -0.4253 0.3377 0.6375 survived=no
Version Control Run TODO Problems Terminal Services Profiler Build Dependencies
All files are up-to-date (moments ago)
13:34 LF UTF-8 4 spaces