**《大数据分析》第三次实验报告**

201250070郁博文

1. **PCA**

**1．PCA简介**

PCA（主成分分析）是一个常用的数据降维的方法。使用该方法可以快速找出原数据集中较为重要的属性。

**2．个人理解**

假设我们要对一个三维的数据集进行降维。首先，我们需要对数据集进行去中心化，具体就是计算每个维度上的数据均值，然后将每个维度上的数据减去该均值。在该三维空间中，就是先找到原始数据的均值的那个点，把所有数据进行平移，使得该均值点与坐标系原点重合。这一步是为了后续计算方差等数据时，数据是一个被规约化的数据，可以拿来直接计算。也为了防止主成分计算错误。接下来是计算协方差矩阵，然后通过SVD（奇异值分解）的方法来计算协方差矩阵的特征值和特征向量。这一步比较抽象，我对于这一步的理解如下：

使用上文中去中心化的三维坐标系。我们要做的就是将坐标系中的点（实例）进行拟合。想象有一根穿过坐标轴原点的线将坐标系中的点投影到这根线上，计算投影点与原点之间的距离的平方。然后以此类推，计算坐标系中所有点的这个距离的平方。然后旋转这根线，重复这个过程，直到找到该距离平方和最大的线，这条线就是新的主成分，换而言之就是新的坐标轴。之后，在这根线的正交平面上继续重复上述过程，直到找到全部的两两正交的这样的线。这个过程实际上也是计算方差的过程。我们之所以要让方差最大，实际上是因为，我们要让坐标系中的点投影到新的主成分坐标轴上后尽可能分散，而方差越大这个分散程度就越大。事实上，距离平方和就是PC1（主成分1）的特征值，该特征值的平方根就是PC1的奇异值，PC1的方向向量就是PC1的特征向量，使用这个特征向量就可以计算出PC1关于原维度的公式。

如果我们把PC1的特征值除以维数-1，就得到了方差解释率，这个数值的意义在于告诉我们该主成分在占有的信息差异。一个PC对应一个方差解释率。如果我们把每个PC的方差解释率加起来就是累计方差解释率。通常累计方差解释率在95%以上就可以反映原数据信息了，也就是原数据适合降到几维。描绘方差解释率的图表叫碎石图，另一种说法是碎石图趋于平缓时，那个主成分数就是合适的。

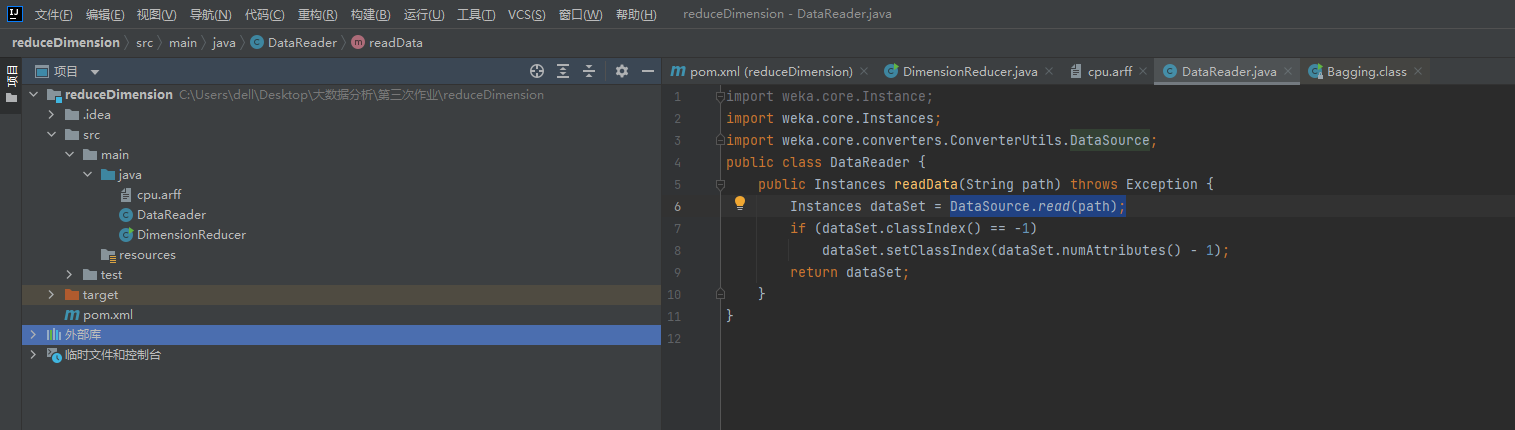
以上步骤对应到数学求解中其实十分简单，就是对协方差矩阵就行奇异值分解求解特征值和特征向量，然后把特征值从大到小排序，选择最大的K个。只不过理解这一步花费了笔者很大精力

1. **实验结果**

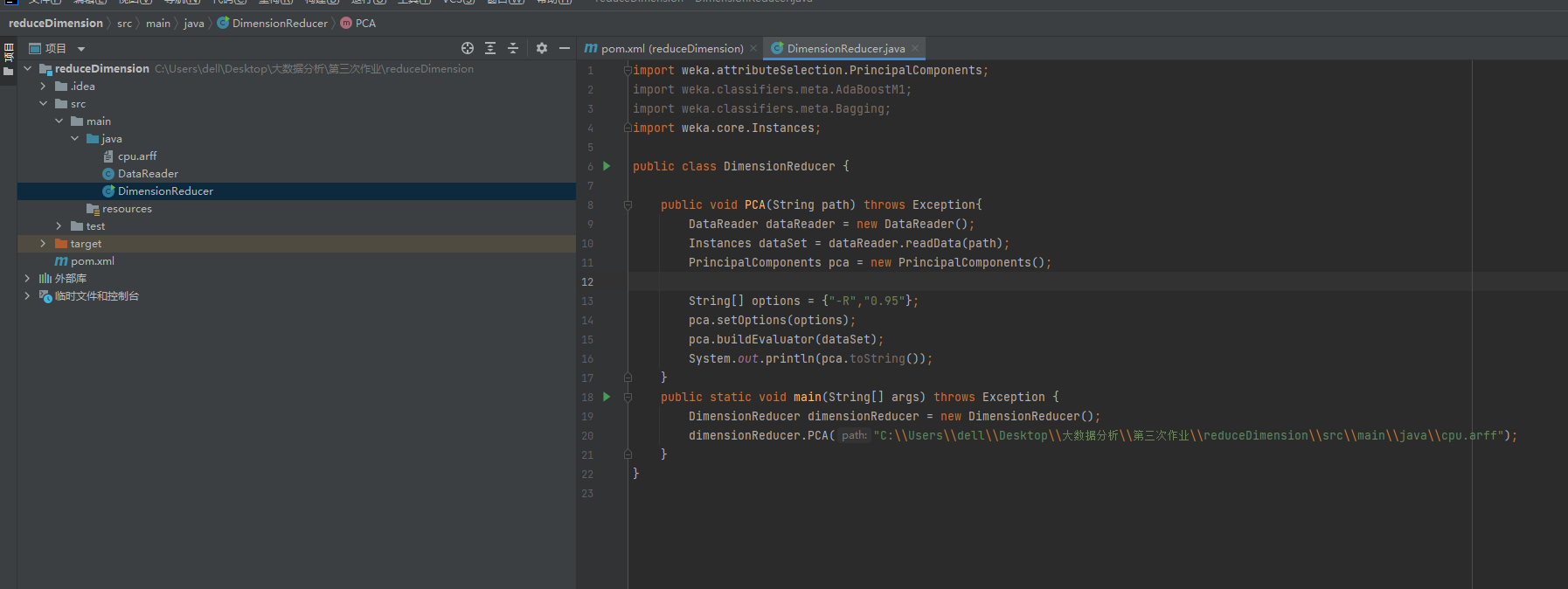
编写的实验代码如下：

这个类封装了数据集文件的读写

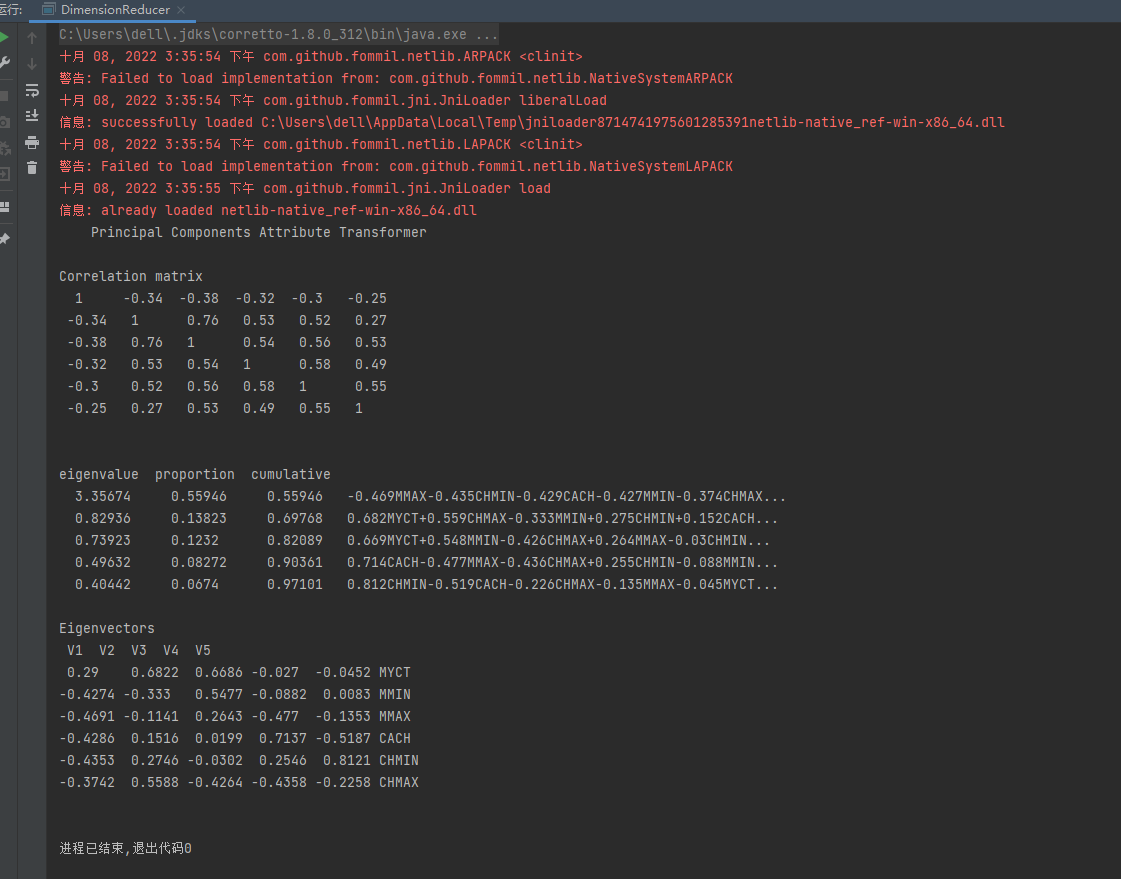
原数据集的最后一个属性是用于分类的，因此读入数据时需要特殊标记，否则会对结果产生影响。



这个类是PCA的实现。Weka的PCA方法提供了参数，可以自行修改累计方差解释率的阈值，我们这里采用默认值0.95.



运行结果如下：



第一张表就是协方差矩阵，第二张表就是特征值和方差解释率。最右边一列就是每个主成分的计算公式，该公式基于第三章的特征向量表格。