1 秒待つ：最適な認知資源配分のためのブースト設計
One-Second Boosting: Intervention that Promotes the Optimal Allocation of Cognitive Resources

香川 璃奈[1]，白砂 大[2]，池田 篤史[3]，讃岐 勝[1]，本田 秀仁[2]，野里 博和[4]
Rina Kagawa[1], Masaru Shirasuna[2], Atsushi Ikeda[3], Masaru Sanuki[1], Hidehito Honda[2], Hirokazu Nosato[4]

[1]筑波大学，[2]追手門学院大学，[3]筑波大学附属病院，[4]産業技術総合研究所
[1]University of Tsukuba, [2]Otemon Gakuin University, [3]University of Tsukuba Hospital, [4]National Institute of Advanced Industrial Science and Technology
kagawa-r@md.tsukuba.ac.jp

概要

昨今の社会情勢の変化やクラウドソーシングプラットフォームの隆盛により、分業の成果物の質の担保が課題となっており、成果物の質を向上させるための介入の需要は高い。我々は、作業者には適切な思考時間が存在するという仮定に立ち、医療画像に異常所見の有無を付与する際に、画像を提示してからある一定の時間､回答できない状態にする介入(ブースト)の効果を検証した。医師(N=730)を対象とした行動実験により、画像を提示してから 1 秒間(介入がない場合の思考時間の中央値より短い時間)だけ回答できない状態にすると正答の期待値が上昇したことを確認した。

キーワード : ブースト, 資源合理性, 機械学習, 医療画像

1. 背景・目的

昨今の社会情勢の変化やクラウドソーシングプラットフォームの隆盛により、特にインターネット上における非同期的な分業の重要性は今後も増すであろう。その一方で、分業の成果物の質の担保が課題となっており[1]､成果物の質を向上させるための介入の需要は高いと考えられる。

我々は、できるだけコストをかけずに、かつ、作業対象を問わず適用できる介入が必要だと考えている。そこで、選択において明確なかたちで情報を理解させることを促し、その結果として継続的に認知能力を引き出すことを目的とした介入であるブースト(boosting)[2]に着目した。

本研究では、インターネット上の分業の例として、統計的学習用の正解ラベル付きデータセット作成に焦点をあてる。統計的学習の技術開発のためには、一般的に、正確にラベルが付与された正解データセット（対象データと正解ラベルのセット）が大規模に必要となる。そのラベルが 1 対 1 で決定されるような明確なラベル付与タスクであれば問題ないが、正解ラベルを規定する基準が一意に定まっていないタスクの場合、正解ラベルを均一に大量に揃えることは難しい。特に、医療データに代表されるような専門性の高いデータが対象になる場合、病変の有無を診断するようなタスクのラベル付与基準は作業者の知識や経験に依存する場合が多く、それらを補うための十分な教育にコスト（時間や費用）が掛かる。近年の統計的学習の需要の高まりを考えると、大量な学習データ作成のためのラベル付与作業にコストを掛けることは望ましくなく、できるだけコストを掛けずに作業品質を向上させるための介入方法の需要は高いと考えられる。

ここで我々は、資源合理性[3]の観点を鑑み、作業者がラベル付けの問題に取り組む時間の延長による正答率の向上と回答者のメンタルワークロードの負荷はトレードオフの関係にあると予想した。

そこで本研究では、作業者が能力を発揮できる最適な時間のかけ方があるという前提に立ち、問題を提示してからある一定の時間、回答できない状態にする、という簡便な制約をブーストとして設定することで作業者集団の作業能力が向上する、という仮説を検証した。

2. 方法 (行動実験)

本研究では、医療画像(膀胱鏡画像)に対して正常(画像中に病変を含まない)か異常(画像中に病変を含む)かを回答する 2 値分類タスクを題材に実験を実施した(図 1)。問題を提示してから回答できるまでの時間制約を、本研究では待ち時間と呼ぶ。問題を提示してからの待ち時間がラベルづけの品質に与える影響を検討した。

実験参加者: 約 730 名の医師が参加し、最終的に 404 名の医師が回答を完了した(男性 354 名、女性 50 名、Mage = 50.3、SDage = 12.0)。

図１　実験刺激の概要

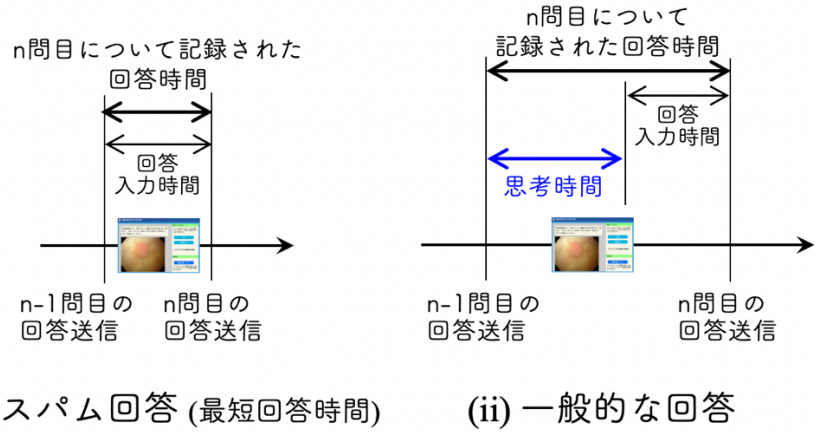

図２　回答時間と思考時間の計算方法の概要

刺激・手続き: 延べ 7 種類(待ち時間なし、1.0、2.0、3.0、4.0、5.0、6.0 秒)の待ち時間を、実験に参加した医師に対してランダムに割り当てた。全ての実験参加者は、3 問のチュートリアルのあとに 100 問の 2 値分類タスクに回答した。ただし、参加者の判断で任意のタイミングで中途離脱が可能であり、離脱したあとに再開はできない。

実験はすべて web 画面上で行われた。実験参加者である医師には、実験設定、病変枚数、待ち時間による介入を設けていることなどは明らかにしていない。各問題間には固視点が 1 秒間呈示された。

本実験は、筑波大学医学医療系医の倫理委員会の承認(承認番号 1616)を得ている。筑波大学に知財登録(筑大産知財第 19-608 号)されている画像 103 枚 (正常と異常が 50 枚ずつ) を実験に利用した。

分析: 待ち時間の設定がデータ品質に与える影響を評価するため、正答の期待値を算出し、アノテーション品質の評価基準として利用した。多くのアノテーターを募集でき、１つのデータに対して複数のアノテーターを割り当てられる場合には、個々のデータに対する正答数の期待値が高いことが望まれると考えたためである。そこで、本実験では、待ち時間 t (秒) における正答の期待値を$Perform(t)$と呼び、待ち時間 t (秒)における平均正答率$Acc(t)$と待ち時間 t (秒)における回答を完了した人数 $n(t)$ に基づき、以下と定義した。

$$Perform(t) = Acc(t) \times n(t). \quad (式\ 1)$$

さらに、回答者の思考過程をより直接的に分析するため、回答時間から思考時間を算出して、解析に利用した[4]。実験で使用したシステムにおいて回答者が１問に回答するまでの思考過程を図 2 に示す。このとき、記録された回答時間は、タスクが表示されてから回答者が答えを決定するまでの思考時間と、決定した答えのボタンをクリックする回答入力時間に分割することができる。ここで、オンライン実験には、一定数のスパム回答（例えば、クリックを連打するなど、問題に対して能動的な思考が伴わない回答）が含まれることを避けられない特性を利用して、記録された回答時間のうち最短の時間が回答入力時間(図2(i))であるとみなした。また、一般的にクリック動作にかかる時間には回答者間で大きな差がないことから、回答入力時間は一定であると仮定し、図 2(i)と図 2(ii)の差分から、各回答の思考時間を算出した。

表 1: 参加者のデモグラフィックデータと結果の概要.

待ち時間 (t)	100 個のタスクの完答者数$n(t)$	女性	年齢	$Acc(t)$	$Perform(t)$
なし	**66**	13.6%	50.0 (± 12.9)	68.50%	45.21
1.0	65	9.2%	50.9 (± 10.6)	71.14%	**46.24**
2.0	55	14.5%	51.2 (± 10.6)	72.02%	39.61
3.0	61	14.8%	51.5 (± 12.0)	69.81%	42.58
4.0	58	12.1%	49.1 (± 13.4)	69.92%	40.55
5.0	50	14.0%	49.9 (± 13.0)	**72.50%**	36.25
6.0	49	10.2%	50.3 (± 11.4)	71.14%	34.86

3. 結果・考察

待ち時間を 1.0 秒以上設定することで、待ち時間がない場合と比較して平均の正答率が上昇する傾向が認められた。さらに、待ち時間がない場合と比較すると、1.0 秒待ちにおいては$Perform(t)$が高くなる一方で、2.0 秒以上の待ち時間を設定すると待ち秒数が長くなるほど中途離脱者が増加した結果として$Perform(t)$の値が減少することを確認した(表 1)。さらに、延べ 7 種類(待ち時間なし、1.0、2.0、3.0、4.0、5.0、6.0 秒)の待ち時間の各条件における思考時間の中央値は順に、1.52、1.40、1.30、1.43、1.38、1.36、1.35 秒であった。このことから、1.0 秒待ちという介入は、待ち時間を設定しない場合の思考時間の中央値(1.52 秒)よりも短い時間、と解釈することができた。

本研究で提案した、待ち時間を設定しない場合の思考時間の中央値よりも短い時間だけ回答できないようにする、という介入は、タスクを問わず簡便に実現できる介入であり、多様なタスクに応用できる可能性がある。今回得られた結果がタスクに依存した結果ではない事を確認するために、より多様なタスクを利用して背景知識が異なる回答者(看護師など)での同様の検討も必要である。思考時間におけるより詳細な思考過程を計測するために、オンライン環境下で簡便に計測できるマウス軌跡などのデータから回答者の思考過程を推測する技術開発にも需要があると考える。

謝辞

本研究成果の一部は、国立研究開発法人新エネルギー・産業技術総合開発機構(NEDO)の委託業務(JPNP20006)、JST 未来社会創造事業の助成(JP19211284)の結果得られた。

文献

[1] Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., & Allahbakhsh, M. (2018). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Comput. Surv.*, *51*(1), 1–40.

[2] Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. Perspectives on Psychological Science, 12(6), 973–986.

[3] Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. Top. Cogn. Sci., 7(2), 217–229.

[4] Dutilh, G., & Rieskamp, J. (2016). Comparing perceptual and preferential decision making. Psychonomic bulletin & review, 23(3), 723–737.