目 录

从游戏 logfile 预测数学成就与推理能力：机器学习的应用

摘 要 计算机过程数据分析技术在测量领域得到了迅速发展，使用机器学习尽可能地挖掘过程数据是一种新的思路，本文在该思路下进行了尝试。本研究以一款流行的游戏推箱子为例，基于 360 名初一、初二学生的推箱子过程数据，应用机器学习技术预测学生的数学成绩分类以及瑞文推理测验得分分类。我们从推箱子的过程数据提取出一系列指标包括第一步用时占比、平均执行时间、执行过程波动、与最优路径重合比例、冗余步数等作为特征（预测变量）；分别取数学成绩、瑞文推理测验的年级前 25%或年级后 25%，构造得二分结果变量。我们使用随机森林模型，在 70%的样本中进行格点搜索以获得最优参数。训练的模型在余下 30%评估集上预测数学成绩分类最高能获得 83.07%的精确率、73.70%的准确率、73.33%的召回率以及 75.57%的 F1 得分；预测推理能力分类最高能获得 76.11%的精确率、65.72%的准确率、63.10%的召回率以及 65.01%的 F1 得分。模型具有较好的预测效果。

关键词 推箱子 机器学习 随机森林

Predict Math Achievement and Raven Reasoning Ability from Log Files of a Game: An Application of Machine Learning

Abstract Computer log file analysis has been developing rapidly. Using machine learning to further explore the log file data is a new thought. This paper contributes to this trend by using machine learning to classify a student's math grades and Raven reasoning ability from the log file of a popular game Sokoban. We extract a series of indexes from the log file as the features (predicting variables), including the proportion of planning time, the average execution time, the variance of execution time, the proportion of coincidence of one's path with the optimal path, the proportion of duplicated moves, etc. We take the highest 25% and the lowest 25% of math grades and reasoning ability respectively to construct two binary outcome variables. Random Forest Models are built by grid search strategy in a 70% randomly-picked subsample to find the optimal hyperparameters. When predicting math grades, the trained model can at most achieve an 83.07% precision, 73.70% accuracy, 73.33% recall and 75.57% F1 score. With respect to reasoning ability, the scores are 76.11%, 65.72%, 63.10% and 65.01% respectively. The performance of our model is satisfactory.

Key Words Sokoban Machine Learning Random Forest

1 引言

随着计算机技术的发展，计算机过程数据分析技术 (computer logfiles analysis)在测量领域迅速发展并且得到了日益加强的重视。计算机过程数据分析技术是指通过追踪、分析被试在计算机上完成认知任务过程中的操作行为、操作时间等信息提取测量指标 (Veenman, Bavelaar, De Wolf, & Van Haaren, 2014; Veenman, Wilhelm, & Beishuizen, 2004)。

用于测量领域，计算机过程分析技术相较于传统的技术有诸多显而易见的优点。首先，传统的测量方法或者基于被试的自我报告、或者基于主试的对于被试行为的编码，而过程数据的分析依赖于客观指标，能够更好地达到标准化 (Veenman & Alexander, 2011)；其次，传统的测量方法，尤其是在测量元认知能力时，会有较强的侵入性：如如观察、出声思考技术 (Pressley & Afflerbach, 1995) 等，在这些条件下被试完成任务的能力一定程度上会受到影响；而过程数据分析则具有无侵入性，被试完成任务的过程中不会受到打扰，能够在最自然地状态下表现 (Veenman et al., 2014)；最后，过程分析技术成本较低，可以对多个被试同时施测，数据分析通过计算机自动化完成，因此相较于传统的测量技术如出声思考法以及眼动追踪技术 (Kinnunen & Vauras, 1995)等更加的省时。

目前，计算机过程分析技术主要应用于元认知能力的测量中。一个典型的应用是 J. Li (2015)等人使用推箱子游戏的过程数据测量被试的计划性 (Planning)。推箱子是个经典的日本游戏，界面见下图。玩家需要控制小人将所有箱子推入到指定位置。小人只能向前推一个箱子，不能与箱子重合。被试被要求用尽可能少的步数完成游戏。计划性是指个体设置短期、长期目标并且产生、选择实现目标的策略的高级思维过程 (Schraw & Moshman, 1995)。他们从被试的过程数据中提取了被试在进行第一步之前的时间与总时间的比例作为计划性的指标，该比例越高反映出被试的计划性越强。他们的研究表明该指标具有很好的信效度，能够作为一个可靠的计划性测量工具。其他的过程分析技术在测量中的应用也见 Veenman et al. (2014)和 Veenman (2013)等研究。

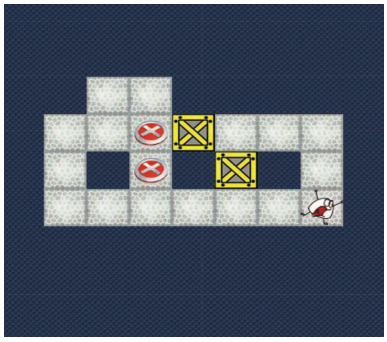

图 1 推箱子游戏的一个例子

然而，目前已有的计算机过程分析技术并非完善。首先，在已有的计算机过

程分析技术中，研究者通常从过程数据中提取出一个或多个指标，而浪费了潜在可挖掘的更多的信息；其次，在评价指标的有效、以及使用指标进行预测时，过去的研究往往依赖于传统的线性模型，如线性相关以及最小二乘回归技术。线性模型在表达能力上受到限制，往往不能获得最优地评估、预测；此外，多数计算机过程分析技术致力于提取出纯净的指标用以反映某个单一的认知过程，然而这些指标往往并不只是纯粹地包含一个元认知成分，而是多个元认知成分的混合，因此测量所得结果更大程度上是一个混合的“元认知技能”(metacognition skills) (Veenman, 2013)。

如果放松对指标纯净程度的要求，仅仅是广泛地测量“元认知技能”的话乃至综合认知能力，机器学习(Machine Learning)技术能够对以上问题提供一个可行的解决方案。机器学习能够使用多个指标，建立更为复杂的非线性模型，尽可能好地拟合现实。

这里涉及的机器学习技术主要属于监督学习 (Supervised Learning)的范畴。在一个常见的监督学习的情景下，我们有一个结果变量（比如，个体的元认知技能）以及一组特征（即预测变量，比如从推箱子的过程数据中提取得到的指标）。我们有一组数据作为训练集，在训练集中我们能够同时观察到结果变量与特征。使用这组数据我们能够训练出一个模型，或者称之为“学习器”。训练好的模型能够从只有特征而无法观测到结果变量的数据中预测结果变量。一个好的模型能够准确地预测出结果变量(Friedman, Hastie, & Tibshirani, 2001)。

机器学习在测量领域已经有不少应用。由于社交网络上样本量大，且信息丰富，使用社交网络数据预测心理指标是机器学习技术应在测量中应用的传统领域。Wu, Kosinski & Stillwell (2015)等人使用 LASSO 模型，从 86220 名被试在 Facebook 网站上的“喜欢”行为预测他们的大五人格特质。他们的结果显示，使用模型预测的结果与被试自我报告的人格特质的相关性（r=0.56）要高于由被试的好友对被试的人格特质评价(r=0.43)。此外，模型预测结果也有更高的外部效度以及内部评分一致性。Guan et al. (2015)等人使用随机森林模型，用微博上的用户信息预测被试是否属于高自杀风险人群，达到了 70%的召回率，但精确度仍低于 30%。类似的研究还包括 Kosinski (2013b)， Kosinski et al. (2013a)， Liu & Zhu (2016)等。

另一类使用机器学习的常见领域是使用移动传感器收集的生理指标进行预测。Zhang et al. (2016)使用了生理指标预测被试的情绪分类。研究者让 123 名被试佩戴内置加速传感器的智能可穿戴设备用于记录被试的移动过程。被试首先自由行走，该部分被传感器记录为中性。之后被试被分为两组，分别观看电影以启动被试的高兴与生气情绪，观看完电影后被试再次行走 1 分钟，行走时的方向、

加速度同样被记录下来。研究者们从记录所得的数据中提取出指标，使用机器学习预测该段行走是来自于中性、高兴还是生气组。预测准确率能达到 81.3%。这方面研究同样可见 Li et al. (2016)等。

以上研究或者使用可智能穿戴设备获得的生理指标，或者使用社交网络上的大数据。据我们所知，尚未有人使用游戏过程数据应用机器学习的方法建立测量指标。在本研究中，我们使用推箱子范式，尝试建立模型从推箱子的过程数据预测被试的数学成绩以及瑞文推理能力测验得分。使用这两个结果变量有以下考虑：推箱子游戏中的每个指标都有可能涉及到不止一种元认知、认知能力，无论使用某个单一的元认知能力或是认知能力得分作为结果变量均无法充分利用推箱子的过程数据。而数学成绩、瑞文推理能力得分作为结果导向测量工具，同样涉及了更为综合的认知、元认知能力，尤其是数学成绩，因此能够达到更好的拟合效果。

2 方法

本研究的工作分为 4 个部分：数据采集、数据预处理、特征提取、模型训练。

2.1 数据采集

2.1.1 被试

本研究总共涉及了 395 名被试，最终有效观测值为 360 个。被试为首都师范大学附属中学第一分校的初一、初二学生，其中女生 184 名，男生 211 名。初一平均年龄为 13.2 岁，最小为 11.9 岁，最大为 15.1 岁；初二学生平均年龄为 14.6 岁，最小为 14.1 岁，最大为 16.3 岁。

2.1.2 结果变量

本文涉及的被预测变量主要包括学生的数学学业成绩以及瑞文标准推理测验的得分。

瑞文标准推理测验(Standard Progressive Matrices, SPM, Raven, 1989)：本测验用于测量一般智力中的推理能力，共计 60 个条目，每答对一题计一分，满分 60 分，总分即为推理能力得分。在本研究所涉及的被试中，最低分 2 分，最高分 60 分，平均 46.2 分，标准差为 8.5 分。

我们将构造了瑞文得分的二分变量作为最终的预测变量：我们取瑞文推理测验前 25%得分的学生记为 1，后 25%得分的学生记为 0，其他学生不予使用。最终使用的被试数是 180 名。

我们同时获得了他们 3 次数学测验的成绩。每次数学测验均为年级内统一施

测，因此在年级内具有可比性。我们将三次数学成绩取平均分以获得对学生数学能力较为准确的估计。平均数学成绩的平均分为 64.9 分，标准差为 19.9 分，最低 4.7 分，最高 95 分。我们按照年级分层，取每个年级数学成绩前 25%的学生标记为 1，后 25%的学生标记为 0，其余学生不予使用，最终使用被试为 180 名。

2.1.3 推箱子

我们分两批收集了被试完成推箱子任务的过程数据。推箱子游戏基于 JAVA 平台实现。学生被要求使用尽可能少的步数完成任务，指导语一直显示在屏幕上。两个年级使用的是同一套题目，包括 20 关，由 Li et al. (2015)筛选获得，平均难度为中等水平，同时也具有一定的难度差异。所有受测者完成关卡的顺序均相同。我们施测使用的程序自动记录了他们每一步的方向、以及所用的时间。

2.2 特征提取

从推箱子游戏中收集到的原始数据并不规范，不能直接作为特征加入模型中，我们需要数据中提取出能够刻画被试能力的特征。

程序记录的一个典型的玩家顺利通过某一关的数据如下图所示。通常，在第一步之前会有较长时间的思考，在思考完成后则有较快的按键反应，操纵小人通关。除了过程数据外，程序同样记录了玩家在本关是否通过，完成了几个箱子，是否主动放弃本关或超时等信息。

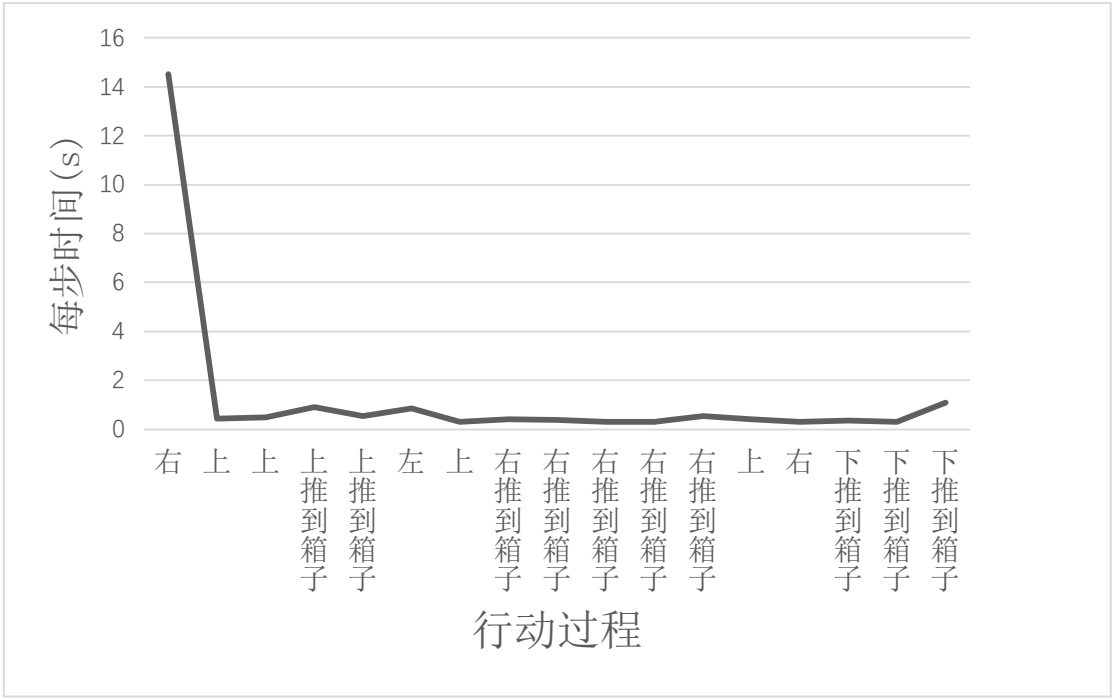

图 2 一个典型的行动过程

我们从过程数据中提取了以下几个特征用于模型预测：

2.2.1 第一步所用时间

由上图可见，被试通常会分配较长时间在第一步之前的思考中。第一步所用时间反映了被试进行计划的能力，即计划性(J. Li et al., 2015)。参照文献，我们计算出在被试在每一关中在第一步之前所用时间占总时间比重，作为刻画被试计划能力的指标。

除此外，我们同样用第一步时间除以总步数、第一步时间除以平均执行时间（定义方法见下）以及对上述变量取其对数形式，构造出多个变量以刻画其计划性。

2.2.2 执行间思考

不少被试在执行过程中会停下来思考，反映在数据中即是其在执行过程中某一步用时较其他时间异常变高。为了刻画这种波动性，我们计算了被试除了第一步之后各步用时的标准差。

我们记录了时间超过平均值一个标准差以上的步数占总步数的比例作为被试在执行过程中的思考次数的指标。

2.2.3 执行速度

我们剔除掉包含思考的行动后（包括第一步），将余下的行动时间求平均。这部分时间刻画了被试在无需思考的情况下对计划的执行的单步时间。

2.2.4 冗余步数

根据被试的行动路线我们可以求出被试每一步行动后的整个地图的状态。我们计算出其路径中重复的状态（即从箱子到小人完全的相同，不包括将箱子推动后小人撤回）占最优路径中总状态数的比例。其中，对于一个状态重复多次，我们只记为 1，以避免玩家在两个状态间来回“踱步”产生误差。

2.2.5 与最优路径重合比例

通过广度优先搜索算法，我们可以求出每一关的最优路径。我们计算了被试的路径与最优路径重合的比例。具体而言，我们计算被试路径的状态的集合与最优路径状态集合的交集，计算交集占最优路径状态集的比例，同样排除掉“来回踱步”的情况。

2.2.6 与最优路径相差步数

我们计算了玩家步数与最优步数（玩家步数-最优步数）的差异，将其也作为

一个指标加入模型中。

2.2.7 每题是否成功、放弃

每道题的结果有 3 种状态：成功通过、被试放弃以及超时。我们用 2 个二分变量刻画这 3 种情况。此外，对于没有通过的关卡中，程序也报告了被试完成的箱子占总箱子的比例，同样被我们作为一个特征加入到模型中。

2.2.8 数据预处理

原始数据观测值以每一关为一个观测值，但在训练模型时我们需要以被试为单位观测值。直接将长型数据转换为宽型数据（即对于每个被试每一关都提取出上述特征，总共特征数是单关特征数的 20 倍）存在以下问题：一，由于样本规模不大，特征过多不利于模型训练，容易过拟合；二、同一个指标在成功失败两种状态下可能有不同的意义。以与最优路径相差步数为例，失败的状态下，玩家的步数通常会低于最优路径步数，玩家即放弃游戏，数字越大表明玩家越坚持；而在成功的状态下，数字越大则反应玩家的步数偏离最优路径越远。

为了克服以上问题，我们将以上特征均以成功与否为分组在每个个体内求平均值，即对于以上每个特征，我们都构造出了 2 类：一类用于描述该特征在成功下对数学成绩以及瑞文得分的预测能力，另一类刻画了在失败状态下对数学成绩以及瑞文推理得分的预测能力。

对于是否成功、是否放弃两个二分变量，直接在各组间求平均容易忽略掉每道题的难度信息，因此我们使用因子分析从其中提取出两个因子作为特征用以训练模型。

以上主要涉及特征的描述统计见下表。

表 1 提取出特征的描述统计

变量	平均值	标准差	最小值	最大值
失败组				
(第一步用时/平均执行时间)	22.71	24.26	2.52	198.34
ln(第一步用时/平均执行时间)	2.31	0.82	0.81	4.97
完成箱子比率	0.33	0.08	0.00	0.57
第一步用时/总时间	0.22	0.12	0.04	0.76
ln(第一步用时/总时间)	-1.92	0.60	-3.31	-0.29
思考步数占比	-2.39	0.23	-3.04	-1.69
平均执行时间	0.64	0.15	0.37	1.33
执行间波动	2.15	1.20	0.35	10.52
重复步数占比	0.07	0.03	0.00	0.20
与最优步数相差	-5.75	9.45	-23.36	65.78
与最优路径重合步数占比	0.17	0.04	0.04	0.32
成功组				
(第一步用时/平均执行时间)	24.36	23.81	2.65	168.97
ln(第一步用时/平均执行时间)	2.49	0.78	0.92	4.95
第一步用时/总时间	0.25	0.14	0.04	0.77
ln(第一步用时/总时间)	-1.77	0.61	-3.18	-0.27
思考步数占比	-2.61	0.27	-3.53	-1.64
平均执行时间	0.48	0.11	0.33	1.18
执行间波动	1.17	0.76	0.20	5.43
重复步数占比	0.03	0.02	0.00	0.16
与最优步数相差	7.65	5.45	0.00	52.67
与最优路径重合步数占比	0.71	0.14	0.17	1.06

2.3 模型训练

瑞文推理与数学成绩的模型训练策略一致。我们首先随机划出 30%的样本作为评估集，70%的样本用于交叉验证以及超参数搜索。我们使用基于 Python 3 的 scikit-learning 包(Pedregosa et al., 2011)提供的随机森林模型进行训练。

我们首先随机划出 30%的样本作为评估集，剩下 70%的样本用于交叉验证以及超参数搜索。在 70%的样本中，我们使用了 n-fold 策略：我们将样本再次随机分成 4 组，依次选择其中一组作为验证集，其他三组作为训练集。我们在训练集上训练特定参数的模型，将获得的模型在验证集中测试，计算模型得分（评

分方式见下节）。4 个轮次后，每组样本均有 3 次作为训练集，1 次作为验证集。我们将 4 组中获得的得分求平均，为对应该超参数的模型在交叉验证组中的得分。

随机森林模型具有多个参数可供调整，不同的参数设置会极大地影响模型的拟合效果。我们关注的主要参数为：最大特征数、最大深度、最小分裂样本量、拟合器数量。最大特征数表示在寻找一个最优的分裂过程中需要考虑的特征数；最大深度表示一个决策树最大的深度，达到该深度后即停止分裂；当一个节点上的样本量低于最小分裂样本量时该节点即停止分裂；拟合器数量是指在随机森林中的树的数量。

我们采用网格搜索的策略寻找出最优的参数。我们提供了各个超参数的搜索范围（即网格），对每一种超参数组合都进行一次交叉验证，选择出交叉验证中平均得分最高的超参数组合。我们搜索的范围为：最大特征数由 5 至 16；最小分裂样本量由 2 至 10；最大深度由 2 至 8，拟合器数量包括 5、10、50、160 四种。因此总共需要进行 3024 次交叉验证。

经过 3024 次交叉验证后，在验证集上得分最高的超参数组合即为最优参数组合。但是通过交叉验证获得的得分并不能准确评估该模型的有效性：因为超参数搜索本身也相当于拟合的过程，容易出现过拟合。因此对于该模型的评估需要在评估集上测试。

首先，我们使用获得的最优参数，用交叉验证中涉及的所有 70%的样本对随机森林模型进行拟合。之后我们用拟合后的模型以及评估集中的特征对数学成绩分类、瑞文分数分类进行预测，用预测得到的分类与真实分类作比较，计算得分。由此得到的得分是对模型预测能力比较准确的评估。

2.3.1 模型评估

对于一个分类任务的表现有多种评估方式。首先，对于预测结果可以分为 4 类：真实阳性(True Positive)，即预测为阳性且实际上也为阳性；虚假阳性(Fake Positive), 即预测为阳性但实际上为阴性；真实阴性(True Negative), 即预测为阴性实际上也为阴性；虚假阴性(Fake Negative)，即预测为阴性但实际上是阳性。

表 2 分类表现评估表

	预测为阳性	预测为阴性
实际为阳性	TP	FN
实际为阴性	FP	TN

最常用的指标有以下几种：

- 准确率(Accuracy)：准确率是最为简单、直接的一个指标，为正确预测的观测

值数量除以总观测值数量，即 (TP+TN)/(FP+FN)；

- 精确率(Precision, P)：精确率描述在预测为阳性的样本中，真实阳性的比例，即 TP/(FP+TP)；
- 召回率(Recall, R): 召回率描述在所有实际阳性样本中，预测为阳性的比例，即 TP/(TP+FN);
- F1： 精确率与召回率两者显然是存在一些矛盾：追求精确率则会牺牲一些召回率，反之反是。F1 是召回率与精确率两者之间较为平衡的一个指标，公式为 F1 = RP/(R+P)。

我们在超参数搜索中分别以 4 种计分方式为目标，即对应每个计分方式均找到一个能使其最大化的超参数组合，以满足不同的预测需要。

3 结果

训练所得模型的在评估集上的表现见下表：

表 3 模型预测结果

最优化目标	F1	精确率	召回率	准确率
数学成绩				
F1 优先	71.14%	79.35%	71.11%	68.02%
精确率优先	75.57%	83.07%	73.33%	73.70%
召回率优先	73.09%	81.06%	71.78%	70.62%
准确率优先	71.65%	80.19%	69.67%	69.44%
推理能力				
F1 优先	63.83%	74.40%	61.19%	63.46%
精确率优先	63.72%	75.51%	59.17%	65.03%
召回率优先	65.01%	74.91%	63.10%	64.21%
准确率优先	64.22%	76.11%	59.05%	65.72%

第一列表示在超参数搜索时评估模型使用的标准，即最优化目标。之后每一列表示相应的模型在评估集上相应指标的得分。

结果表明，在超参数搜索阶段采用不同的最优化目标对结果的影响不大。用该模型预测数学成绩，能够达到 80%左右的精确率，而 F1、召回率、准确率都接近 70%。对于推理能力的预测结果稍弱于对数学成绩的预测，平均能达到 65%左右的 F1, 75%左右的精确率，60%左右的召回率，以及 63%左右的准确率。模型结果尚可。

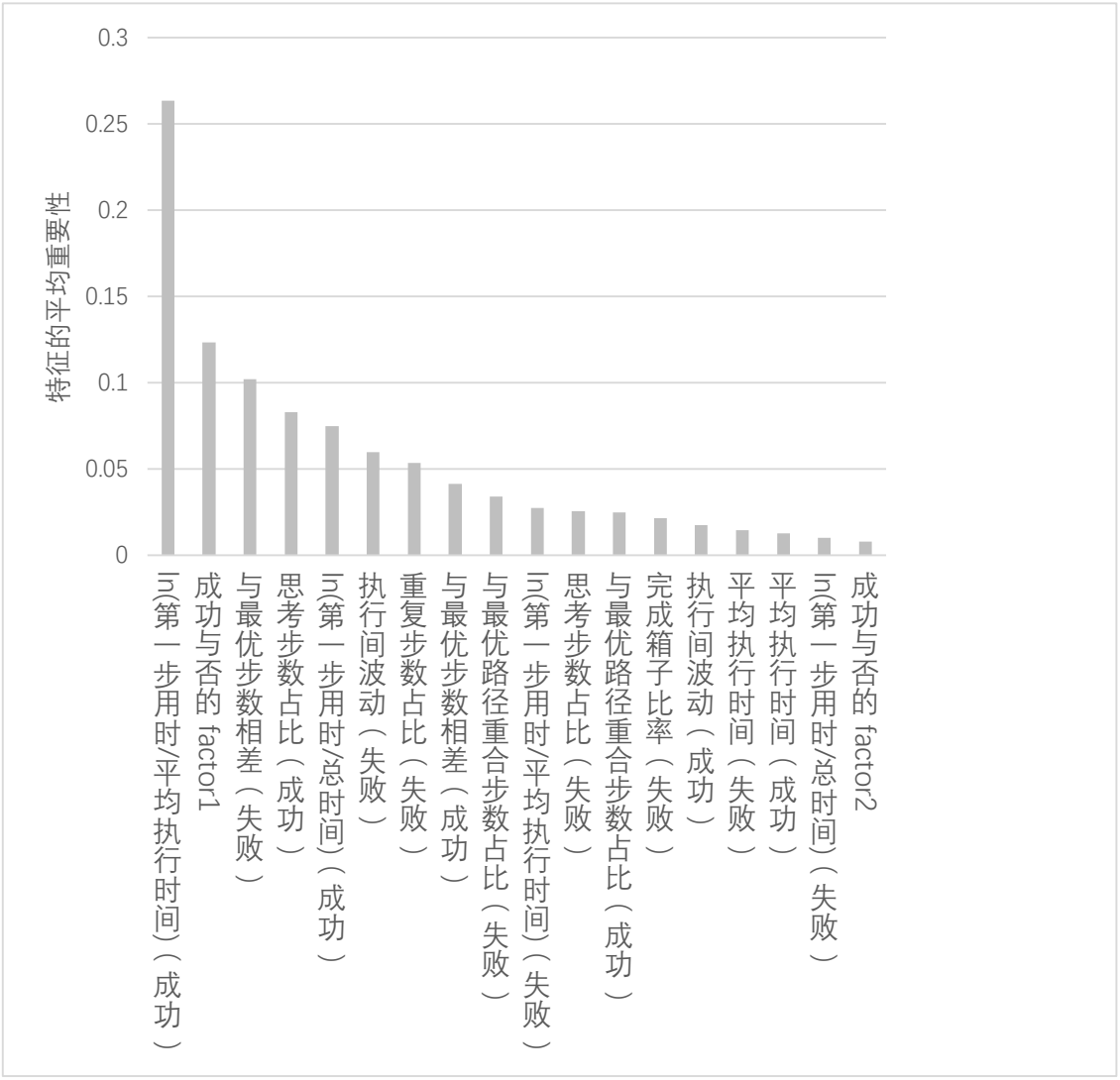

图 3 模型中特征平均重要性（前 10 位）

上图为一个模型中平均重要性前 10 的特征。特征重要性定义为（标准化后）由该特征减少的基尼不纯度 (Tan, 2006)。由图可见，第一步时间与平均执行时间的比值取对数(在成功组的平均值)在模型中的平均重要性最高，这进一步验证了 Li et al. (2015)所提出的该指标的有效性。其次重要的特征是从成功与否中通过因子分析提取出来的第一个因素。之后，与最优解相差的步数、思考时间占比、执行间波动、重复步数占比等因素均对模型有不小的贡献，证实了这些特征的有效性。

4 讨论

在本研究中，我们从推箱子的过程数据挖掘信息，使用随机森林模型预测了学生的数学成绩分类以及智力分类，预测效果尚可：我们的模型对预测数学成绩达到了 80%的精确率以及 70%左右的命中率与召回率。这表明计算机过程数据分析结合机器学习能够训练出较好的测量工具。

当然，也必须得承认，由于我们的样本中仅包含得分最高的 25%与最低的25%的被试，差异较大，在这种背景下 80%的精确度并非尽善尽美；特别是，当样本推广至所有被试后由分类任务转化为对连续变量的预测时，难度会进一步提高。我们的模型仍有提高的空间。在实际环境中，该预测结果恐怕仍然难以拿来应用，但这个结果说明使用机器学习模型结合过程数据是一个可行的策略。

与已有的研究成果，尤其是用社交网络预测人格特质(Youyou et al., 2015)等一系列研究相比，我们的预测准确性仍然偏低。这其中有可能有两方面的原因：第一，我们的研究样本量受到限制。由于社交网络的传播效果极其广泛，可以收集到非常大的样本，因此能够应用更多复杂的模型。而我们的研究受到样本量大小的限制，特征数量、模型选择上都受到限制，以避免过拟合。其次，社交网络上的信息量要更为丰富，这些信息与人格特质也有较为紧密的联系；而游戏的过程数据更多的是在一个侧面反映被试的能力，与结果变量距离更远。单从预测结果上，我们的准确率也低于 Li et al. (2016)用可穿戴设备预测情绪的，部分也是因为情绪与步行有着更密切的联系。

因此，针对以上问题，有几个环节可以改善。

一，增大样本量，提高被试之间的异质性。本研究中涉及的样本均来自于同一个中学，具有较强的同质性。在训练中使用的样本量低于 200，模型训练时为了避免过拟合，必须减少特征数量，限制了模型的拟合能力。当样本量足够大时，更多的特征能够发挥作用，提高模型的预测能力。目前所使用的推箱子工具是在windows 操作系统上实现。未来可以将其移植到线上，同时通过社交网络进行传播,迅速扩大其样本量。

二、使用与推箱子游戏更为密切的结果变量，或者使用包含更多信息量的游戏。本文所使用的结果变量：瑞文推理能力与数学成就，一定程度上能由游戏数据刻画，但相对而言其联系并非是直接的。使用一组联系更加密切的游戏与结果变量能够极大地提高预测效果。

三、提取更多的特征。之前的研究主要集中研究第一步用时占比的预测能力上。本研究进一步肯定了第一步用时占比的有效性，同时也提取了更多的指标，包括重复状态比例、思考次数等。相信过程数据中仍然有更多的信息可以去挖掘，在之后的研究中可以提取、检验更多的特征。此外，随着样本量、异质性的增大以及不同的被预测变量，部分在本研究中效果不佳的指标或许在今后的研究中能够更为有效。

四、寻找更有效的机器学习模型：在本研究探索阶段，我们还尝试了支持向量机(SVM)、决策树、逻辑斯蒂回归等模型，随机森林提供了最好的拟合效果。这些模型仍然属于常规模型，针对我们特殊的数据结构，包含嵌套分层的模型或

许能够有更好的效果。

参考文献

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning.

Guan, L., Hao, B., Cheng, Q., Yip, P. S., & Zhu, T. (2015). Identifying Chinese Microblog Users With High Suicide Probability Using Internet-Based Profile and Linguistic Features: Classification Model. *JMIR Mental Health*, *2*(2), e17. http://doi.org/10.2196/mental.4227

Kinnunen, R., & Vauras, M. (1995). Comprehension monitoring and the level of comprehension in high-and low-achieving primary school children's reading. *Learning and Instruction*.

Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., & Graepel, T. (2013a). Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning*, *95*(3), 357–380. http://doi.org/10.1007/s10994-013-5415-y

Kosinski, M., Stillwell, D., & Graepel, T. (2013b). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805. http://doi.org/10.1073/pnas.1218772110

Li, J., Zhang, B., Du, H., Zhu, Z., & Li, Y. M. (2015). Metacognitive planning: Development and validation of an online measure. *Psychological Assessment*, *27*(1), 260–271. http://doi.org/10.1037/pas0000019

Li, S., Cui, L., Zhu, C., Li, B., Zhao, N., & Zhu, T. (2016). Emotion recognition using Kinect motion capture data of human gaits. *PeerJ*, *4*(6), e2364. http://doi.org/10.7717/peerj.2364

Liu, X., & Zhu, T. (2016). Deep learning for constructing microblog behavior representation to identify social media user's personality. *PeerJ Computer Science*, *2*(1), e81. http://doi.org/10.7717/peerj-cs.81

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Pressley, M., & Afflerbach, P. (1995). Verbal protocols of reading: The nature of constructively responsive reading.

Raven, J. (1989). The Raven Progressive Matrices: A review of national norming studies and ethnic and socioeconomic variation within the United States. *Journal of Educational Measurement*.

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review*.

Tan, P. N. (2006). Introduction to data mining.

Veenman, M. V. (2013). Assessing metacognitive skills in computerized learning environments. In International handbook of metacognition and learning technologies (pp. 157-168). Springer New York.

Veenman, , Bavelaar, L., De Wolf, L., & Van Haaren, M. G. P. (2014). The on-line assessment of metacognitive skills in a computerized learning environment. *Learning and Individual Differences*, *29*(C), 123–130. http://doi.org/10.1016/j.lindif.2013.01.003

Veenman, , Wilhelm, P., & Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction*, *14*(1), 89–109. http://doi.org/10.1016/j.learninstruc.2003.10.004

Veenman, M. V., & Alexander, P. (2011). Learning to self-monitor and self-regulate. Handbook of research on learning and instruction, 197-218.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036–1040. http://doi.org/10.1073/pnas.1418680112

Zhang, Z., Song, Y., Cui, L., Liu, X., & Zhu, T. (2016). Emotion recognition based on customized smart bracelet with built-in accelerometer. *PeerJ*, *4*(2), e2258–14. http://doi.org/10.7717/peerj.2258

致谢

这份毕业论文太像是我四年学习的总结：大一的时候我在信科学计算机，大二以来我同时学习经济与心理。大一的训练让我有扎实的编程基础，让我能够顺利写完整个项目的代码；经济学的训练让我对社会科学的应用统计和计量方法有深入的理解和熟练运用，处理数据时应心得手；信科与经济的训练提供了technique，而心理学的训练赋予了我在做这个研究的时候的 sense：我能理解这个题目的意义，也能判断什么特征可能是有效的，什么特征是无关的。3 方面训练，缺了任何一方面，这个研究恐怕都无法如此顺利地执行下来。

尽管现在，我已经决定专精计算机与经济，放弃在心理学上进一步深造了。这篇毕业论文也恐怕是我和心理学一个告别。

做出这个决定的原因非常多，有功利的考量，也有情怀的追求。但需要强调的是，我并没有失去对人类行为研究的兴趣；相反的是，经济学让我看到了其对人类行为研究中更强的解释力：一个理性人假设，就已经能刻画人的绝大多数行为模式了。

在完成这篇毕业论文的过程中，需要感谢的人太多。

最要感谢的是我的导师黎坚老师。他从我刚转到心理学院时便接纳我，带我做研究；到我决定主要做经济学时是他都是充分的理解、支持；在我离开组会一整年，毕设开题紧迫，仓促间找到他时，他也没有因为我这一年的离开对我失去信心，依然给了我一个非常新颖又诱惑的想法，并在过程中给了非常宝贵的指导；

其次要感谢的是好友张凌祺，他提供了强有力的机器学习的技术指导，堪称我的第二导师。从跟他提起这个项目起，他便提供了大量的机器学习的学习资源，以及给了完善的解决方案，帮助我在理论上到技术上从懵逼到入门；我在项目完成过程中，有关特征的选择、模型的设定、调整无不与其探讨，他给了我非常多宝贵的意见；张同学即将前往宾大攻读计算神经科学的博士；差一点就能在一个城市，非常遗憾，但好在也并不远，机票不贵。

以及好友聂卓、任昶宇、杨笑寒、沓钰琪、曾莹、王震。孤军奋战 3 年半，终于在最后半年找到了组织。能有一群人一起浪不难，酒肉朋友随便找；能有一群人一起早起学习也简单，豆瓣打卡小组就能够满足；但是能有这么一群人一起浪完一起学习，能开车也能聊学术，实在是个小概率事件。多亏了他们，做毕设的时间过得很愉快。很荣幸、很幸运能在本科的尾巴上遇到他们。