

**计算机教研组课程考试论文**

基于度量学习与聚类分析方法的

社会分层研究初探

课 程 名 称： 机器学习 。

专 业： 社会学 。

学 号： 20181052113 。

姓 名： 李凌浩 。

成 绩： 。

阅卷教师签字： 。

年 月 日

基于度量学习与聚类分析方法的

社会分层研究初探

社会分层（Social Stratification）是对社会系统结构化不平等的描述，它反映的是社会上各类物质性和象征性的资源在不同的人当中的分布情况[[1]](#footnote-1)。社会学研究通过这一角度来观察社会结构、描述资源分配形态和评估社会不平等程度。传统的社会分层研究属于经验社会学范畴，主要依靠定性和定量研究方法来对社会资源的不同部分进行考察，依靠社会学家自己的理论背景提出不同的分层维度，如个人收入、受教育水平、个人财产等等，这是得以进行进一步研究的前提。但近年来，统计学上的度量学习，以及能够进行无监督学习的聚类分析方法的出现，可能会为社会分层研究提供新的思考进路。本文试图就以上两种统计学习方法，结合大样本抽样调查数据，在社会分层结构确立方面做出尝试和初步探索。

在已有的中文文献中，从社会地位角度研究其与政治参与关系的实证研究尚不充分，而西方学界普遍相信对社会地位尤其是受教育程度因素与政治参与的正相关关系。但国内外学界对政治参与与社会地位作为两个独立概念的研究已经相当充分且在不断增多。对社会地位，国内外学者大多将其看作社会对某个人或某个群体的综合性价值评价，并将其作为社会分层的其中一个维度（李春玲，2005；李培林，1996），而社会地位的测量在西方学界主要以运用职业声望及职业对应的平均收入和教育水平求出职业声望得分作为社会经济地位指数进行声望分层测量的方式（Duncan，1961；Blau & Duncan，1967）为主流，在国内则依据改进的职业声望和社会经济地位指数计算公式作为较普适的标准来衡量个人或群体的社会地位，同时，也更强调了权力和工作单位因素对中国人的声望地位的重要影响（李春玲，2005：98-99、101）。

二、研究设计

本文选取中国社会科学院中国社会状况综合调查（Chinese Social Survey，CSS）数据展开分析。对社会地位与政治参与概念的操作化将根据CSS问卷设计与数据结构的特征进行调整。

（一）社会地位的测量

出于社会地位测量的全面性，本文选取李春玲对社会地位的研究结果，根据职业声望得分作为衡量社会地位高低的参数，其获得的回归方程为：

职业声望Y=11.808+3.349\*平均教育年限+0.573\*平均月收入(百元)+16.075\*是否是最高管理者+11.262\*是否是中层管理者+3.738\*是否是基层管理者+8.942\*是否在党政机关工作+6.841\*是否在事业单位工作-5.694\*是否在企业单位工作-26.655\*是否从事受歧视职业

该回归方程是对影响人们职业声望评价的因素的解释。根据这些因素，可以在CSS问卷中找到对应问题及变量（具体问题题干、题肢、选项及操作化处理可参见附表1）。对比回归方程与CSS问卷对应变量的特征，本文对该方程做出如下调整：

1. 将方程中的“平均受教育年限”与“平均月收入”两项替换为CSS问卷调查中对被访者学历水平与年收入的有关数据，对于未上学的被访者，教育年限按0年计；小学学历的被访者，教育年限按6年计；初中学历的被访者，教育年限按12年计；高中学历的被访者，教育年限按15年计；中专学历的被访者，教育年限按16年计；职高技校学历的被访者，教育年限按15年计；大学专科学历的被访者，教育年限按20年计；大学本科学历的被访者，教育年限按19年计；研究生学历的被访者，教育年限按22年计；

对于被访者的年收入，主要从事非农工作的被访者，以其主要从事的非农工作的平均月收入作为被访者的平均月收入；主要从事务农工作的被访者，以其个人年总收入按月平均作为被访者的平均月收入；

2. 基于CSS问卷有关变量对企业事业单位中管理者的分类，将方程中“是否是最高管理者”的权重对应CSS问卷题B4e中选择选项1的被访者；方程中“是否是中层管理者”、“是否是基层管理者”两项的权重合并，对应选择题目选项2的被访者；选择题中选项3的被访者此项不得分；

实际上，从方程系数上看来，管理者层级高低，实际上对职业声望指数的影响并不是线性的，从基层管理者到中层管理者对指数的影响程度有一个跃迁。故此对方程中“是否是中层管理者”、“是否是基层管理者”两项权重的合并方法，本文采用将变量数值化并运用最小二乘法进行非线性回归拟合的方式（具体程序实现过程请参见附件）。将“基层管理者”、“中层管理者”、“高层管理者”数值上分别标记为1、2、3，根据散点图，直观上选择拟合的函数为：

其中，为管理层级的标记数值，为权重系数。

根据最小二乘法，可以得到（数值对应四舍五入，下同）：

故而得到“基层管理者”与“中层管理者”合并的权重为：

故以7.851作为CSS问卷中选择对应题目选项2的被访者的得分权重；

3. 出于对受歧视职业的种类、程度等可能存在历时性变化的考虑，暂不将其纳入本文对职业声望的测算中；

4. 将职业声望得分作为社会地位得分。

综上，本文对个人社会地位测算的方程如下：

公式1

社会地位Y=11.808+3.349\*被访者折算教育年限+0.573\*被访者折算平均月收入(百元)+16.075\*管理活动情况为“只管理别人，不受别人管理”+7.851\*管理活动情况为“既管理别人，又受别人管理”+8.942\*是否在党政机关工作+6.841\*是否在事业单位工作-5.694\*是否在企业单位工作

（二）度量学习与聚类分析的应用

对社会声望分层研究做出度量学习和聚类分析角度的考察，是本文的主要目的所在。从实证分析的角度看，社会分层本质上是一个对社会中的人或群体进行分类的过程，是社会各形式差异和不平等差异的结构性描述。确定社会分层的分界线和区隔是社会分层研究的主题之一。故而，度量学习和聚类分析能够适用于社会分层领域，一方面，度量学习作为有监督学习能够与现有理论和分层描述相适应，采用现有分层结构的标的；另一方面，聚类分析作为无监督学习的代表，能够在社会地位的经验性指标的基础上，做出统计意义上的结构性划分。将度量学习和聚类分析应用到计量社会分层研究中，是有其优越性的。所以，本文选取KNN算法和层次聚类分析算法，试图在算法层面上考察社会分层的结构性特征。

三、统计结果及分析

1. 数据描述

出于算法需要，本文只选择CSS调查中针对所有对应问题有所回应的数据。在刨除了存在拒答、不适用、不清楚具体答案等情况的数据后，一共获得了4475个有效样本。样本对有关社会地位问题回答的整体情况可见下表。

|  |  |
| --- | --- |
| 样本各社会地位指标情况 | |
| 受教育年限（折算） | |
| 0 | 514 |
| 6 | 1181 |
| 12 | 1399 |
| 15 | 497 |
| 16 | 176 |
| 19 | 340 |
| 20 | 333 |
| 22 | 34 |
| 从事工作情况 | |
| 只从事非农业工作 | 1730 |
| 以从事非农业工作为主 | 248 |
| 只从事农业工作 | 2210 |
| 以从事农业工作为主 | 286 |
| 管理情况 | |
| 只管理别人 | 102 |
| 既管理别人，也受别人管理 | 727 |
| 只受别人管理 | 1149 |
| 从事非农工作所在公司/单位 | |
| 在党政机关工作 | 99 |
| 在事业单位工作 | 219 |
| 在企业单位工作 | 382 |
| 其它 | 64 |

可见，在各个社会地位指标中，样本以主要或完全从事农业经营为主，也即本文定义的农业人口；教育经历集中在6-12年，以中小学学历为主；整体月收入（农业人口按年收入按月平均计算）平均值为2362.62元，标准差为5937.97；非农业人口样本（即主要或者完全从事非农业劳动者）的管理情况以“只受别人管理”为主，占总数的一半以上，而“只管理别人”的被访者只占有效样本的5%；在非农工作所在单位/公司情况中，个体户、民办事业单位或社区居委会、村委会等自治组织占比最少，在事业单位、企业单位工作的被访者数量大致相当，而在党政机关工作的被访者仅仅略多于在个体户、民办事业单位或社区居委会、村委会等自治组织工作的被访者。总体来看，有效样本的各个社会地位指标基本符合常识预期，CSS问卷数据的效度有一定保证。

根据公式1，在样本社会地位指标情况对样本社会地位指数进行了赋分，并对样本的社会地位得分进行标准化转换，并重新映射为大于等于1的正值。最后，对各样本的得分取整。最终取得的社会地位得分经过了两次标准化转换。第一次标准化转换的公式是：

其中，X是某个样本的初始社会地位指数得分，S为样本社会地位指数标准差，δ为样本社会地位指数平均得分，n为样本数。转换并去除1个极值（得分为3279分）后，4474个样本的得分在1至721分之间，平均分为23.55，标准差为45.62，最小值为-23.14。故而，第二次标准化转化的公式为：

从而得到各样本的正值得分，最后进行四舍五入的转换，最终得到标准化的数据，并绘出柱状图[[2]](#footnote-3)。图形显示，样本得分集中在6分左右，低分（得分低于平均分）样本占比远超过得分较高样本，整体数据呈现程度不低的右偏态分布。也就是说，尽管在2017年，社会地位较低的被访者仍然占据CSS调查访问的相对多数，并且相对平均分而言，存在大量地位得分极低者（1分至6分）。这一点符合李强提出的中国社会结构的“倒丁字型”结构模型的特征[[3]](#footnote-4)。从本文选取的样本来看，中国社会结构差异在近年仍然存在，结构紧张尚未充分缓解。在社会地位指标得分的基础上，本文仿照李春玲文章的做法，对得分不同的样本进行了层次划分处理。具体分样本得分为0至5分、6至35分、36至65分、66至95分、96至121分、122至160分、161分以上共7层（下称样本社会地位层次），层次递增代表社会地位等级提高[[4]](#footnote-5)。同样，分层后的数据仍然呈现右偏态分布。

2. 度量学习：基于KNN算法的分析结果

KNN（K-NearestNeighbor）算法，即邻近算法，是一种常用的监督学习方法，也是度量学习的基本算法。其基本思路是通过对测试样本的某种距离度量找出训练集中预期最靠近的k个训练样本，基于这k个“邻居”的信息进行预测。但KNN算法是“懒惰”（Lasy）的，只有在给出测试数据之后才会进行算法学习过程，除此之外只是在储存数据特征。故而，k近邻算法在社会分层研究中的应用意义更多地在于说明“社会层次”的结构性存在[[5]](#footnote-6)。k近邻学习的理想情况要求训练样本的“密采样”（Dense Sample），这一点往往通过数据降维实现。本文运用k近邻算法，以上述样本社会地位指标为数据特征，社会地位层次为数据标签，划分测试集和训练集进行学习。但以样本社会地位层次为数据标签的处理有其特殊性，即社会地位层次标签实际上是由公式1得来，样本特征与数据标签之间的关系是显式的。

调整样本特征量纲与公式1符合后进行k近邻分类测试结果如下表所示[[6]](#footnote-7)：

|  |  |
| --- | --- |
| 保证显式关系的k近邻学习 | |
| 近邻数 | 模型得分 |
| 1 | 0.64 |
| 2 | 0.57 |
| 3 | 0.61 |
| 4 | 0.63 |
| 5 | 0.63 |
| 6 | 0.62 |
| …… | …… |
| 496 | 0.29 |
| 497 | 0.28 |
| 498 | 0.28 |
| 499 | 0.28 |
| 500 | 0.28 |
| 最优近邻数 | 1 |

可以看出，在不宜进行数据归一化和主成分分析的情况下，因数据特征不满足密分布，k近邻算法的预测效果不理想。但一定程度上仍然可以说明，社会地位的结构性差异是客观存在的。样本社会地位指标在概率上（0.64）可以由其近邻样本预测。

3. 聚类学习：基于密度聚类的分析结果

基于KNN算法在样本密分布和特征完备性上的困难，本文采用了无监督学习的方式，运用密度聚类方法（Density-Based Clustering）建立不依赖公式1提供的社会地位层级的模型。密度聚类方法依靠样本之间的分布的紧密程度（空间上的密集）来确定聚类结构。这也是本文选取聚类学习方法作为主要分析手段的原因：一方面，这一思路十分符合通常情况下人们对社会分层结构的“人以类聚、物以群分”的直观理解；另一方面，聚类分析属于无监督学习方法，密度聚类学习不需要样本标签，从而使密度聚类可以与传统分层研究结果相互对照[[7]](#footnote-8)。

本文选取DBSCAN（Density-Based Spatial Clustering of Applications with Noise）算法进行密度聚类，其基本思路是构建若干个在其欧式距离邻域ε内至少包含*MinPts*个样本的核心对象（Core Object）。所有在各自ε邻域内有交集的核心对象组成一个聚类簇。不属于任何簇的样本被判定为噪声。进行DBSCAN密度聚类分析的关键就在于确定模型的两个参数ε和*MinPts*的取值。

本文通过经验判断此二个参数的最佳取值，在本数据集中，ε可以理解为用欧式距离衡量的两个样本之间的社会差异性，而*MinPts*则是数据集中社会差异性较小的样本集至少应当包含的数目。对于ε，ε的迅速扩大可能意味着社会差异的陡增，这往往是跨越了阶层的结果。故可以以数据集中的第一条数据与其它所有数据的欧式距离为例，观察距离的突增情况。对于*MinPts*，则根据“社会地位层次情况柱状图”得出层次人数最小值，并以此值为基础适当扩大遍历区间（见下表），最终得到最优模型参数ε=157，*MinPts*=40，此时轮廓系数为0.5799[[8]](#footnote-9)，聚类分析拟合效果较好。

进一步地，将DBSCAN聚类的结果与公式1提供的社会地位标签比较。DBSCAN聚类结果和社会地位标签的同异，有DBSCAN标签相同且社会地位标签相同、DBSCAN标签不同且社会地位标签不同、DBSCAN标签相同而社会地位标签不同、DBSCAN标签不同且社会地位标签相同四种情形，记四种情形的样本数为TP、FN、TN和FP[[9]](#footnote-10)，各自对应的样本数如下：

|  |  |  |  |
| --- | --- | --- | --- |
|  | Ture | FALSE | Ratio |
| Positive | 407490 | 725988 | 0.05 |
| Negative | 6053468 | 12829730 | 0.94 |
| Ratio | 0.32 | 0.67 | 1.00 |

至此，可以认为，DBSCAN算法给出的样本聚类模型和已有模型有一定相似性，不同的是，DBSCAN算法给出了更细致的簇划分。

四、结语

本文只是对统计学习方法在社会分层结构研究中的可能应用进行了初步尝试。研究过程中，度量学习和聚类分析，尤其是基于DBSCAN方法的无监督聚类学习，展现了较高的探究社会分层结构的能力和潜力。但对于多维度展现的社会属性，统计学习算法的不可视化和黑箱性质表现得十分明显，相较于传统的基于基本理论给出的社会分层评价指标，统计学习算法的结果几乎不具有可解释性。需要指出的是，本文只是根据社会分层的基本过程（即划分群体）选择的算法，经过研究的探索，可以建议在算法上的进一步尝试神经网络算法进行对各参数可解释的分层研究。

进一步地，机器学习模型的建立乃至算法的评价，是无法脱离社会学理论而进行的。统计学习在社会分层研究乃至整个社会科学研究中的运用需要一个尚未确立的哲学方法论基础：是作为传统质性工具的补充，还是新的量化方法？或许对于直接进行基于计算机和互联网的解释学，机器学习用于经验研究的形式和可行性还需要进一步探索。

1. 李春玲、吕鹏，社会分层理论， [↑](#footnote-ref-1)
2. 数据柱状图参见附件“社会地位指标得分柱状图” [↑](#footnote-ref-3)
3. 李强.“丁字型”社会结构与“结构紧张”[J].社会学研究,2005(02):55-73+243-244. [↑](#footnote-ref-4)
4. 分层层次柱状体参见附件“社会地位层次情况柱状图” [↑](#footnote-ref-5)
5. 这一点涉及机器学习方法运用于哲学社会科学领域经验研究中的方法论问题，本文在结语部分有稍做简单探讨。 [↑](#footnote-ref-6)
6. 事实上，本文在研究中的确进行了归一化和主成分分析的尝试，其效果符合此处推断，相对于简单进行量纲修正的k近邻算法的正确率（64%），进行了Max-Min和Standardize归一化的训练测试集只能获得32%左右的正确率，而进行了PCA降维后的训练测试集则只能获得最高30%的正确率（n\_components=5），这一测试暗示了公式1的有效性。详见所附ipynb文件。 [↑](#footnote-ref-7)
7. 但本文中样本特征的选取仍然输与传统社会分层理论的研究范畴。在研究的一开始怎样描述样本、构建样本特征，需要凭研究者运用自身常识和理论进行判断，这也是所有机器学习方法的共性； [↑](#footnote-ref-8)
8. 详见ipynb文件。 [↑](#footnote-ref-9)
9. 基于算法特殊性，这里的TF、TN等只做记法标识而并无其原来的衡量算法指标的含义。 [↑](#footnote-ref-10)