Statistics and Application 统计学与应用, 2018, 7(4), 445-449
Published Online August 2018 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2018.74051

Hans汉斯

The Statistical Analysis and Demonstration of Evaluation Scheme for Undergraduates in College

Yuan Zhou

Central China Normal University, Wuhan Hubei
Email: 1053610875@qq.com

Received: Aug. 2nd, 2018; accepted: Aug. 20th, 2018; published: Aug. 27th, 2018

Abstract

This project investigated the opinions and suggestions on current evaluation scheme among undergraduates in CCNU via questionnaire, and then quantized the importance of related indices in order to gain the average satisfaction of each index in students' minds. On the other hand, for the sake of scientificity, the raw data of the undergraduates enrolled in 2014 in the school of mathematics and statistics were chosen and handled, with such statistical methods as factor analysis, principal component analysis and cluster analysis. By comparing the advantages and disadvantages of different methods, the final scheme and name list of evaluation based on raw data will be given. This research production can be naturally extended to all the data-based evaluations.

Keywords

Evaluation, Principal Component Analysis, Factor Analysis, Cluster Analysis

对高校学生评优评先方案的统计学分析与论证

周　源

华中师范大学，湖北 武汉
Email: 1053610875@qq.com

收稿日期：2018年8月2日；录用日期：2018年8月20日；发布日期：2018年8月27日

摘　要

本项目首先通过问卷调查的方式，调查了华中师范大学本科学生对现行评优评先方案的满意度和建议，

文章引用: 周源. 对高校学生评优评先方案的统计学分析与论证[J]. 统计学与应用, 2018, 7(4): 445-449.
DOI: 10.12677/sa.2018.74051

并将评优评先各指标的重要性进行量化，以期得到各指标在学生认知中的平均满意度。另一方面，为保障三好学生名单的科学性，本项目基于2016年数学与统计学学院2014级本科生三好学生评选原始数据，采用因子分析，主成分分析，聚类分析等几种统计学方法，将原始数据处理，得到了完全基于数据本身得到的三好学生名单。通过比较不同方法的优劣，最终给出了基于数据的三好学生评选的方法和名单。这个结果可以很自然地进行推广，用于一切基于数据的荣誉评选中。

关键词

评优评先，主成分分析，因子分析，聚类分析

1. 引言

通过调查发现，现今在高校、中小学以及企业内等，对荣誉称号(或奖金奖学金等)的评定往往都是基于量化数据的加权和来决定的。相关评选工作仍处于起步阶段，相关评价指标模糊且粗糙[1]。实际上，对于各类量化数据而言，其权重的值是由管理者直接给出的，在一定程度上缺乏客观性和科学性，群众对其评选方案的满意度也不一定高。如何才能使得评选方案更加合理和科学，这是值得大家思考的。本文通过收集一定的真实可靠地数据，利用统计学的思想和方法来探讨这种评优评先方式的合理度。

本文将以华中师范大学数统学院本科生的三好学生评选方案为例，对此问题进行较为具体的分析和论证。一方面，本项目通过调查问卷来比较全面地获取同学们对该评优评先方案的满意度和建议，在分析数据的同时结合这些问卷中的因素进行全面考虑，从而使得评优评先方案更接近同学们的期望，在一定程度上保证评选方案的公平和同学们的满意度。另一方面，本项目仅仅依赖于各类量化数据的数值，除了通过主成分分析、因子分析、聚类分析(包括层次聚类、k 均值聚类和密度聚类)传统方法之外，还利用了较为前沿的机器学习理论，得到了更加客观的三好学生名单。这些方法可以用于改进三好学生评选过程，更一般地，也可以用于一切基于数据的荣誉评选。

2. 研究分析过程

2.1. 调查问卷分析

项目组通过实体问卷和网络问卷两种方式，随机搜集了华中师范大学本科生总计 372 人的数据，基本涵盖了各个院系、各个年级的学生，保证了数据来源科学客观。

问卷结果显示，学生对现行评优评先方案的满意度仍有提升空间，并且更多的人更加获选者个人的学术能力，这体现在：

①理想中的学分绩平均占比达到 68%，同时有四分之一以上的人认为学分绩应该占比 80%

②科研创业、获奖情况、发表文章论文等重要性相较学生干部加分、志愿活动加分、认证考试加分(问卷中的认证考试指大学英语四、六级考试、计算机等级考试等)更高。

③同时，有 59.5%的学生认为，挂科导致三好学生评比一票否决制度不合理。

基于上述调查情况，学生对现行评优评先方案满意度仍有提升空间。本项目将会把学生的相关意见纳入考虑，并以 2016 年 9 月数学与统计学学院 2014 级本科生数学师范专业的三好学生数据进行分析，

试题构建更合理的解决方案。

2.2. 主成分分析

对标准化以后的数据，使用 R 中的 princomp 函数，得到了数据集中的主成分。通过忽略高阶主成分，对数据进行了降维处理，保留的低阶主成分对原始数据的方差贡献率已经达到 0.9661639，较为充分地保留了原始数据的信息。对保留的主成分，将其主成分得分依据其方差贡献率进行了线性加权，得到了每个学生最终的主成分得分。进而可以通过排序得到基于主成分得分的三好学生名单。

在原来的考核标准里，学分绩是十分重要的一个指标。但是通过主成分分析得到的主成分中，有些与学分绩无关，有的甚至与学分绩呈现负相关。综合考虑各异常项，都与社会实践、志愿服务、学生干部、班团荣誉等与社交能力相关的因素有较强的正相关关系。社交能力在现实社会中发挥着很重要的作用，但是在评优评先系统中，却总是将学分绩摆在第一位，而忽略了学生其他的能力。

按照主成分分析的结果，学分绩所占的重要性降低，这与问卷调查过程中，学生普遍认为学分绩更为重要的满意度不符合，故导致最后加权结果意义不大。

学分绩相关的主成分贡献率过低是这是由于原始数据中只包含了学分绩达到评选下限的学生的数据，缺少学分绩较低、没有参评资格的学生，导致了学分绩的方差较小，从而和学分绩相关的主成分贡献率较小。因此想要解决这一困境，最佳的方法就是搜集参评范围所有学生的全部数据。

2.3. 因子分析

使用 R 中的 factanal 的函数，根据样本数据给出方差最大的载荷因子矩阵，得到了对原始数据贡献率最大的五个因子，五项因子的贡献率分别为 0.130，0.125，0.120，0.107 和 0.072，和主成分分析的想法类似，将因子贡献率作为权重导出因子得分的加权和。进而可以对每个学生的因子得分进行排序，得到基于因子分析的三好学生名单。

容易发现，在名单中较靠前的学生其因子得分同样较高，说明按照教务处评出的三好学生名单相对合理。同时，各因子贡献率都接近，可以避免因为数据缺失导致的结果不科学或不显著。通过得到的因子加权得分，能够对所有参评的学生进行排序，得到一份较科学评优评先的名单。但是，数统学院评优评先的前期过程中，并没有获取每一位学生的信息，而是只统计了上报学生的数据。这对本项目也产生了一定的负面影响。

2.4. 聚类分析

在解决实际问题过程中，将多样本对象分类时，依据单因素分类不足以全面综合的描述其类别，往往要考虑多方面因素进行分类[2]。在面对缺少学分绩较低的学生数据的情况下，可以考虑对有资格参评的学生进行聚类分析，期望得到能获评三好学生的类别与不能获评的类别。主要方法有层次聚类，K 均值聚类和密度聚类三种。

2.4.1. 层次聚类

按照层次聚类的结果，我们能将所参评的 89 个同学分成两组，分别有 46 和 43 个同学。由于在问卷调查中学生更侧重于学分绩、个人学术能力，因此我们计算这两类的平均得分，发现第一类平均学分绩显著高于第二类的平均学分绩。由此我们可以认定第一类为三好学生名单。

2.4.2. k 均值聚类

考虑使用 kmeans 方法聚类数据集，由于前面层次聚类方法得到的结果，因此初步取簇数为四，得到了四个类别的学生，并调用 barplot 函数绘制了每个簇中心的条形图，同时绘制了簇散点图，对于不同簇

的观测点使用不同的颜色进行可视化处理。观察这些数据点可以发现相同颜色的数据点在大多数平面内的投影都大致集中在一个区域内，可以认为这种聚类方法在处理这些数据上具有很好的说服力。由于变量的个数大于两个，因此无法通过数据图像来展示聚类的过程。此时，可以通过使用二元聚类图来将变量减少为两个主成分，再次利用组件来展示聚类的过程，方便更好地理解各个簇内数据点的特征。

在后期优化过程中，考虑到取簇数为四操作较为复杂而且并不完全准确，因此，可以使用距离平方和来确定哪一个 k 值能够得到最好的 k 均值聚类效果。进一步得到了优化的 K 均值聚类，导出了基于 K 均值聚类的两类学生名单。与层次聚类相同，通过比较平均学分绩，第一类学分绩显著高于第二类，因此第一类学生是通过 K 均值聚类得到的三好学生名单。

2.4.3. 密度聚类

由 R 语言 FPC 包中的 kNNdistplot 函数，观察函数的拐点，得到了最优的 EPS 值，并使用 dbscan 函数，导出了数据点所属的簇类。进一步分析得到了基于密度聚类的三好学生名单。

通过密度聚类，数据被划分成三类。第一类有 21 个样本，这 21 个在现实中都被评为了三好学生；该样本第二类有 65 个样本；第三类是数据的噪点，分别是原始数据中的第 1，第 14，第 26 个样本，他们在现实中都被评为了三好学生。但是深入分析可知，第一个样本在加分总和上很突出，第 14 个样本在学生干部方面很突出，第 26 个样本在论文方面很突出，因此在算法运行过程中这三个样本被当作噪声处理，但实际上，这三名学生表现都很优异，因此仍将 1，14，26 号作为三好学生样本。

2.4.4. 聚类方法的评估

官方给出的评选名单是 1~55 号学生。将聚类所得的名单两两比对(表 1 中的数字代表在两种方法得到的名单中均出现的人数)。

通过分析发现，三种聚类方法以及原始名单尽管存在总人数差异，但上述对比结果均取得了极高的相似度。因此可以认为，通过上述聚类分析得到的方法是充分合理的，而总人数会有略微的差异，可以根据不同的三好学生指标数量选择不同的聚类方法进行。

3. 结论

上述通过聚类分析得到的名单是只基于原始数据产生的，不同名单之间相似度均很高，针对不同的三好学生指标数可以通过上述统计方法得到不同数量的名单结果进行三好学生评选，充分保证科学性和客观性，同时在一定程度上符合了大部分学生重视学分绩的观念(体现在处理噪点和确定分出的类别哪一类是三好学生上)。

这样的过程只需简单更改程序和导入数据就可以基于不同的基础数据得到不同的三好学生名单，具有普适性和一般性。针对噪点和类别的确定，同样可以根据不同学校、公司、企业的侧重点，按照学分绩、个人能力优先等进行操作。

Table 1. Comparison among the lists given by different cluster analysis methods
表 1. 各聚类方法所得名单的对比

	层次聚类	k 均值聚类	密度聚类	原始名单
层次聚类	46	30	24	45
k 均值聚类	30	32	22	32
密度聚类	24	22	24	24
原始名单				55

但是在项目中仍然存在影响结果的重要人为因素——原始数据的评分，包括学分绩的评定(不同老师平时成绩给分高低差异较大)，加分项的加分比例等因素，这些因素的解决仍然需要进一步进行分析。

参考文献

[1] 赵静. 基于绩效评估的高校教师评先评优机制研究[J]. 长春教育学院学报, 2016, 32(8): 44-46.

[2] 公丽艳, 孟宪军, 刘乃侨, 毕金峰. 基于主成分与聚类分析的苹果加工品质评价[J]. 农业工程学报, 2014, 30(13): 276-285.