専門的データとシミュレーションを用いた Hyper Question の有効性の検討

田丸陽稀[a] 藤崎樹[b] 植田一博[c]

要約

個々人の意見を集約した結果が時に高い精度を記録する現象を集合知と呼ぶ．こうした集合知に関する研究の中で，成績の良いメンバーを抽出し，その意見を集約することで意思決定の精度をさらに高めようとする手法を少数選抜と呼ぶ．本稿では，少数選抜の一手法である Hyper Question の精度と限界を実際の実験参加者のデータに基づいて検討した．Hyper Question は，成績優秀者の回答の類似性に着目し，正答を利用せず成績優秀者を抽出する手法を提供する．本稿ではまず，Hyper Question で提案される少数選抜方法が従来検討されてこなかった二択課題で，かつ回答データがデンスな場合で多数決と同程度の精度を示すことを明らかにした．次に，計算機シミュレーションを用いて問題の選択肢数や回答者集団の能力を変化させた上で，従来手法の代表例の多数決と比較することで，その精度を検討した．結果，問題の選択肢が多い時，また回答者の能力の分散が大きい時，高い精度を示すことを明らかにした．

JEL 分類番号： C69 D70 D82

キーワード：集合知， 少数選抜， ヒューマンコンピュテーション

[a] 東京大学教養学部学際科学科 tamaru-haruki327@g.ecc.u-tokyo.ac.jp

[b] 東京大学大学院総合文化研究科学術研究員 bpmx3ngj@gmail.com

[c] 東京大学大学院総合文化研究科教授 ueda@g.ecc.u-tokyo.ac.jp

1. はじめに

集団は，時に極めて優れた意思決定を行う場合があることが明らかとなっている．例えば Galton(1907)は，牛の体重を予測するイベントで，参加した人々の回答を集めて分析した．結果，参加者の回答の集約値（平均値など）がどの参加者の回答よりもよい精度を示し，正解に極めて近いことが明らかとなった．このように，集団が高い正確性を記録する現象は集合知(“The wisdom of crowds”; e.g. Surowiecki, 2004)と呼ばれ，情報技術の発展とともに集団の意見を集めやすくなった近年，盛んに研究が行われている．

こうした集合知研究の中で，近年よく進められている分野の一つが，少数選抜と呼ばれるものである．これは，集団の中から成績の良い少数のメンバーを選抜することによって集合知の精度をさらに高めるという手法である (e.g. Budescu and Chen, 2015)．このような少数選抜は，問題の正答があらかじめ分かっていれば，成績優秀者を特定することを可能にする (e.g. Von Ahn et al., 2008)．しかし，正答のわかっていない問題に対しては，利用することができないという問題を抱えている．

このような課題を克服する手法も近年検討されており，その例の一つとして Li et al. (2017)の提案した Hyper Question が挙げられる．この手法は，問題に詳しい回答者（つまり成績優秀であろう回答者）同士の回答が類似することに注目している．例えば，ポケモンに関する問題に回答する時，ポケモンに詳しい回答者の回答は，ポケモンに詳しくない回答者の回答と比較して，おおむね類似するであろう．こうした特徴を利用するべく，Hyper Question ではいくつかの問題をセットにし，個々の問題における回答が正答かどうかではなく，セット単位での回答者の回答の類似性によって，問題に詳しい回答者を抽出することを試みる．この手法について，Li et al. (2017) は次のように説明する．まず，全体で Q 個の問題から k 個の問題の集合（k-Hyper Question と呼ぶ）を作ることを考える．このとき全体で $\binom{|Q|}{k}$ 個の k-Hyper Question が作られる．次に，この k-Hyper Question に対して統計的分析を行う（例えば，多数決を用いる Hyper-MV なら k-Hyper Question に対して多数決を行う）．こうして k-Hyper Question を決定する．次に，選ばれた k-Hyper Question を元の単一問題にデコードする．以上のようにして成績優秀者の解答のみを選抜し，最後に単一問題に対して統計的決定を行う（Hyper-MV なら多数決）．以上のような Hyper Question は，従来良い精度を残してきた多数決と比べても高い精度を誇ることが明らかとなっている．

このように Hyper Question は正解を用いない優れた少数選抜手法であるが，いまだ検討の余地があると考えられる．まず，先行研究では，ポケモンや中国語などに関する問題を用いていた．それに対してより，専門的な問題，例えば CT スキャン画像の判断といっ

た課題は扱われていない．また，問題も多択（5 択など）なものが検討されてきた．一方で，Gigerenzer and Goldstein (1996) や Hertwig et al. (2008)に代表され，人間の意思決定の文脈で盛んに研究されてきた，2 択の問題については分析されていない．さらに，回答者が回答をスキップすることを許容するスパースな回答データを扱っており，デンスなものは扱われてこなかった．加えて，回答者集団がどのような特徴を持っていればその精度がさらに上昇する（低下する）かについて十分検討されていない．本稿では，Hyper Question が，上述した様々な条件で，従来手法の代表例である多数決と比べて，どの程度の精度を発揮するかを検討する．

2. 実験

2.1. 専門的な問題を用いた Hyper Question と多数決との比較

まず，専門的な問題対象に，Hyper Question と従来手法の代表例である多数決との比較を行った．Kurvers et al. (2016)で用いられた皮膚癌のデータを用いた．具体的には，医者 40 人に CT スキャンの画像 108 枚を見せ，そこに癌細胞が写っているかを判断してもらい，実際に癌細胞があったかどうかを回答させたデータである．したがって，このデータは，Li et al. (2017)で検討されたような，多択（5~6 個の選択肢からなる）問題の回答データではなく，2 択の中から必ず答えを選択させた結果得られたものである．加えて，回答をスキップすることを許容しないため，回答データはデンスなものである．

このデータを，多数決と Hyper Question を用いた手法（以下，Hyper-MV と表記）を用いて分析し，それぞれの正答率を算出した．なお，メンバーを選抜する問題のセットとして，2 問からなるセットを設定した．

2.2. 結果

結果を以下の表 1 に示す． Hyper-MV の正答率は 0.91 であり，多数決の正答率である 0.90 と同程度の精度であることが明らかとなった．

表 1

手法	正答率
多数決	0.90
Hyper-MV	0.91

このように，2 択問題に対するデンスな回答データをもち，かつ専門的な事象を扱った問題では，Hyper Question は多数決と同等の精度を示すことが判明した．

次節では，こうした Hyper Question の比較がどこまで保たれるかをより詳しく検討する．具体的には，多様な状況を題材に，コンピュータシミュレーションを通じて多数決と Hyper-MV の比較を行う．

2.3 コンピュータシミュレーションを用いた多数決と Hyper-MV の比較

前節では，二択の問題については Hyper-MV は多数決と同等の精度であった．しかし，従来の検討では，多択の問題について，Hyper-MV が多数決に対してより良い精度を示している．そこで，次にどのような条件まで Hyper-MV の多数決に対する優位性が保たれるかを検討する．ここでは，回答者の能力と問題の選択肢を変化させることで検討する．そのためコンピュータシミュレーションを用いて，Hyper-MV と多数決の比較を行う．具体的には，まず回答者集団内に専門的な知識を持つ回答者（専門家）と専門的知識を持たない回答者（非専門家）を仮定する．回答者は全体で 20 人であり，そのうち専門家の人数を{2,4,6}人の中からランダムに選んだ．専門家の正答率はランダムに{0.8,0.9,0.95,1.0}から選ばれ，非専門家の正答率は正規分布からランダムに選ばれた（こうした設定は (Li et al., 2017) に倣った）．この正規分布の平均（0 から 1 まで 0.1 ずつ増加）と分散（0 から 0.25 まで 0.05 ずつ増加）を変化させ，その時の両手法の精度を検討した．以上の条件の下で，非専門家集団の正答率の平均と分散，および問題の選択肢数を変化させることで，Hyper-MV と従来手法である多数決の比較を行った．シミュレーションとしては，平均，分散を決定した後，各条件で 100 回ずつ行い，その平均値をそれぞれの手法の正答率とした．

2.4 結果

結果を図 1 に示す．問題の選択肢 n が大きいとき，あるいは分散 v が大きい時は，Hyper-MV が多数決に対し優位性を示している．具体的には，非専門家の正答率が低い時に特に Hyper-MV の精度が多数決よりも優れている．しかし，問題の選択肢が少なくなるにつれ，Hyper-MV の精度が多数決に近しくなることが読み取れる．

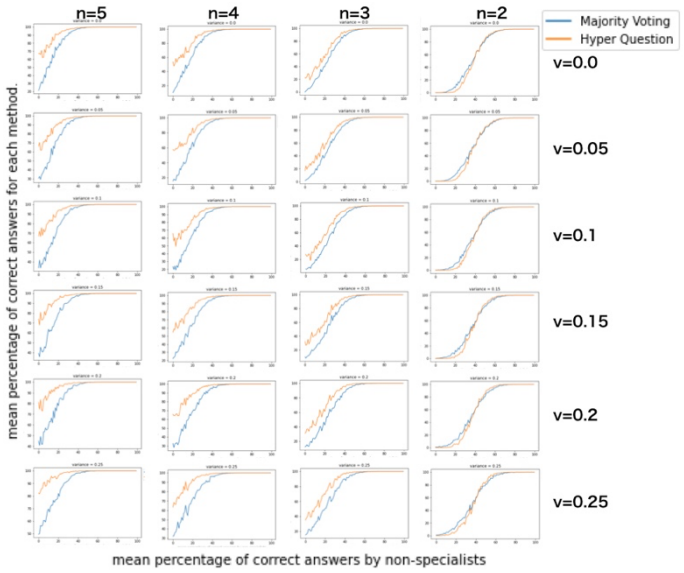

図 1　Hyper-MV と多数決をコンピュータシミュレーションにより比較したグラフ．各グラフの横軸は専門的知識を有しない回答者集団の正答率，縦軸は手法の正答率，青線は多数決，オレンジ線は Hyper-MV をそれぞれ表している．また，各グラフは，横方向に問題の選択肢数（左から 5 択，4 択，3 択，2 択），縦方向に専門的知識を有しない回答者集団の正答率の分散（上から v=0，0.05，0.1，0.15，0.2，0.25）に応じて並べられている．

3. 考察

本研究では，まず実データに基づいて Hyper-MV と多数決を比較した後，さらにコンピュータシミュレーションで両手法の精度を比較検討した．その結果，実データによる比較では Hyper-MV が多数決とほとんど同等の結果を出した． 次に，シミュレーションを用いて Hyper-MV と多数決の検討を行った．結果，選択肢が多い時は Hyper-MV は多数決に対し優位性を示すことが判明した．しかしながら，選択肢の数が減少するにつれその優位性は減退し，特に 2 択の時は多数決と同程度の精度であることが明らかとなった．この内容は実データの結果と一致する[1]．また，シミュレーションにおいて，回答者の正答率の分散が大きい時，Hyper-MV が多数決に対して優位性を示すことが判明した．こうした結果から，Hyper Question の短所と長所を検討したい．

まず，Hyper Question の短所として，選択肢が少ない時の精度が減退することが挙げられる．これは次のようなメカニズムだと考える．非専門家が回答を間違える時，その回答は正解でない選択肢のどれかに均等に分散する．それゆえ，専門家以外の回答に適度に差が生まれ，専門家の回答のみが選択される．しかし，正解でない選択肢の数が減少すると，非専門家の回答同士も類似し始める．結果として，正答率の低い回答者の回答を選択してしまい，手法の成績が低下する．Hyper Question は計算量が多いため，多数決と同等の精度であれば多数決を使う方が効率的となる．一方で，Hyper Question の長所として，非専門家の正答率の低い時に多数決よりも高い精度を発揮することが挙げられる．問題の選択肢数が少なくなった時にも，こうした長所を発揮することができれば，Hyper Question をより多様な状況で扱うことができよう．手法の改善のためには，こうした方向の開発が有効であると思われる．

4. 結論

本稿では，少数選抜の手法の一つである Hyper Question について検討した．具体的には，人の判断，意思決定をよく説明するとされる 2 択課題に対するデンスな回答データについ

[1] 但し，実データは回答者全員が医者であり，シミュレーションのように非専門家がいないことについては注意する必要がある．

て，実データを用いて検討した．その結果，そのようなデータに対して，Hyper Question が多数決とほぼ同等の精度を示した．さらに，より広い条件についてコンピュータシミュレーションを行った結果，同様な結果が得られ，Hyper Question の適用範囲の広さが示唆された．特に，選択肢の数が多い方が，また専門的知識を持たない回答者の正答率の分散が高い時に，多数決よりも高い正答率を示すことが明らかになった．今後は，コンピュータシミュレーションにおいてあまりよい結果を示さなかった条件に対しても優れた精度を発揮する手法の開発が望まれる．具体的には，問題選択肢数が少ない時においても，非専門家の正答率が小さい時に良い精度を残すような手法の開発が望まれる．

引用文献

Budescu, D. V. and Chen, E., 2015. Identifying expertise to extract the wisdom of crowds. Management Science, 61(2), 267-280.

Galton, F., 1907. Vox populi (the wisdom of crowds). Nature, 75(7), 450-451.

Gigerenzer, G. and Goldstein, D. G., 1996. Reasoning the fast and frugal way: models of bounded rationality. Psychological review, 103(4), 650.

Hertwig, R., Herzog, S. M., Schooler, L. J. and Reimer, T., 2008. Fluency heuristic: a model of how the mind exploits a by-product of information retrieval. Journal of Experimental Psychology: Learning, memory, and cognition, 34(5), 1191.

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., Giuseppe, A., Iris, Z. and Wolf, M., 2016. Boosting medical diagnostics by pooling independent judgments. Proceedings of the National Academy of Sciences, 113(31), 8777-8782.

Li, J., Baba, Y. and Kashima, H., 2017. Hyper questions: Unsupervised targeting of a few experts in crowdsourcing. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 1069-1078.

Surowiecki, J., 2004. The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations, Doubleday

Von Ahn, L., Maurer, B., McMillen, C., Abraham, D. and Blum, M., 2008. Recaptcha: human-based character recognition via web security measures. Science, 321(5895), 1465-1468.