2016C The Goodgrant Challenge

问题描述

此问题旨在对美国多所高校进行量化评分，确定各所学校有效利用私人资金的潜力和估计投资回报率，以帮助goodgrant基金会确定最佳的投资策略。

数据处理

本问题数据较为复杂，且数据缺省值较多，同时，也存在数据错误的情况，纵观各优秀论文，其往往采用以下策略：

1.删除数据缺少较多的学校，认为其并不希望获得资助。

2.通过拟合、聚类等方式补充缺省值，但与此同时，本问题缺少时间序列数据，拟合较为困难。

个人对数据处理的想法：

1.通过未缺省的数据进行聚类（k-means,BP,模糊综合聚类），通过取每一类的平均值进行处理。

2.同时，对各项指标进行相关性分析（皮尔森相关系数/相关系数矩阵）的方法，使用高相关度的数据进行拟合，同时，可以与聚类数据进行对比分析。（高相关度数据处理意味着数据可以降维，此处缺省值的补充基本是为降维做准备）

同时，各个优秀论文普遍将数据分为两类：即一类用于选择学校，一类用于确定投资回报率。

量化模型的建立

指标选取

针对量化模型的指标选取，本题优秀论文采取了如下策略：

1.PCA

2.基于Lasso回归的模型选取（此方法较为新颖，但问题在于如何处理y）

学校选择模型与投资回报率的确定

1. AHP确定各指标权重或使用PCA给出的权重
2. 归一化计算综合得分

投资规模的确定

1.通过人为确定的ln函数方程/一些特殊的方程确定

2.通过收集多年的相关指标的数据，建立时间序列模型。优秀论文多采用灰色预测模型（受限于数据量较少）

3.建立多目标规划模型，并求解多目标规划问题

本题的难点

1.数据量较大，且数据特征较多，初始的人工处理较为复杂

2.为确定投资规模，需要自行寻找相关数据，或者使用某些特定的函数

本题关键

本题最主要的问题是将问题简化到可以处理的程度，以解决数据缺省较多，数据没有时序特征等问题