基于武汉高校学生性格分析及匹配问题

徐木子，倪可欣，廖瑜蕾

（湖北经济学院，湖北 武汉 430205）

【摘 要】：性格的分析和匹配问题长期以来受到世界各国学者研究。与此相比，中国在这一领域还没有进一步的发展。基于以往的研究，介绍了分析不同性格以及匹配的方法。性格特征的原始信息由发放给武汉部分高校的400多份问卷收集而来。我们使用二维坐标轴来确定性格的标准，并将从问卷中收集到的信息数字化。采用K-means编程来对数据进行分组。然后通过欧氏度数，我们找到不同性格之间的距离范围以便衡量每种性格的人对于不同人格的偏好。

【关键词】：DISC性格分析；聚类算法

【中图分类号】：G4 **【文献标识码】**：A **【DOI】**：10.19311/j.cnki.1672-3198.2018.20.068

0、引 言

心理学分析研究由心理医生西格蒙德•弗洛伊德提出。他在人格结构的意识形态中提出了三个发展阶段一“自我”，“本我”，“超我”。Sigmund Freud根据自己的研究，将人格划分为8类。该理论被应用于医疗领域，用以治疗患有精神疾病的患者。之后，其他科学家和学者也对性格进行细分，这些理论构成了性格分析测试的基础，旨在找出测试者确切的性格类型。MBTI考试主要为企业提供就业指导，并且为测试人员提供了 5 点需要遵循的条例，如“回答问题的时候保持放松状态”。由于测试者需要遵循这些，MBTI测试结果很容易受主观因素影响。另一个名为Enneagram（九型人格）的测试模型与MBTI测试不同。该模型将人格分为“完美型”，“助人型”，“成就型”，“自我型”，“理智型”，“疑惑型”，“活跃性”，“领袖型”和“和平型”。它与其他模型相比突出的一点在于它深入到人们的思考模式而不仅是外部行为。此外，这个理论表明，当某些元素发生变化时，性格特征也会发生变化，从而建立了不同个性之间的关系。然而，经过多次测试，该方法的准确性受到限制。三个主要原因可以解释。一个是对于导师的依赖，一个是算法的缺陷，最后一个是测试问题的冗杂。与 Enneagram（九型人格）测试模型相比，DISC 测试减少了性格划分的维度，并通过对数据进行数字化来区分最突出的特征与最不突出的特征。更多的测试中的问题数量是相当适度的。根据研究，发展成为成熟步骤的大多数性格测试旨在帮助个人选择合适的工作和企业聘请合适的人员。虽然这些测试开始涵盖其他领域，比如交友市场。在当代，互联网成为了匹配朋友的新方式，打破了人们之间的距离界限，因此在市场上出现了大量的相关应用。尽管如此，这种应用极少考虑到用户的性格特征，不利于用户间长期的交往。这篇学术论文的目的是寻找性格分析和匹配的方式，为高校学生构造一个良好的交友环境。此外，该方法还可以用于学校的小组学习，以尽量减少群体成员之间的潜在矛盾。

1、研究程序

1.1 设定性格标准

考虑到Enneagram（九型人格）测试的不足，我们决定减少性格的维度。基于DISC的理论，我们以多维尺度回归的方法作为问卷设计基础，设立二维坐标，将问卷中设定的性格信息数据化，以供后期匹配分析。其中一个评估人的主动性（更接近外

向的个性或内向的个性）另一个衡量人的灵活性（更倾向于稳定的情况或变化的情况）。由于这两个维度的测试以前已经被测试过，我们主要对测试题目进行了改动。

我们基于二维模型将性格划分为 4 种。开朗支持型人格被定义为“OS/SO”型，活泼机灵性被定义为“OA/OA”型，沉静稳健型划为“IS/SI”型，冷静思辨性则为“IA/AI”型。

1.2 设计问卷

问卷中有 15 个问题。前 14 个是多项选择题，每个问题包括四个选项。七个问题旨在衡量主动性的维度，而左侧则用于测量灵活性的维度。每个选择都反映每个人物的程度。为了使测试结果变得更加精确，每个问题都包括一个特定的情况。在 Enneagram（九型人格）测试中，间隔表是评估性格的主要方式。然而，考虑到频率副词的划分容易产生误会，我们这里采用 TF—IDF 算法，把性格信息转换为权重向量。此外，考虑到测试人员是大学生，这些问题的场景设计均为学生的日常生活。最后一个问题用于确认测试者的偏好性格。我们使用数字表来设计问题，可以简化测量。在提交调查问卷之前，我们选取了自愿测试的三位学生，完成过后我们向周围的学生询问了相关问题，并将其与测试结果进行对比。在此基础上，纠正了一些具有争议的问题。

我们以多维尺度回归作为问卷设计基础，设立二维坐标，将问卷中设定的性格信息数据化，以供后期匹配分析。

1.3 信息权重转换

（1）TF-IDF 算法的具体原理如下。

第一步，计算词频，即 TF 权重（Term Frequency）。

词频（TF）=每个词语在性格选择中出现的频率。

由于每个人对理想型性格的诉求不同，我们采取“标准化”词频的处理方式，以便不同文本的比较，将文本中单个研究关键词除以文本中出现频率最高的词的出现频数或者文本的词数总体之和：

词频（TF）=问卷和调查中单个研究关键词出现的次数/总词数

第二步，计算 ID 权重，即逆文档频率（Inverse Document Frequency），需要建立一个语料库（corpus），用来容纳性格特征的选择。逆文档频率（IDF）越高，那么这种性格选择出现于问卷和调查中的分布就会越集中于一个点，说明这个描述词在内容属性能力方面的区分能力越强。

第三步，计算 TF-IDF 值（Term Frequency Document Frequency）。

TF-IDF=词频（TF）×逆文档频率（IDF）。

根据计算可以分析出 TF-IDF 值越高，则该表现性格特征的词语在问卷和调查中出现的次数就越多（成正比）。反过来说，某个词的出现频率和被选择的次数越多，则 TF-IDF 值就越大。逐个算出被选择和填的每个性格特征信息的 TF-IDF 值，并对这些值根据大小来排出顺序，最大的值就是要提取的性格描述中出现的次数最多的关键词。

（2）生成 TF-IDF 向量的具体步骤。

①用 TF-IDF 的计算公式，寻找被测试的人选择和描述的性格特征频率最高的五个描述词。

②这些被选出的五个描述词性格特征描述词，组成一个共同的集合，并算出在集合中每个性格特征描述词的词频。若是没有，则记该词词频为 0，这个词语被提及的次数越多，则这个词频的数值越大。

③根据公式算出各个性格特征词的 TF-IDF 权重向量：

TF-IDF=词频（TF）×逆文档频率（IDF）

（2）数据统计。

①计算各个性格特征描述词出现的总次数，并通过比较大小单独列出最热门的五个性格特征描述词。

②对各个年龄的理想型的性格进行分类计数，通过排序得出不同年龄层对理想型性格要求排名前 5 的性格特征。

③对同类职位的理想型的性格进行分类计数，通过排序得出不同类别职位的人对理想型性格要求排名前 5 的性格特征。

（4）聚类中心分类结果。

用 TF-IDF 算法最后选出五个性格描述词后，借助 K-Means 分类可以得到聚类中心，运用 KNN 算法，得出与聚类中心相匹配的五个其余类别，并通过对出现频数的估计，衡量聚类中心词的范畴：

①算距离：确定聚类中心，算出范围内的性格描述词与其自身的 TF-IDF 权重向量的距离。

②找邻居：挑出最靠近聚类中心的十五个性格描述词。

③做分类：依据分类的界限，对聚类中心进行分类。

1.4 收集和分析数据

我们在互联网上发送了大约三百份问卷，并在湖北经济学院发了一百份。其中有五十份问卷不能使用由于答卷者含糊其辞。剔除掉无用数据后，我们获取了一些有价值的信息。在分组和匹配过程中运用了四种方式。

（1）二维坐标轴。

二维坐标轴是 DISC 测试模型的理论，它通过二维尺度来评估性格。该方法的好处是可以直接通过数字反映性格。我们将每个选项代表的数字求和，根据具体数字进行评估。每个维度数据最大的反映了其性格。为了计算一般结果，将个人的数字加起来平均。同样，将个人偏好性格的数字相加后再平均。找出每个人的偏好个性。根据结果，我们发现，77.17%的人喜欢“10”字。此外，大约 30%的组是“AO/OA”性格。

（2）多元回归模型。

多元回归分析方法是研究多个自变量与一个因变量间是否具有某种线性或非线性的关系的统计学研究方法，旨在分析多个自变量与因变量之间可能具有的数量关系，以便于分析自变量对于因变量的影响，达到优良的预测效果。多元回归模型的一般模型为：

$$Yi=\beta_0+\beta_1 X1_i+\beta_2 X2_i+\cdots+\beta_k Xk_i+\mu_i i=1,2,\cdots,n$$

其中 k 为解释变量的数目，$\beta_j(j=1,2,\cdots,k)$ 称为回归系数。通过对自变量与因变量的研究，发现二者具有线性关系，因此此次回归模型归结为多元性回归。设 y 为应变量，X_1，X_1…X_k 为自变量，多元回归模型为：

$$Y=b_0+b_1x_1+\cdots+b_kx_k+e$$

其中 b_0 为常数项，b_1，b_2 …b_k 为回归系数。

本文基于文献回顾以及深入调查，选取了九个对于因变量有一定影响的因素，依次为活跃程度，规律性，应激反应，抗压能力，计划性，爱好广泛程度和专注度。为测试这些变量与学生性格的关系，我们对于这些数据进行了处理。由于用变量的对数形式能更好的估计自变量与因变量的百分比变化，因此建立了对数模型加以分析：

$$\text{Character}=b_0+b_1x_1+\cdots+b_kx_k+e$$

将对于这九项指标的测试数据进行预处理后经过计算，活跃程度和规律性的数值接近显著值，因此选为主要变量。

（3）K-Means 算法。

K-means（Mac Queen，1967）是解决众所周知的聚类问题的最简单的无监督学习算法之一。该过程遵循一种简单的方式，通过先验固定的一定数量的簇（假设有 k 个簇）来分类给定的数据集。主要思想是定义 k 个质心，每个集群一个。由于不同的位置会导致不同的结果，这些质心应该以灵活的方式放置。所以，更好的选择是让它们尽可能的远离彼此。接下来的步骤是把属于给定数据集的每个点和它关联到最近的质心。当没有点需要处理时，第一步就完成了，并且早期进行了组合。在这一点上，我们需要重新计算 k 个新质心作为上一步产生的聚类的重心。在制定出这些新的重心之后，必须在相同的数据集点和最近的新质心之间进行新的测量。已经生成了一个循环。作为这个循环的结果，我们可能会注意到，k 个质心一步一步地改变了它们的位置，直到没有进行更多的改变。换句话说，重心不再移动了。最后，该算法旨在将目标函数最小化，在这种情况下是平方误差函数。目标函数：

$$J=\sum_{j-1}^{k}\sum_{i-1}^{n}||x_i^{(j)}-c_j||^2$$

$||x_i^{(j)}-c_j||^2$ 用来计算数据点 $x_i^{(j)}$ 与聚群中心的距离。

该算法由以下步骤组成：

①将 K 点放入由聚类对象表示的空间中。这些点代表初始组质心。

②将每个对象分配给具有最接近质心的组。

③分配所有对象后，重新计算 K 个质心的位置。

④重复步骤 2 和 3，直到质心不再移动。这产生了将对象分离成可以计算要最小化的度量的组。

基于 K-means 的理论。数据被分为 4 个类别，因为 K 被设置为 4.然后选择作为组中项目的圆的质心。重复这一步骤多次，重心固定下来了，代表了整个组群的水平。

（4）匹配方法。

欧几里德度量是一种计算 m 维空间中两点的实际距离的方法。在二维坐标轴上，可以用来测量点之间的距离。

$$\rho(A,B)=\sqrt{[\Sigma(a[i]-b[i])\hat{}2]}(i=1,2,\cdots,n)$$

基于欧几里德度量，我们测量测试人员的性格与偏好性格之间的距离。经测量，距离范围在 0.5 和 2.7 左右的（最大值为 10），可以通过考虑图来适应人物的匹配。定义我们通过特定数字知道的性格并将其设置为圆的中心，0.5 和 2.7 分别代表最小和最大半径。重叠部分是可能涵盖偏好性格的区域。

2、结 论

本文介绍了分析和匹配个性的创新方式。根据前期计划的理论，计算程序分为三个部分。我们通过“二维坐标轴”将问卷信息数字化，然后通过多维回归模型测试出显著性影响因素，运用 K-means 程序将数据分组为四组。最后，通过欧几里德度量匹配组。根据测试结果，我们发现一些问题不能清楚地划分性格，所以测试问题后期还需调整。总之，该方法可以用于匹配不同个性的学生。对于大学教育工作者，可以将学生分成不同的类别，以减少学生之间的不必要冲突，增加团队的凝聚力。

[**参考文献**]：

[1]罗石涌.敏感者的生命色彩——卢西安·弗洛伊德及其艺术的主观解读[J].美术大观，2010，（3）.

[2]曹玉峰.论九型人格在企业人才招聘中的应用[J].人力资源管理，2012，（11）.

[3]李学明.基于信息增益与信息熵的 TFIDF 算法[J].计算机工程，2012，（8）.

[4]周爱武.一种改进的 K-MEANS 聚类算法[J].微型机与应用，2011，（21）.’

[5]李涛.多元线性回归与 LOGISTIC 回归分析的正确应用[J].临床荟萃，2009，（15）.

[6]施培蓓.初始化独立的谱聚类算法[J].计算机工程与应用，2010，（25）.