Eliciting Thinking Hierarchy without a Prior

Abstract

当我们在参考群体的智慧时，我们通常会根据解决方案的支持人数来对答案（结果）进行排名，尤其是在我们无法验证答案本身是否正确的时候。——**注：**可以简单理解成是“少数服从多数原则”。

然而，当大多数人犯有系统性错误的时候，这样的答案排名方式有可能是非常糟糕的情况。——**注：**即并不是多数人所认为的答案就一定是正确的，也很有可能正确答案只掌握在少数或极少数人的手中。

因此，本文所讨论的一个基本的问题就是：我们能否在没有任何先验的情况下，在问题的所有答案中建立一个等级制度，找到那些可能不受大多数人支持的、但更正确的答案？而这些答案可能来自于思维更为老练的人。

而为了解决上述的基本问题，本文提出了 1）一个**新的模型**来描述人们的思维层次；2）两种**不需要任何先验的算法**来学习人们的思维层次；3）基于前两点的理论框架，一种新的基于开放式回答的**实用型众包方法**。并且，除了理论论证之外，还进行了四项实证性质的群体研究，并表明：a）通过上述方法获得的排名靠前的答案的准确性远远高于多数投票(例如，在一个问题中，多数答案得到 74 个回答者的支持，但正确答案只有 3 个回答者的支持。而本文所提出的方法在没有任何先验条件的情况下，却将正确答案排在了最高位)；b）本文所提出的模型具有很高的拟合优度，特别是对于那些排名靠前的答案是正确的问题。

相关说明：本文是第一个在一般问题解决场景中提出已经有经验验证的思维层次模型的研究；也是第一个提出一种实用的基于开放式回答的众包方法的研究，并且该方法击败了同样没有任何先验条件的“多数投票原则”。

Introduction

群体的智慧已经被证明比个人的智慧能带来更好的决策和解决问题的能力，特别是当我们（大众）没有足够的先验知识来判断个别专家的意见时。

此外，多元性回答也是最受人们欢迎的集合大众意见的方式之一，且这些不同的观点也通常根据受大众的欢迎程度来进行排名。然而，当大多数人都带有系统性偏见的时候，这种方式却可能是非常糟糕的。例如下面，这是本文进行的一个现实研究，这些回答来自于多位大学生。具体问题如下：

如下左图所示，已知圆 A 的半径是圆 B 半径的 1/3，那么当圆 A 绕着圆 B 转一圈再回到起点时，此时圆 A 总共旋转了多少次？

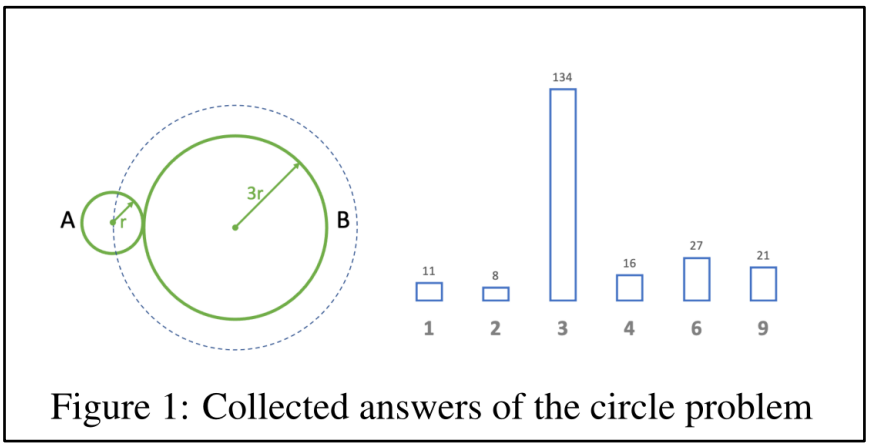
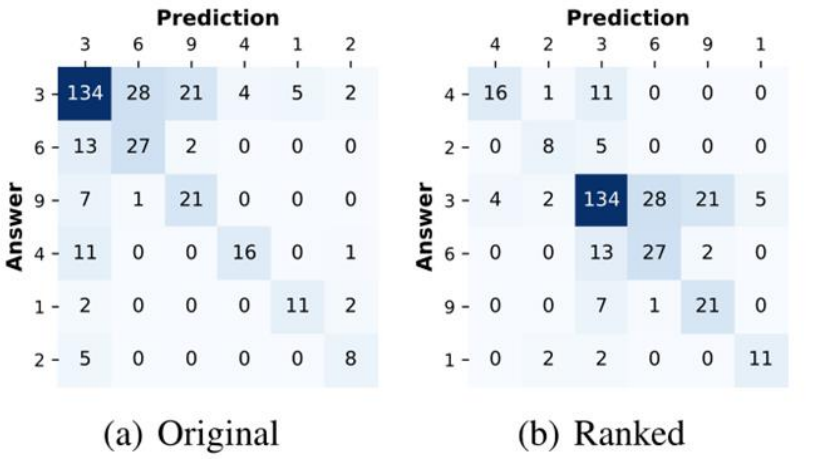


Figure 1: Collected answers of the circle problem

如上右图所示，在该项研究收集到的答案中，其中有 11 人认为“1”是对的，8 人认为“2”是对的，134 人认为“3”是对的，16 人认为“4”是对的，27 人认为“6”是对的，21 人认为“9”是对的。虽然多数人的答案是“3”，即大圆的周长与小圆的周长之比。但正确答案却是“4”，而它只有 16 个人支持。**——注：这说明用“少数服从多数原则”来判断正确决策的方法并不总是成立的。**

但如果我们有了一定的先验知识，比如知道每个回答者的专业水平（**说明：**例如上面这是一道数学问题，假设我们知道某位回答者他是数学系或是文学系的），那么，我们就更有可能确定正确的答案是什么，因为相关专业水平更高的人他的回答可信度也更高。或者换句话说，他的思维层次更高，能瞬间明白这个问题的本质。反之，则不然。但是，有时候获得先验知识的条件是相当昂贵且困难的，尤其是在新领域内。



(a) Original (b) Ranked

Figure 2: The empirical results of the circle problem.

此外，在上述问题的另一项调查中（回答者除了需要回答自己觉得正确的答案外，还要去预测其他人最可能会回答什么样的答案【可以跟自己的答案一样】），如上左图所示，在回答“3”是对的 134 人中，仅有 4 人认为其他人会回答“4”，而在回答“4”是对的 16 人中，却有 11 人认为其他人会回答“3”。而本文注意到，这在没有任何先验知识的情况下，是一个非常有趣且关键的现象。——**注：**上述实验现象其实可以说明两点：1）选“4”的回答者觉得这道题不简单，大部分人不会像自己一样选对，因为如果问题很简单，比如是“1+1=？”，那么得出“2”的人也会认为别人也能得出“2”的结果；2）选“4”的回答者在一定程度上了解别人的想法，所以知道大部分人可能会选择“3”而不是其他数，但选“3”的回答者却并不一定了解为什么有人会选择“4”。

因此，针对本题以上的结果，可以得出以下一些结论：1）本题目并不简单，但大部分人不知道；2）结果“4”才是正确的答案；3）大多数人的思路可能存在共同的错误(系统性偏见)，才导致他们选“3”；4）选“4”的人能够看破选“3”的人的思路，但选“3”的人却完全不懂选“4”的人。——**注：**实际上，这些结论可以类比高考，学霸做题，也许能大概分析出出题人想考什么，设置了哪些问题陷阱，而学渣做题，觉得自己都会做，但考出来的分数可能并不高（只是单纯的一个比喻，方便形象理解）。

因此，本文总结的关键观点是：**1）高（专业）水平的人往往有正确的答案；2）高（专业）水平的人往往可以预测低（专业）水平的人的答案（思想），但反之则不然。**

关键问题：

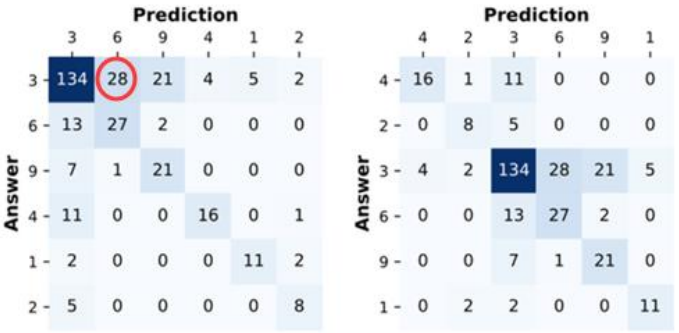
利用上述观点，可以构建一种有趣的思维层次结构（参考**思维层次理论 CHT**），即答案(策略)之间存在着一种等级关系，一定程度上也反映了回答者的思维水平。而本文的目标是：在没有任何先验知识的情况下，建立一个实用的方法(算法)来学习思维层次模型；且基于思维层次模型，可以有效地对答案进行排序，使得排名较高的答案可能来自于更老练(思维水平更高)的人，而这些答案可能不被大多数人支持。——**注：**简单来说，本文就是想在完全不知道一群回答者专业水平如何的情况下，通过群体统计的方式来计算获得正确的答案，而不是简单地依靠“少数服从多数原则”。

而本文之所以想要的是答案等级排序而不是一个最好的答案其实有很多原因，其中关键的两点是：1）对于一些主观问题（例如，为什么酒吧椅子高?），可能存在不止一个高质量的回答，而完整的思维层次结构提供了更丰富的结果。2）答案之间的等级关系有助于更好地理解人们是如何思考的，尤其是当试图引出人们对一项政策的看法时，这一点显得非常重要。

解决办法：

参考其他相关研究，本文遵循同时寻求答案和预测的框架，并将其扩展为一种更实用的基于开放回答的范例。该范例要求一个唯一的、但开放的回答结果，并要求每个回答者都有自己的答案和对其他人答案的预测。例如，在上述的“圆”问题中，回答者需要提供自己的答案：“4”，以及对他人的回答预测：“3”。

接着，依据上述的回答结果，本文构建了一个“答案—预测”矩阵 A，用来记录特定的“答案—预测”对的人数（例如，如下图(a)显示 28 人自己的回答是“3”，并预测其他人的回答是“6”）。而需要注意的是，矩阵 A 的对角线是仅自己回答结果的人数之和，例如 134 人，即不考虑对他人的预测。



(a) Original (b) Ranked

Figure 2: The empirical results of the circle problem.

为了更好地了解思维层次，本文提出了一个新的思维层次模型，用于描述不同思维水平的人是如何回答问题的。在该模型中，回答者自己的答案和预测的他人答案的联合分布取决于描述人们思维层次的潜在参数。并且本文表明，当给定回答者自己的答案和预测的他人答案的联合分布时，就可以通过解决非负矩阵分解问题中的一个新变体来推断潜在的思维层次，简称非负同余三角化(NCT)。接着，在 NCT 分析的基础上，本文提供了两种简单的答案排序算法，并表明：在适当的假设下，算法能够有效学习潜在的思维层次。

最后，本文假设由范例收集的“答案—预测”矩阵是回答者自己的答案和预测的他人答案的联合分布的代理（替代）。然后，本文在“答案—预测”矩阵上实现了基于 NCT 的答案排序算法，而默认算法是通过**最大化矩阵上三角区域中元素的平方和**对答案进行排序的。此外，在另一个变体版本中，为了允许不同的答案具有相同的复杂程度，答案被分割为压缩矩阵，而变体算法则是通过最大化压缩矩阵的上三角形区域中元素的平方和来对答案进行排序的。

而除了上述理论框架外，本文还通过在数学、围棋、常识、汉字发音等各个领域向人们提问进行了实证性质的研究。

例：“圆”问题的实验结果。在上述的“圆”问题中，实证研究收集了朴素的“答案—预测”矩阵（如上图(a)），并根据本文所提出的算法对其重新进行了排序（如上图(b)）。可以发现，在本文所提出的算法计算下，答案“4”被认为是等级最高（最正确）的答案。

Learning Thinking Hierarchy

1）思维层次模型的定义；

2）非负同余三角化(NCT)的实现；

3）利用“答案—预测”联合分布推断思维层次的算法实现；

4）假设“答案—预测”联合分布的代理。

本文的理论框架主要分为如上四个部分，具体实现方法参照原文，这里不作说明。——论文链接：https://arxiv.org/abs/2109.10619

Studies

本文选取了数学题（35 个）、围棋死活题（30 个）、常识题（44 个）和汉字读音题（43 个）四种类型的问题进行实验。在四种类型的 152 个问题中，算法在 134 个问题中得到了正确答案，而“多数原则”仅答对了 116 个问题，其具体结果如下表所示。

Type	Total	Our algorithm(Default)	Our algorithm(Variant)	Plurality voting
Math	35	**29**	**29**	24
Go	30	**28**	**28**	23
General knowledge	44	**41**	**41**	35
Chinese character	43	**36**	35	34

Table 1: The number of questions our algorithms/baseline are correct.

而如上所示，可以发现：将本文所提出的算法与基线（多数投票）进行比较，不论是默认算法还是变体算法，它们在所有类型的研究中都击败了多数投票，且默认算法略好。此外，在 152 个问题中，变体算法在 138 个问题中输出了与默认算法相同的层次结构。而仅在一个问题中，变体算法的最佳结果是错误的，但默认算法是正确的。

除此之外，本文还计算了算法的缺失匹配指数，发现算法输出正确答案的问题的缺失匹配度更小，从而更好地拟合了思维层次模型。因此，可以将缺失拟合指标作为算法的可靠性指标。

各类型研究的案例分析：

参看原作者组的论文解读链接：http://cfcs.pku.edu.cn/news/241025.htm

Discussion

本文提出了一套新的信息汇总算法。该算法无需任何先验信息，只需要每个人在回答后额外提供对别人答案的预测。并且，本文还通过实验证明了在大多数情况下，即使大部分人都错了，该算法仍然可以汇总得到正确答案。除此之外，本文所提出的算法还可以对人们提供的答案进行“思维分层”，展现人们的“思考路径”。

此外，当对一项政策进行群体意见询问时，利用思维层次信息，可以更好地理解群体意见。然而，就负面影响而言，在充分了解人群的思维层次的情况下，也可能更容易实施媒体对社会舆论的操纵。因此，一个有趣的未来方向是探索思维层次信息的实际影响。

原作者组演示汇报链接：https://weibo.com/tv/show/1034:4778519457366120