DEIM Forum 2018 G6-4

超問題：専門知識を要するクラウドソーシングタスクの回答統合法

李　吉屹† 馬場　雪乃†† 鹿島　久嗣††,†††

† 山梨大学コンピュータ理工学科　〒 400-8511 山梨県甲府市武田 4-3-11
†† 京都大学大学院情報学研究科知能情報学専攻　〒 606-8501 京都府京都市左京区吉田本町
††† 理化学研究所革新知能統合研究センター　〒 103-0027 東京都中央区日本橋 1-4-1
E-mail: †jyli@yamanashi.ac.jp, ††{baba,kashima}@i.kyoto-u.ac.jp

あらまし　クラウドソーシングの品質管理の文脈において様々な回答統合法が多数提案されているが，多くの方法は多数派の意見を強調するように働くため，大多数のワーカが正しく回答できないような難しい問題では正しい回答を導くことができない．本研究では，多数決が失敗する難しい問題での回答統合に対する有効なアイデアとして「超問題」を提案する．超問題とは，複数の問題をまとめてひとつの問題とみなしたものである．専門家は超問題への回答において非専門家よりも合意する可能性が高いため，専門家が多数派となる可能性が高まる．人工データと実データを用いた実験で，専門家が少数しか存在しないような状況において，提案法が効果的であることを示した．
（なお，本論文は既発表文献 [9] に基づく.）

キーワード　クラウドソーシング, ヒューマンコンピュテーション, 品質管理, 回答統合

1. はじめに

不特定多数の人々に作業を依頼する手段であるクラウドソーシングは，科学やビジネスの様々な場面で活用されている．クラウドソーシングの作業者（ワーカ）は，能力・注意力の不足，意欲の欠如など様々な理由により正解を提示しないことが多いため，品質管理はクラウドソーシングの大きな問題である．この問題への対処法の一つは，冗長性の導入である．つまり，複数の異なるワーカに同じ問題を割り当て，その回答を統合して信頼性を向上する手法である．クラウドソーシングで依頼される典型的な作業の一つに，画像へのタグ付けや科学的質問への回答など，特定の質問に対して複数の候補から一つの回答を選択させる多肢選択問題がある．多肢選択問題に対して，多数決のような単純な統合手法に加え，より洗練された統計的手法が提案されている [1,2,6,10,13,15–19]．ワーカの能力と正解はどちらも未知であるため，これらの手法では，ワーカの能力と正解を相互に推定する．つまり，多くの問題で多数派となるワーカは能力が高いとみなし，また，そのようなワーカの回答は正解である確率が高いとみなすことで，能力と正解を推定する．これらの回答統合法は，多数派の意見を強調していると言える．正答するワーカが多数派となる場合では，多数派の意見を強調する方法は有効であるが，図 1 に示すような高度な専門知識を必要とする問題では，誤答するワーカが多数派となる．このような場合には，能力の低いワーカを誤って「能力が高い」と推定することになり，従来の回答統合法は失敗してしまう．

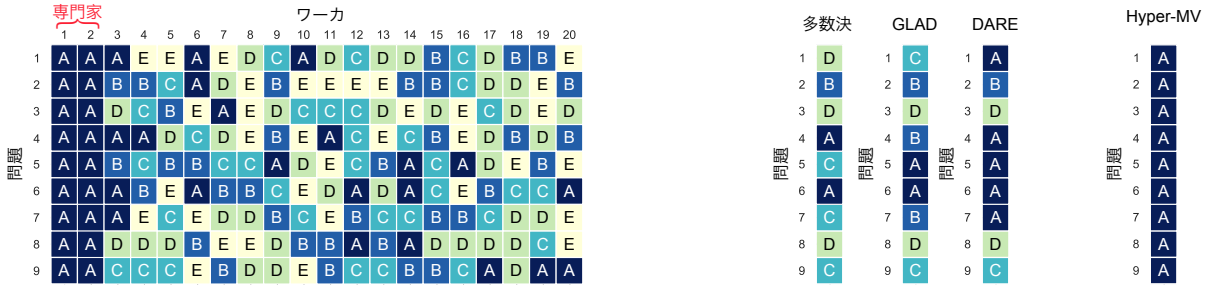
図 2(a) に，従来の回答統合法の失敗例を示す．この例では，9 件の質問に対して 20 人のワーカが回答している．20 人のうち，ワーカ 1 とワーカ 2 は常に正答し，残りの 18 人はランダムに回答する．図 2(b) に，従来の回答統合法の適用結果を示す．誤答するワーカが多数派となるため，多数決や統計的手法（GLAD [18] と DARE [1]）では，いくつかの問題で正解を得られない．

> 質問：次の薬のうち，長期使用によりクッシング症候群を引き起こす可能性が最も高いものはどれか？
>
> （a）ヘパリン（b）インスリン
> （c）テオフィリン（d）プレドニゾロン

図 1: 多数派が誤答する可能性が高い専門的な問題の例

我々は，多数派が誤答するような難しい問題を対象にした，新しい回答統合法を提案する．ここで，正答率の高いワーカを専門家，正答率が低いワーカを非専門家と呼ぶ．難しい問題においては，非専門家が多数派となるため，従来の回答統合法は失敗する．我々の提案法のキーアイデアは，個々の問題ではなく，複数の問題に着目することである．専門家は多くの問題に正答するが，言い換えると，ある複数の問題を考えたときに，専門家はその全てで正答する確率が高い．一方で非専門家は，個々の問題にたまたま正答する可能性はあるが，全てで正答する確率は低い．また，専門家は全ての問題で正答しやすいため，専門家同士の回答は非専門家同士に比べて一致しやすい．このような，複数の問題に対して，専門家同士が非専門家同士よりも合意しやすいという観察にもとづき，複数の問題をまとめて一つの問題とみなした「超問題」を導入し，超問題の上での多数決法である Hyper-MV を提案する．図 2(c) の例では，提案法がすべての問題に対して正しい統合結果を出力することを示している．また，超問題は，多数決以外の手法にも適用できる．GLAD [18] と DARE [1] に超問題を組み込んだ，Hyper-GLAD と Hyper-DARE を提案する．

人工データと実データの両方を使用した実験により，我々の手法が，少数の専門家しか存在しない場合に高い精度を達成す



(a) (b) (c)

図 2: 従来法が失敗する例：(a) ‘A’〜‘E’ の 5 つの選択肢を持つ 9 件の多肢選択式問題（全て正解は ‘A’）に対し，20 名のワーカから回答を集める．20 名中 2 名は専門家であり，全ての問題に正答する．残りのワーカは，5 件の選択肢からランダムに一つ選んで回答する．(b) 多数決では正解は得られず，より高度な統計的手法（GLAD [18]，DARE [1]）でも多くの場合に失敗する．(c) 一方，提案法（Hyper-MV）では，全ての問題で正解を得ることができる．

ることを確認した．特に我々の手法は，まず全てのワーカを少数の質問でテストし，その結果を利用して専門家を見つけるようなシナリオで特に有用である．

本論文の貢献は以下の 3 点にまとめられる：

（1） 多数決が失敗する，多肢選択問題のクラスを示した．

（2） 少数の専門家しか存在しない場合に専門家の意見を強調するためのアプローチ「超問題」を提案した．

（3） 超問題に基づく回答統合法として，Hyper-MV，Hyper-GLAD，Hyper-DARE を提案し，専門家が少数しか存在しない状況における提案法の有効性を示した．

2. 関 連 研 究

クラウドソーシングにおける回答統合法は広く研究されており，ワーカの能力や問題の難易度を推定して正解推定に用いる統計的手法が多数提案されている [1, 2, 6, 7, 10, 13, 15–19]．これらの手法は多数派のワーカの意見を強調する傾向があり，正しく答えられる専門家が少数の場合には失敗することがある．

専門家が少数の場合に，補助情報を利用して専門家を発見する手法が研究されており，ワーカの学歴などの属性に基いて専門家を見つける手法や [8]，問題の説明文を利用してトピックごとのワーカの能力を捉える手法 [11]，ワーカが検索エンジンで用いるクエリ情報を用いて専門性を推定する手法 [4] などが提案されている．また，正解既知の問題を用意して，それに対する正答率で専門家を見つける手法も広く用いられている [3, 5, 12]．しかし，補助情報や正解既知の問題を常に用意できるとは限らない．我々は，回答情報だけから専門家を見つけるという汎用的な問題設定を対象にする．

提案法は，専門家同士の回答は相関するという前提に基づくが，マルチラベル分類においても，ラベル同士の相関に関する同様の前提に基づく手法が提案されている．提案法で用いる「超問題」のアイデアは，いくつかのクラスをまとめたメタクラスを考えるマルチラベル分類手法 RAkEL [14] と類似している．

3. 問 題 設 定

本論文では，典型的なクラウドソーシングタスクである多肢選択問題を扱う．多肢選択問題では，ワーカは与えられた選択肢の中から一つを選んで回答するよう求められる．また我々は，ワーカが一つの多肢選択問題だけに回答するのではなく，複数の問題に回答する場合を対象とする．このとき，全ての問題で選択肢が共通する場合と，共通するとは限らない場合の二種類が考えられる．前者の例は「画像に猫が写っているかどうか」を答えるタスクである．このタスクでは，どのような画像でも選択肢は「(a) 猫が写っている, (b) 猫が写っていない」となる．後者の例としては，ある日本語に対応する外国語を選択させるタスクがある．「ドイツ語で『はい』を意味するのはどちらか？(a) Ja，(b) Nein」「ドイツ語で『ありがとう』を意味するのはどちらか？(a) Bitte，(b) Danke」という二つの問題では，選択肢が異なっている．本研究では，より汎用的な設定である，後者の「選択肢が共通するとは限らない」場合を対象とする．

本研究で扱う回答統合問題は，次のように定式化される．ワーカ集合 $\mathcal{W}$ と問題集合 $\mathcal{Q}$ が与えられており，各問題 $q \in \mathcal{Q}$ について，選択肢集合 $\mathcal{C}_q$ が存在する．ワーカ $w \in \mathcal{W}$ の問題 $q \in \mathcal{Q}$ に対する回答を $l_{wq} \in \mathcal{C}_q$ で表す．また，回答集合を $\mathcal{L} = \{l_{wq}\}_{w \in \mathcal{W},\, q \in \mathcal{Q}}$ で表す．我々の目的は，$\mathcal{L}$ が与えられた下で，各問題 $q \in \mathcal{Q}$ の正解 $t_q \in \mathcal{C}_q$ を推定することである．

4. 超問題に基づく回答統合

4.1 超 問 題

従来の回答統合法は，多数派の意見を強調する性質があるために，多数派が誤答する難しい問題では失敗することがあった．我々は，このような問題において，正しく回答できる少数の専門家の意見を強調し正しい回答統合結果を得るために，「超問題」とそれに基づく回答統合法を提案する．

超問題は，問題集合 $\mathcal{Q}$ の部分集合である．特に，要素数 k の超問題を「k-超問題」と表記する．例えば，問題集合として $\{1, 2, 3, 4\}$ が与えられたときに，その 3-超問題は，$\{1, 2, 3\}, \{1, 2, 4\}, \{1, 3, 4\}, \{2, 3, 4\}$ である．また，超問題に対する回答を，超問題を構成する単一問題に対する回答の連結で表す．つまり，あるワーカが問題 1 に ‘A’, 問題 2 に ‘B’, 問題 3 に ‘C’ と答えている場合，超問題 $\{1, 2, 3\}$ に対する回答は ‘ABC’ となる．

専門家は超問題を構成する全て問題において正答する可能性

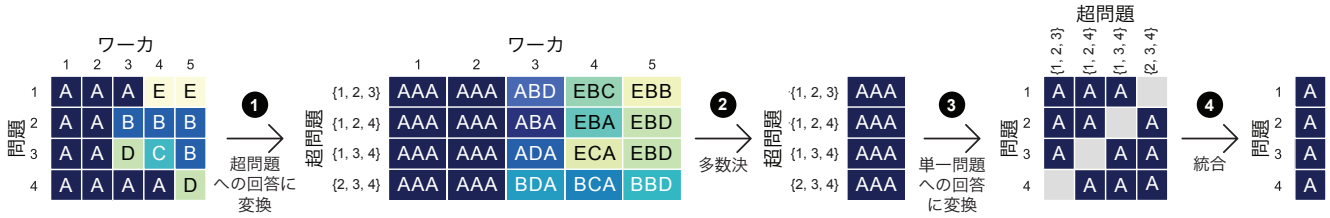


図 3: Hyper-MV の例：(1) k-超問題を構築し，単一問題に対する回答を超問題に対する回答に変換する．ここでは，$k = 3$ の例を示している．(2) 超問題への回答に対して多数決を実施する．(3) 各超問題に対する多数決解から，各単一問題の多数決解への投票を得る．(4) 多数決を実施し，各単一問題に対する最終的な答えを得る．

が，非専門家よりも高い．つまり，一連の問題に対する専門家の正答率と非専門家の正答率の差は，単一問題のそれよりも大きくなり，専門家が多数決で可能性が高くなる．いま，常に正答する完璧な専門家と，常にランダムに回答する非専門家が，それぞれ複数いるとする．‘A’ か ‘B’ で答えるような選択肢が二つの問題において，専門家の正答率は 100% になり，非専門家の正答率は 50% となる．専門家は常に正答するため，ある専門家ペアを選んだときに，その合意解は常に正答と一致する．一方で，非専門家ペアは，25% の確率で誤答で合意してしまう．つまり，例えば正解が ‘A’ である問題に対して両者が ‘B’ と答える確率が 25% となる．非専門家の数が多い場合，このような好ましくない偶然が起こりやすくなり，従って多数決結果が誤答となる可能性が高くなる．

選択肢が ‘A’ と ‘B’ である問題が複数ある場合を考える．全ての問題で正解は ‘A’ だとする．このとき，ある二つの問題を選んだときに，ある非専門家ペアが両方の問題で ‘B’ と誤答する確率は 6.3% となる．超問題を用いて表すと，この 2-超問題に対する回答は ‘AA’，‘AB’，‘BA’，‘BB’ の 4 種類であり，両者が ‘BB’ で合意する確率が 6.3% ということである．さらに 3-超問題を考えると，両者が誤答 ‘BBB’ で合意する確率は 1.6% に減少する．一方で専門家ペアは，2-超問題であっても 3-超問題であっても，誤答 ‘BBB’ で合意する確率は 0% である．このように超問題を導入すると，非専門家同士が誤答で合意する確率を減らすことができ，多数派が誤答となる状況を回避しやすくなる．

4.2 超問題に基づく多数決 (Hyper-MV)

超問題を用いた具体的な回答統合法として，Hyper-MV を提案する．Hyper-MV では，超問題に対する回答で多数決を取ることで，専門家の意見を強調する．

Hyper-MV では，まず問題集合 $\mathcal{Q}$ から $\binom{|\mathcal{Q}|}{k}$ 件の超問題を構築する．ある k-超問題 $\{q_1, q_2, \ldots, q_k\}$ の選択肢は，$\prod_{\kappa=1}^{k} |\mathcal{C}_{q_\kappa}|$ 種類となる．また，単一問題に対する回答を，超問題に対する回答として変換する．あるワーカ w の超問題 $\{q_1, q_2, \ldots, q_k\}$ に対する回答は，$(l_{wq_1}, \ldots, l_{wq_k})$ の連結として表される．Hyper-MV は，各超問題への回答に対して多数決を実施し，超問題に対する多数決解を単一問題に対する多数決解として復元する．ある単一問題が複数の超問題に含まれるため，単一問題に対する多数決解が複数生成される．最後に，それらを多数決によって統合し，各単一問題に対する最終的な答えを得る．

以上の手続きを，図 3 で示した具体例を用いて説明する．この例では，5 人のワーカが 4 件の問題に対して回答している．4 件の問題全てで選択肢は ‘A’, ‘B’, ‘C’, ‘D’, ‘E’ である．Hyper-MV は，まず k-超問題を構築し，単一問題に対する回答を超問題に対する回答として変換する．この例では $k = 3$ としており，$\{1,2,3\}, \{1,2,4\}, \{1,3,4\}, \{2,3,4\}$ の 4 件の超問題を構築する．また，例えば 3 人めのワーカの回答を，超問題 $\{1,2,3\}$ に対する回答 ‘ABD’ として変換する．次のステップでは，超問題の上で多数決を実施する．4 件の超問題全てで，‘AAA’ が多数決解となる．3 番目のステップでは，各超問題の多数決解から，各単一問題の多数決解への投票を得る．例えば，超問題 $\{1,2,3\}$ の多数決解 ‘AAA’ から，問題 1, 2, 3 のそれぞれに対する ‘A’ という投票を得る．最後のステップでは，この投票に対して多数決を実施し，各単一問題に対する最終的な答えを得る．この例では，全ての問題で正答が ‘A’ であり，5 人のワーカのうち 2 名が専門家で常に正しい答えを返している．単純な多数決は問題 2 で失敗するが，Hyper-MV を用いると全ての問題で正解を獲得できる．

超問題の 1 回めの多数決は，ほかの回答統合法で置き換えることができる．ワーカと問題の難易度を考慮した統計的回答統合法である GLAD [18] と DARE [1] で置換えた手法を，それぞれ Hyper-GLAD, Hyper-DARE として提案する．なお，D&S [2]，BCC [7]，CommunityBCC [15] 等の，混合行列を用いてワーカの能力を表現する回答統合法は，全ての問題で選択肢が共通することを前提としているため，本研究の問題設定には適さない．

4.3 超問題のランダムサンプリング

Hyper-GLAD と Hyper-DARE では，超問題の数が多くなるとパラメータ数が増え処理にパラメータ推定に時間が掛かる．計算時間の削減法として，超問題のランダムサンプリングが考えられる．各単一問題が含まれる超問題の数の偏りを減らすため，以下のようなランダムサンプリング法を用いる．n 件の単一問題をランダムに並び替え，先頭から k 件の単一問題を選んで超問題とする．$\lfloor \frac{n}{k} \rfloor$ 件の超問題が生成されたら，再度単一問題をランダムに並び替え同じ処理を実行する．この処理を r 回繰り返し，最終的に $r \cdot \lfloor \frac{n}{k} \rfloor$ 件の超問題が生成される．5.2.3 節では，超問題のサンプリング数が統合結果の正答率に与える影響を検証する．

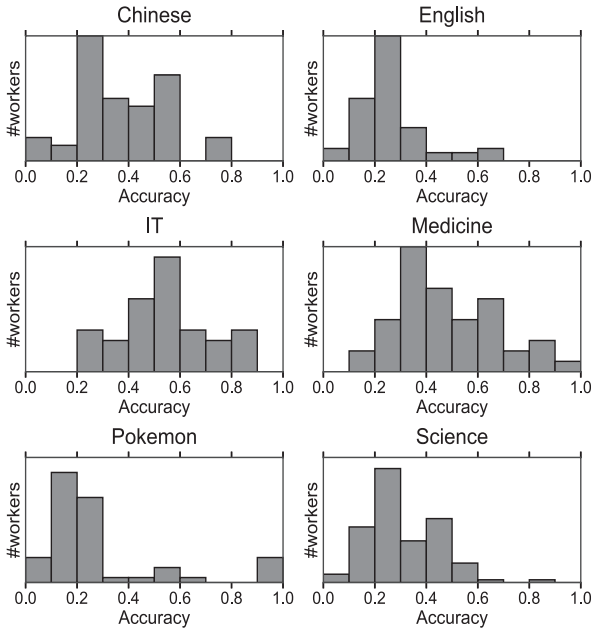


図 4: 実データにおけるワーカの正答率の分布

5. 評価実験

5.1 人工データでの実験

5.1.1 実験設定

どのような場合に Hyper-MV が多数決よりも優れているかを検証するために，人工データを用いた実験を行った．20 人のワーカが 20 件の問題に回答している場合を対象とした．全ての問題で選択肢は 5 件とした．非専門家はランダムに回答するものとした．専門家の人数 n_e が $\{2,4,6\}$，専門家の正答率 p_e が $\{0.8,0.9,0.95,1.0\}$ の場合にそれぞれについて，回答を生成した．各 (n_e,p_e) について，回答データは 100 種類生成した．

超問題のサイズは $k=5$ に，ランダムサンプリングのパラメータは $r=100$ に設定した．

5.1.2 結　　果

表 1 に，各手法の統合結果の正答率の平均と標準偏差を示す．Hyper-MV は，特に専門家の数が少ない場合，多数決よりも優れていることが確認できた．20% 以上のワーカが専門家であれば専門家の正答率が 80% の場合でも，Hyper-MV を用いることでほとんどの場合に正解を獲得できる．

5.2 実データでの実験

5.2.1 データセット

表 2 に示す 6 分野の専門的な問題を用意し，クラウドソーシングで回答を収集してデータセットを作成した (注1)．‘Chinese’ データセットは中国語の語彙に関する問題で，‘English’ は Graduate Record Examinations（GRE）で用いられている，語彙に関する問題である．‘IT’ は基本情報技術者試験で過去に出題された問題，‘Medicine’ は看護師国家試験で用いられた薬効および副作用に関する問題，‘Pokémon’ はポケモンの英語名に関する問題，‘Science’ は化学と物理に関する問題である．いずれのデータセットでも，全ての問題で，選択肢は異なっている．

クラウドソーシングプラットフォーム「ランサーズ」を用いて回答を収集した．ワーカは，あるデータセット内の全ての問題に答えるように指示された．また，問題と選択肢の提示順は，ワーカごとにランダムとした．図 4 に各データセットにおけるワーカの正答率の分布を示す．いずれのデータセットでも平均的なワーカの正答率は低く，難しい問題となっている．

5.2.2 実験設定

提案法 (Hyper-MV, Hyper-GLAD, Hyper-DARE) に対するベースラインとして，多数決，GLAD，DARE を用いた．GLAD のパラメータ事前分布は，$\alpha \sim \mathcal{N}(1,1)$ と $\beta' \sim \mathcal{N}(1,1)$ に設定した．DARE の事前分布は，$a_p \sim \mathcal{N}(0,1)$，$d_q \sim \mathcal{N}(0,1)$ と $\tau_q \sim \mathrm{Gamma}(1,0.01)$ に設定した．Hyper-MV については，超問題のサイズ k を 2 から 7 の範囲で，Hyper-GLAD と Hyper-DARE については k を 2 から 4 の範囲で変えながら実験を行った．提案法では超問題をランダムサンプリングしている．10 種類の異なる超問題集合をサンプリングし，それぞれについて手法を適用した．

5.2.3 結　　果

表 3 は，提案法とベースラインの正答率を示す．すべてのデータセットで，$k=4,5,6,7$ の場合に Hyper-MV の正答率が多数決を上回った．$k=2,3$ の場合には，‘Chinese’ データセットでは多数決の正答率の方が高くなった．これは，少数の単一問題で構成された超問題では専門家の強調に不十分だったためと考えられる．また，Hyper-MV を $k \geq 8$ にした場合に，正答率が低下することを確認した．現実には専門家が常に正しい答えを返すとは限らず，大きいサイズの超問題に対する回答が一致しない場合もあるためである．

ほとんどの場合で，Hyper-GLAD と Hyper-DARE はそれぞれ GLAD と DARE よりも高い正答率を達成した．この結果は，小さい超問題であっても，GLAD と DARE のワーカ能力推定を改善するのに有効であることを示している．また，$k=5,6,7$ の Hyper-MV は GLAD や DARE よりも優れた，あるいは同等の正答率を達成している．単純な手法である Hyper-MV が，ベイズ推論に基づく複雑な手法である GLAD や DARE に匹敵することは注目に値する．

ワーカが全ての問題に回答していない場合，つまり回答が欠損している場合の各手法の性能を調査した．実データの回答の一部をランダムに取り除いて，各手法を適用した．超問題への回答に変換する場合，超問題を構成する少なくとも一つの単一問題に対する回答がない場合には，その超問題に対する回答が存在しないとみなした．図 6 に，回答のランダムサンプリングを 10 回行った場合の正答率の平均値を示す．欠損率が 0.3 以下の場合，Hyper-MV（$k=3,5$），Hyper-GLAD（$k=3$），および Hyper-DARE（$k=3$）は，それぞれ MV，GLAD，DARE よりも高い正答率を示している．一方，欠損率が増えるにつれ，提案法の正答率は減少する．これは，欠損率が増えるにつれ，超問題に対する回答が欠損し，ある超問題に一人し

(注1)：http://www.ml.ist.i.kyoto-u.ac.jp/en/en-research/li2017cikm で公開されている

表 1: 専門家の人数（n_e）と専門家の正答率（p_e）ごとの，人工データにおける多数決 (MV) と Hyper-MV の正答率の平均と標準偏差．ウィルコクソンの符号順位検定による統計的に有意な（$p < 0.05$）勝者には下線が引かれている． Hyper-MV は，特に専門家の数が少ない場合，多数決よりも正答率が高くなる．

p_e	$n_e = 2$		$n_e = 4$		$n_e = 6$	
	MV	Hyper-MV ($k = 5$)	MV	Hyper-MV ($k = 5$)	MV	Hyper-MV ($k = 5$)
0.80	0.398 ±0.117	<u>0.597</u> ±0.182	0.629 ±0.113	<u>0.929</u> ±0.072	0.840 ±0.077	<u>0.974</u> ±0.038
0.90	0.446 ±0.098	<u>0.838</u> ±0.133	0.722 ±0.092	<u>0.989</u> ±0.028	0.916 ±0.068	<u>0.998</u> ±0.010
0.95	0.460 ±0.112	<u>0.932</u> ±0.058	0.759 ±0.086	<u>0.999</u> ±0.007	0.945 ±0.055	<u>1.000</u> ±0.005
1.00	0.475 ±0.116	<u>1.000</u> ±0.000	0.809 ±0.090	<u>1.000</u> ±0.000	0.977 ±0.035	<u>1.000</u> ±0.000

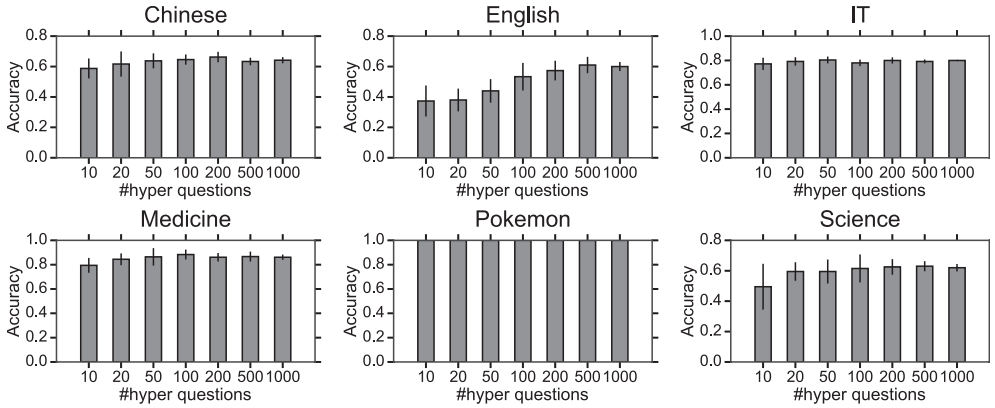


図 5: 超問題の数が Hyper-MV ($k = 5$) の正答率に与える影響．10 種類の超問題集合に手法を適用した場合の，正答率の平均と標準偏差を示している．超問題の数が 200 以上の場合，すべてのデータセットで正答率が安定している．

表 2: 実データの詳細

データセット	問題数	ワーカ数	選択肢数
Chinese	24	50	5
English	30	63	5
IT	25	36	4
Medicine	36	45	4
Pokémon	20	55	6
Science	20	111	5

か回答していない場合が生じるためである．k が大きくなるほど超問題に対する回答が欠損しやすいが，実際に Hyper-MV の $k = 3$ と $k = 5$ の場合を比較すると，$k = 3$ の場合の方が正答率の減少が緩やかである．このことからも，超問題に対する回答の欠損が性能低下の要因であることが確認できる．

最後に，超問題のランダムサンプリングが提案法の性能に与える影響を調査した．図 5 に，超問題のサンプリング数を変化させた場合の Hyper-MV（$k = 5$）の正答率を示す．‘English’ と ‘Science’ 以外では，超問題が少数でも正答率はほとんど低下しなかった．‘English’ と ‘Science’ では，専門家の人数が他のデータセットよりも少なかったため，専門家の意見を強調するために多くの超問題が必要だったためと考えられる．全てのデータセットで，超問題の数が 200 以上であれば正答率は安定している．

6. 議　　論

6.1 回答の欠損

提案法が上手く働かない場合について議論する．まず，図 6 で示したように，提案法はワーカが一部の問題に回答していない場合には正答率が低下する．我々の手法は，まずは少数のテスト用問題を用意しワーカに全ての問題に回答させ，提案法を用いて専門家を発見し，本番の問題は専門家だけに依頼するというような運用シナリオ [8] に適していると考えられる．

6.2 文脈的バイアスと意味的バイアス

提案法は，複数の問題にわたって専門家同士の回答が一致しやすく，一方で非専門家同士の回答は一致しづらいという前提に基いている．しかし，非専門家同士の回答が偶然一致することもあるし，ワーカ同士の結託によって生じるようなバイアスによって，一致しやすくなることも有り得る．回答のバイアスは，文脈的バイアスと意味的バイアスに分類される．文脈的バイアスは，選択肢が提示される文脈や選択肢自体に関係なく生じる．典型的な例は，質問の提示の順序と選択肢のレイアウトによって生じるバイアスである．一方，意味的なバイアスは，選択肢の実際の内容に応じて生じる．バイアスへの簡単な対処法は，各ワーカに提示される質問と選択肢の順序をランダム化することである．このランダム化は文脈的バイアスを減らすこ

表 3: 多数決（MV）と Hyper-MV，GLAD と Hyper-GLAD，DARE と Hyper-DARE の正答率の比較．提案法については，10 回の平均と標準偏差を示す．提案法が競合手法を上回る場合は，下線が引かれている． 提案法は，ほとんどのケースで競合手法よりも高い正答率を示した．

(a) MV vs. Hyper-MV

Dataset	MV	Hyper-MV					
		$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
Chinese	0.625	0.604 ±0.021	0.617 ±0.017	<u>0.642</u> ±0.038	<u>0.637</u> ±0.019	<u>0.671</u> ±0.035	<u>0.646</u> ±0.034
English	0.433	<u>0.460</u> ±0.020	<u>0.513</u> ±0.031	<u>0.523</u> ±0.042	<u>0.597</u> ±0.046	<u>0.563</u> ±0.046	<u>0.497</u> ±0.057
IT	0.720	<u>0.772</u> ±0.031	<u>0.804</u> ±0.022	<u>0.812</u> ±0.018	<u>0.788</u> ±0.018	<u>0.796</u> ±0.022	<u>0.800</u> ±0.018
Medicine	0.667	<u>0.747</u> ±0.019	<u>0.803</u> ±0.015	<u>0.850</u> ±0.022	<u>0.864</u> ±0.026	<u>0.900</u> ±0.022	<u>0.894</u> ±0.017
Pokémon	0.650	<u>0.975</u> ±0.034	<u>1.000</u> ±0.000	<u>1.000</u> ±0.000	<u>1.000</u> ±0.000	<u>1.000</u> ±0.000	<u>1.000</u> ±0.000
Science	0.550	<u>0.555</u> ±0.015	<u>0.580</u> ±0.033	<u>0.620</u> ±0.024	<u>0.620</u> ±0.033	<u>0.605</u> ±0.035	<u>0.650</u> ±0.071

(b) GLAD vs. Hyper-GLAD

Dataset	GLAD	Hyper-GLAD		
		$k=2$	$k=3$	$k=4$
Chinese	0.542	<u>0.696</u> ±0.053	<u>0.775</u> ±0.028	<u>0.750</u> ±0.032
English	0.567	<u>0.663</u> ±0.023	<u>0.713</u> ±0.022	<u>0.730</u> ±0.028
IT	0.720	<u>0.840</u> ±0.000	<u>0.820</u> ±0.020	<u>0.824</u> ±0.020
Medicine	0.694	<u>0.889</u> ±0.000	<u>0.917</u> ±0.012	<u>0.931</u> ±0.014
Pokémon	0.850	<u>1.000</u> ±0.000	<u>1.000</u> ±0.000	<u>1.000</u> ±0.000
Science	0.550	<u>0.645</u> ±0.061	<u>0.645</u> ±0.042	<u>0.690</u> ±0.037

(c) DARE vs. Hyper-DARE

Dataset	DARE	Hyper-DARE		
		$k=2$	$k=3$	$k=4$
Chinese	0.625	<u>0.658</u> ±0.025	<u>0.667</u> ±0.046	<u>0.708</u> ±0.019
English	0.600	<u>0.677</u> ±0.026	<u>0.680</u> ±0.031	<u>0.690</u> ±0.021
IT	0.800	<u>0.816</u> ±0.020	<u>0.804</u> ±0.012	<u>0.828</u> ±0.018
Medicine	0.861	0.833 ±0.000	<u>0.939</u> ±0.021	<u>0.956</u> ±0.014
Pokémon	1.000	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000
Science	0.600	0.600 ±0.000	<u>0.605</u> ±0.015	<u>0.640</u> ±0.054

とができるが，意味的バイアスには効果的ではない．

6.3 ワーカの種類ごとに生じうる回答バイアス

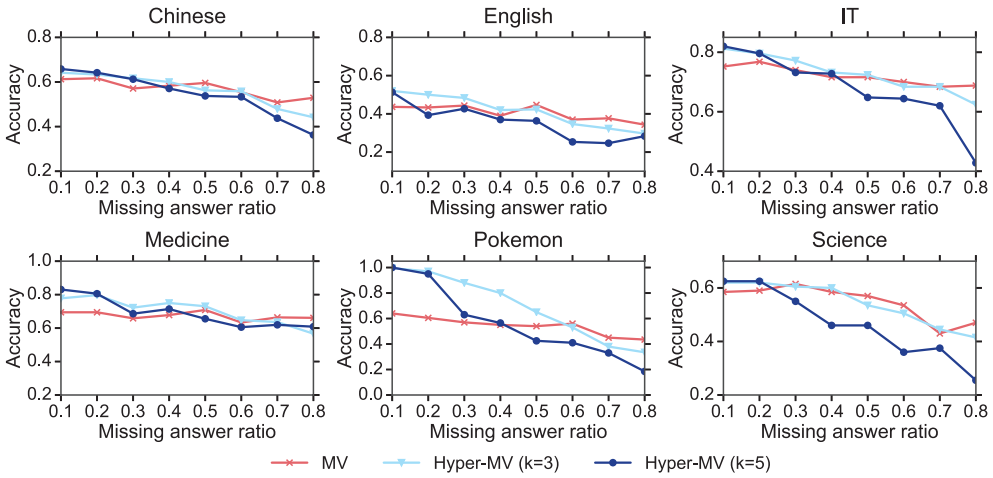
回答統合に悪影響を与えるワーカを 4 種類に分類し，それぞれで生じうる回答バイアスについて議論する．

- 能力は低いが勤勉なワーカ：この種類のワーカは，難しい問題でも精一杯回答しようとし，結果的には文脈的バイアスに基づく回答となる．一方，問題自体が特定の誤答を引き起こしやすいような問題の場合には，意味的バイアスに基づく回答となる．超問題を構成する全ての問題が意味的バイアスを生じさせるものでない場合には，提案法により専門家の意見を強調することで，意味的バイアスの影響を減らすことができる．
- スパムワーカ：常にランダムな選択肢を選んだり，同じ位置の選択肢を選ぶような，選択肢の内容と無関係に回答するワーカである．このワーカは選択肢の内容を見ないため，文脈的バイアスだけが問題となる．
- 悪意のある非専門家：正解は知らないが，わざと間違うように回答するワーカである．同種のワーカ同士で結託しない限り，この種類のワーカの行動は（正解を知らないため）スパムワーカと同じとなり，文脈的バイアスだけが問題となる．
- 悪意のある専門家：正解を知っていて，わざと間違うように回答するワーカである．選択肢の数が二つだけの場合，悪意のある専門家同士は，同じ誤答を選び回答が一致してしまい，回答統合に悪影響を与える．選択肢数を増やすほど，影響を減らすことができる．

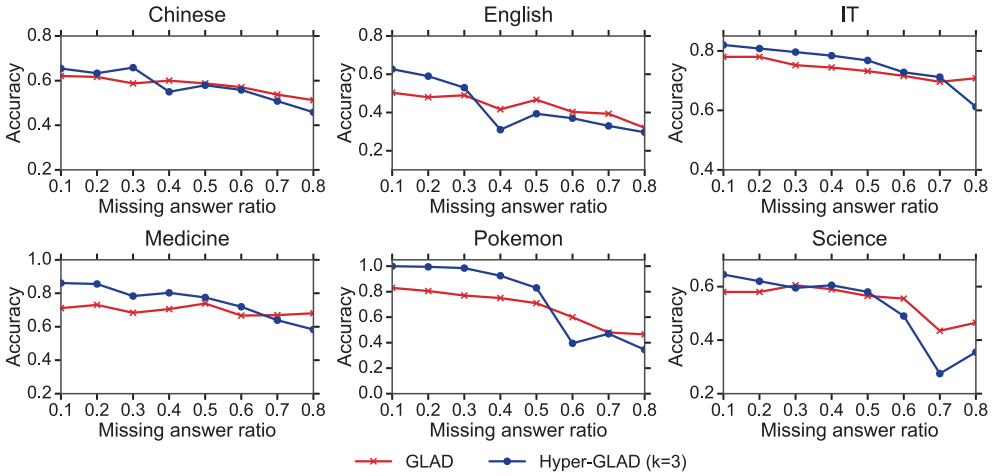
悪意のある専門家・非専門家同士が回答を共有し結託すると，統合結果に意味的バイアスを含めるための攻撃を実行することができる．しかし，ほとんどのクラウドソーシングタスクでは，このような結託攻撃を企むワーカは一般的ではない．まとめると，質問に三つ以上の選択肢があり，悪意あるワーカ同士が結託しない場合には，文脈的バイアスだけが問題になり，質問・選択肢の順序のランダム化が効果的な対処法となる．

7. ま と め

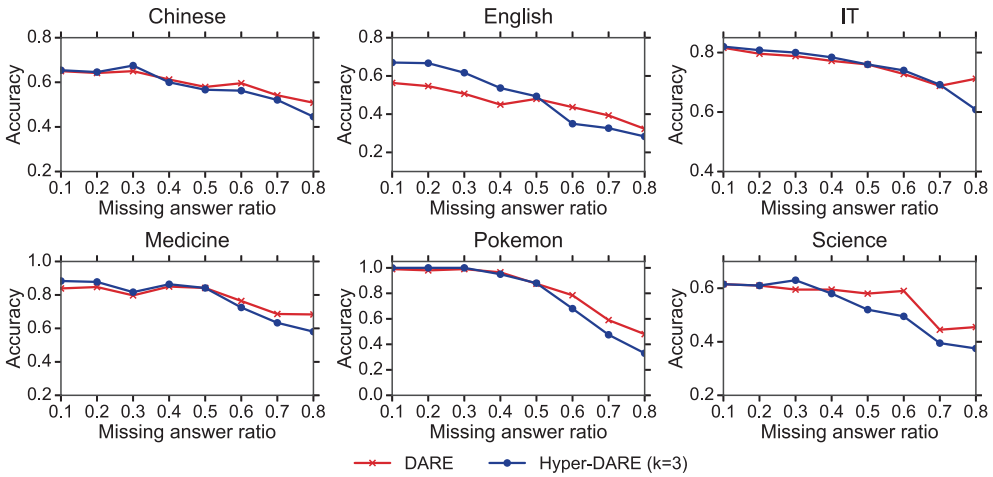
本研究では，多数派が誤答するような難しい問題に対する回



(a) MV vs. Hyper-MV



(b) GLAD vs. Hyper-GLAD (k=3)



(c) DARE vs. Hyper-DARE (k=3)

図 6: 回答の欠損率に応じた正答率の変化．10 回の平均正答率を示している．欠損率が 0.3 以下の場合には提案法はベースラインよりも正答率が高いが，欠損率が高くなるにつれ正答率が低下する．

答統合法を提案した．複数の問題において専門家同士の回答は一致しやすく非専門家同士の回答は一致しづらいという観察に基づき，複数の問題をまとめた超問題を導入し，超問題に基づく回答統合法を提案した．人工データと実データを用いた実験により，提案法は，特に少数の専門家しか存在しない場合に有効であることを確認した．提案法には，いくつかの欠点がある．第一に，回答の欠損率が高いと，統合結果の正答率を大きく損なう場合がある低下する．第二に，正答率は超問題のサイズ（k）に敏感であり，データに適した k の選択が必要となる．これらの欠点への対処が今後の課題である．

文　　献

[1] Y. Bachrach, T. Minka, J. Guiver, and T. Graepel, "How to grade a test without knowing the answers: A Bayesian graphical model for adaptive crowdsourcing and aptitude testing", In *Proceedings of the 29th International Conference on Machine Learning (ICML '12)*, 2012, pp. 819–826.

[2] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm", *Journal of the Royal Statistical Society. Series C (Applied Statistics)* Vol. 28, No. 1., 1979, pp. 20–28.

[3] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor, "Are your participants gaming the system?: Screening Mechanical Turk workers", In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*, 2010, pp. 2399–2402.

[4] P. G. Ipeirotis and E. Gabrilovich, "Quizz: Targeted crowdsourcing with a billion (potential) users", In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*, 2014, pp. 143–154.

[5] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on Amazon Mechanical Turk", In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP'10)*, 2010, pp. 64–67.

[6] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems", In *Advances in Neural Information Processing Systems 24 (NIPS'11)*, 2011, pp. 1953–1961.

[7] H. C. Kim and Z. Ghahramani, "Bayesian classifier combination", In *Artificial Intelligence and Statistics (AISTATS'12)*, 2012, pp. 619–627.

[8] H. W. Li, B. Zhao, and A. Fuxman, "The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing", In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*, 2014, pp. 165–176.

[9] J. Li, Y. Baba, and H. Kashima, "Hyper Questions: Unsupervised targeting of a few experts in crowdsourcing", In *Proceedings of the 26th ACM International Conference on Information and Knowledge Management (CIKM'17)*, pp. 1069–1078, 2017.

[10] Q. Liu, J. Peng, and A. Ihler, "Variational inference for crowdsourcing", In *Advances in Neural Information Processing Systems 25 (NIPS'12)*, 2012, pp. 692–700.

[11] F. L. Ma, Y. L. Li, Q. Li, M. H. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. W. Han, "Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation", In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*, 2015, pp. 745–754.

[12] J. M. Mortensen, M. A. Musen, and N. F. Noy, "Crowdsourcing the verification of relationships in biomedical ontologies", In *AMIA Annual Symposium Proceedings (AMIA'13)*, 2013, pp. 1020.

[13] V. C. Raykar, S. P. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds", *Journal of Machine Learning Research* Vol. 11, 2010, pp. 1297–1322.

[14] G. Tsoumakas and I. Vlahavas, "Random k-labelsets: An ensemble method for multilabel classification", In *Proceedings of the 18th European Conference on Machine Learning (ECML'07)*, 2007, pp. 406–417.

[15] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi, "Community-based Bayesian aggregation models for crowdsourcing", In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*, 2014, pp. 155–164.

[16] F. L. Wauthier and M. I. Jordan, "Bayesian bias mitigation for crowdsourcing", In *Advances in Neural Information Processing Systems 24 (NIPS'11)*, 2011, pp. 1800–1808.

[17] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The multidimensional wisdom of crowds", In *Advances in Neural Information Processing Systems 23 (NIPS'10)*, 2010, pp. 2424–2432.

[18] J. Whitehill, P. Ruvolo, T. F. Wu, J. Bergsma, and J. Movellan, Whose vote should count more: Optimal integration of labels from labelers of unknown expertise, In *Advances in Neural Information Processing Systems 22 (NIPS'09)*, 2009, pp. 2035–2043.

[19] D. Y. Zhou, J. C. Platt, S. Basu, and Y. Mao, "Learning from the wisdom of crowds by minimax entropy", In *Advances in Neural Information Processing Systems 25 (NIPS'12)*, 2012, pp. 2195–2203.