第三次作业-数据降维 实验报告

1. 对选择的方法的理解

我选择的降维方法是主成分分析（Principal Component Analysis)。

主成分分析是一种常用的无监督学习降维算法。这一方法利用正交变换把由线性相关的变量表示的观测数据转换为少数几个由线性无关变量表示的数据，线性无关的变量称为主成分。

在主成分分析中，首先对给定的数据进行标准化，使数据每一变量服从平均值为 0，方差为 1 的分布。之后再对数据做正交变换，原来由线性相关变量表示的数据，通过正交变换变成若干个由线性无关变量表示的数据。新变量是可能的正交变换中变量的方差的和（信息保存量）最大的。 将新变量依次称为第一主成分，第二主成分等。通过主成分分析，可以用主成分近似地表示原始数据，这就是主成分分析降维的基本思想。

2. 对数据集的分析和处理思路

2.1 数据集简介

数据集是关于 cpu 的信息，一共有 7 维，去除一个 label，有六维。

2.2 数据集处理思路

1. 先去除缺失值，由于数据集比较小，所以我人眼看了一遍发现没有缺失值，故跳过这一步。
2. 对数据做标准化，将数据变换到平均值为 0 方差为 1.
3. 运行主成分分析算法。

3. 实验结果

```
    Principal Components Attribute Transformer

Correlation matrix
  1     -0.34  -0.38  -0.32  -0.3   -0.25
 -0.34   1      0.76   0.53   0.52   0.27
 -0.38   0.76   1      0.54   0.56   0.53
 -0.32   0.53   0.54   1      0.58   0.49
 -0.3    0.52   0.56   0.58   1      0.55
 -0.25   0.27   0.53   0.49   0.55   1


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...

Eigenvectors
 V1  V2  V3  V4  V5
 0.29    0.6822  0.6686 -0.027  -0.0452 MYCT
-0.4274 -0.333   0.5477 -0.0882  0.0083 MMIN
-0.4691 -0.1141  0.2643 -0.477  -0.1353 MMAX
-0.4286  0.1516  0.0199  0.7137 -0.5187 CACH
-0.4353  0.2746 -0.0302  0.2546  0.8121 CHMIN
-0.3742  0.5588 -0.4264 -0.4358 -0.2258 CHMAX
```

默认保留的 proportion 是 0.95，可以看到在这个设置下维度从 6 降到了 5，且保存了大部分的信息。

参考资料

1. weka 官方文档
2. 《统计学习方法》. 李航