**运用主成分分析法选择投资的学校**

**摘要**

本文利用题目已给的数据，运用主成分分析法构建了一个对学校投资优先级排名的模型，给出了投资学校的优先顺序。我们的模型分为两个部分：

第一部分，我们进行了对数据的人工筛选及处理。首先根据数据类型提取出整型和浮点型数据，以及学校的名字和其所在的州；然后对其所在州进行编号，使其转化为整型变量。

第二部分，我们运用主成分分析法为所有学校投资优先级进行评分并排序。首先我们先对缺失数据进行补全或弃用；然后利用主成分分析法为各所学校给出投资优先级评分。

**关键词：投资优先级、主成分分析法**

1. **问题重述**

GooGrand基金会是一个慈善机构，旨在帮助提高美国大学生的教育表现。为此，基金会打算从2016年7月开始，每年捐赠100000000美元（相当于一亿美元）到一个合适的学校，为期五年。

GooGrand基金会要求你的团队开发一个模型来确定最佳的投资策略，需要给出各个学校投资优先级的排名，以保证基金会能有最好的回报率。

1. **问题分析**

问题是要我们给出该基金会投资学校奖学金优先级的排名。要得到这个排名，我们要综合考虑学校的地理位置（所在州县）及数据资料中所给的各项目要求。为此，我们采用主成分分析发来为各个学校投资优先级打分，给出排名。

主成分分析[1]也称主分量分析，旨在利用降维的思想，把多指标转化为少数几个综合指标。在统计学中，主成分分析是一种简化数据集的技术。它是一个线性变换。这个变换把数据变换到一个新的坐标系统中，使得任何数据投影的第一大方差在第一个坐标(称为第一主成分)上，第二大方差在第二个坐标(第二主成分)上，依次类推。主成分分析经常用减少数据集的维数，同时保持数据集的对方差贡献最大的特征。这是通过保留低阶主成分，忽略高阶主成分做到的。这样低阶成分往往能够保留住数据的最重要方面。但是，这也不是一定的，要视具体应用而定。

在本题中，由于所给数据中有很多残缺数据，因此要使用主成分分析发需要对残缺数据进行补全或适当舍去。之后在利用主成分分析法就可给出投资优先级评分并给出排名。

1. **模型假设**
2. 不管学校是否受到投资，这些学校的统计资料在五年内保持不变，及保证投资优先级排名不会变。
3. 排除自然因素或外界社会因素对学校的影响，及不考虑学校突然倒闭这种情况的发生
4. **变量说明**

表格 1 变量说明

|  |  |
| --- | --- |
| **符号** | **含义** |
|  | 学校编号 |
|  | 变量编号 |
|  | 第个变量中缺失数据的不充值 |
|  | 补充数据前拥有第个变量数据的学校个数 |
|  | 学校总数 |
|  | 人工筛选后剩下的变量总数 |
|  | 学校的变量的值 |

1. **模型建立和求解**
   1. **数据的人工筛选及处理**

首先，我们对题目所给数据进行人工筛选，仅保留下学校名称、学校所在州县及各个为整型或浮点型的变量，具体数据见附件“beforeNULL.txt”。其次，我们为各个州进行了编号，将浮点型变量转化为整型变量，详见表格2。

表格 2 州县编号

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **州** | **编号** | **州** | **编号** | **州** | **编号** | **州** | **编号** |
| **AL** | 1 | **IN** | 14 | **NE** | 27 | **SC** | 40 |
| **AK** | 2 | **IA** | 15 | **NV** | 28 | **SD** | 41 |
| **AZ** | 3 | **KS** | 16 | **NH** | 29 | **TN** | 42 |
| **AR** | 4 | **KY** | 17 | **NJ** | 30 | **TX** | 43 |
| **CA** | 5 | **LA** | 18 | **NM** | 31 | **UT** | 44 |
| **CO** | 6 | **ME** | 19 | **NY** | 32 | **VT** | 45 |
| **CT** | 7 | **MD** | 20 | **NC** | 33 | **VA** | 46 |
| **DE** | 8 | **MA** | 21 | **ND** | 34 | **WA** | 47 |
| **FL** | 9 | **MI** | 22 | **OH** | 35 | **WV** | 48 |
| **GA** | 10 | **MN** | 23 | **OK** | 36 | **WI** | 49 |
| **HI** | 11 | **MS** | 24 | **OR** | 37 | **WY** | 50 |
| **ID** | 12 | **MO** | 25 | **PA** | 38 | **DC** | 51 |
| **IL** | 13 | **MT** | 26 | **RI** | 39 | **PR** | 52 |

* 1. **缺失数据的处理**

由于所提供的数据集遭受严重的数据丢失可能破坏结果的可靠性和可解释性，因此为了解决这一问题，我们采用两种不同的方法对不同丢失率的数据进行处理。

1. 为了简单起见，我们省略了些数据缺失率超过或等于50%的变量，即不考虑它对投资优先级评分的影响。所剩的变量见表格3。

表格 3 需要考虑的变量名

|  |  |  |
| --- | --- | --- |
| INSTNM | CODE | HCM2 |
| TRIBAL | AANAPII | HSI |
| PCIP09 | PCIP10 | PCIP11 |
| PCIP22 | PCIP23 | PCIP24 |
| PCIP38 | PCIP39 | PCIP40 |
| PCIP47 | PCIP48 | PCIP49 |
| UGDS\_WHITE | UGDS\_BLACK | UGDS\_HISP |
| PPTUG\_EF | CURROPER | NPT4\_PRIV |
| GRAD\_DEBT\_MDN\_SUPP | GRAD\_DEBT\_MDN10YR\_SUPP | RPY\_3YR\_RT\_SUPP |
| PREDDEG | CONTROL | LOCALE |
| NANTI | MENONLY | WOMENONLY |
| PCIP12 | PCIP13 | PCIP14 |
| PCIP25 | PCIP26 | PCIP27 |
| PCIP41 | PCIP42 | PCIP43 |
| PCIP50 | PCIP51 | PCIP52 |
| UGDS\_ASIAN | UGDS\_AIAN | UGDS\_NHPI |
| NPT41\_PRIV | NPT42\_PRIV | RET\_FT4 |
| md\_earn\_wne\_p10 | gt\_25k\_p6 | UG25abv |
| HBCU | PBI | ANNHI |
| PCIP03 | PCIP04 | PCIP05 |
| PCIP15 | PCIP16 | PCIP19 |
| PCIP29 | PCIP30 | PCIP31 |
| PCIP44 | PCIP45 | PCIP46 |
| PCIP54 | DISTANCEONLY | UGDS |
| UGDS\_2MOR | UGDS\_NRA | UGDS\_UNKN |
| RET\_FTL4 | 108题 |  |

1. 对于数据缺失率在50%以下的变量，我们通过简单地用已有值的均值替换缺失值来解决数据缺失问题，即令

补全后的具体数据详见附件“afterNULL.txt”。

* 1. **主成分分析[2]**
     1. **标准化**

利用公式

不全后数据进行标准化处理，得到标准化矩阵，其中

* + 1. **相关系数矩阵**

求标准化矩阵 的相关系数矩阵

其中

* + 1. **主成分**

求解5.3.2中的相关系数矩阵 的特征方程，得到 个特征根，确定主成分。

根据不等式

来确定 的值，使信息的利用率达到85%以上。然后对每个，解方程组得到单位特征向量。

将标准化后的指标转化为主成分

其中 称为第一主成分， 称为第二主成分，……， 称为第主成分。

* + 1. **主成分分析评分**

对 个主成分进行综合评价。对 个主成分进行加权求和，即得最终评价值，权数为每个主成分的方差贡献率。

最终，我们给出了学校的投资有限级顺序，详见附件“”。这里给出前十名，见表格4。

[插入表格]

1. **结论**

学校投资优先级排名详见附件“”，前十名见表格4。

1. **参考资料**

[1] https://blog.csdn.net/zhang\_alongzd/article/details/51674232

[2]https://blog.csdn.net/MATLAB\_matlab/article/details/59483185?locationNum=10&fps=1