原著論文

Original Paper

クラウドソーシングワーカーの段階的育成方法の提案

Crowdsourcing Worker Development based on Probabilistic Task Network

芦川 将之
Masayuki Ashikawa
電気通信大学大学院情報システム学研究科 / (株) 東芝 研究開発センター
Graduate School of Information Systems, The University of Electro-Communications / Corporate Research and Development Center, Toshiba Corporation
ashikawa.masayuki@ohsuga.is.uec.ac.jp

川村 隆浩
Takahiro Kawamura
電気通信大学大学院情報システム学研究科
Graduate School of Information Systems, The University of Electro-Communications
kawamura@ohsuga.is.uec.ac.jp

大須賀 昭彦
Akihiko Ohsuga
（同　　上）
ohsuga@uec.ac.jp

keywords: crowdsourcing, education, bayesian network

Summary

Current crowdsourcing platforms such as Amazon Mechanical Turk provide an attractive solution Crowdsourcing platforms provide an attractive solution for processing numerous tasks at a low cost. However, insufficient quality control remains a major concern. Therefore, we developed a private crowdsourcing system that allows us to devise quality control methods. In the present study, we propose a grade-based training method for workers in order to avoid simple exclusion of low-quality workers and shrinkage of the crowdsourcing market in the near future. Our training method utilizes probabilistic networks to estimate correlations between tasks based on workers' records for 18.5 million tasks and then allocates pre-learning tasks to the workers to raise the accuracy of target tasks according to the task correlations. In an experiment, the method automatically allocated 31 pre-learning task categories for 9 target task categories, and after the training of the pre-learning tasks, we confirmed that the accuracy of the target tasks was raised by 7.8 points on average. This result was comparatively higher than those of pre-learning tasks allocated using other methods, such as decision trees. We thus confirmed that the task correlations can be estimated using a large amount of worker records, and that these are useful for the grade-based training of low-quality workers.

1. は　じ　め　に

不特定多数の人の力を用いて様々な業務処理やサービスを行うクラウドソーシング技術は，大規模データの解析や構築など様々な分野や用途で利用されている．その利用範囲の拡大に従い実際に作業 (タスク) を処理する作業者 (ワーカー) の数も増大しており，将来的にクラウドソーシングにおける作業が社会における一つの就労形態となることが予想される．しかし，そのような傾向にあるにもかかわらず，現状のクラウドソーシングではワーカーに対する育成や労働環境の改善といったサポートが十分であるとはいい難い．これはワーカーが不特定多数であり，補充や変更が容易であることが原因であると予想されるが，このようなワーカーの安易な変更は，ワーカーの経験不足による全体の精度低下やワーカーの不当解雇という問題につながりかねない．

そのため今後のクラウドソーシング運用では通常の労働環境と同様に人材 (ワーカー) の育成が重要になってくると予想される．しかし，クラウドソーシングにおける人材育成には様々な問題がある．特にマイクロタスク型クラウドソーシングではワーカーの数の多さ，ワーカーの匿名性からワーカー個人への対応が難しい．また,「高速」,「低コスト」が利点であるため，コストや時間をかけて人材を育成するのはその利点を失わせてしまう可能性などの問題がある．

我々はこのようなマイクロタスク型クラウドソーシングにおける人材育成の問題に対し，ワーカーがタスクを処理する過程で適切な学習タスクをこなすことで能力を向上させる段階的な学習方式を提案する．このような段階的な学習法式としては，学習支援システム (Intelligent Tutoring System, ITS) における学習モデルをベイジアンネットワークによって表現する研究 [Ueno 00] が提案されており，その有効性が示されている．我々はこのベイジアンネットワークを用いた段階的学習手法のマイクロタスク型クラウドソーシングへの適用を提案する．具体的な手法として，まずワーカーのタスク処理結果からベ

イジアンネットワークを用いてタスク間の関係性の解析を行う．次にタスクを処理することで段階的な学習が可能となるような学習タスクを自動生成する．これによってワーカーの能力の育成を狙う．

このように低品質なワーカーを高品質なワーカーへと育成することで，安易なワーカーの排除を行うことなく，精度向上と同時にワーカーの労働環境を向上させることが我々の狙いである．

このような段階的な学習法を実現するにあたって，外部のマイクロタスク型のクラウドソーシングサービスではワーカーやタスクのコントロールなどサーバ側におけるコントロールが制限される場合が多く，サーバ側における新規機能追加が難しいという問題がある．

そのため，我々はシステム側を自由に変更することが可能なプライベートな環境下におけるクラウドソーシングシステム (PCSS) を開発した [Ashikawa 14]．PCSS は 2011 年 11 月から継続的に運用を続けており，2454 人のワーカーにより 1853 万件のタスクの処理実績がある．

本稿では，学習やクラウドソーシングにおいて機械学習的なアプローチを行った既存の研究に関して紹介し (2 章)，そして，クラウドソーシングにおける段階的学習手法を提案し (3 章)，段階的学習手法に関する PCSS 上の実験とその効果に関する考察を行う (4 章)．最後にまとめと今後の課題に関して述べる (5 章)．

2. 関 連 研 究

クラウドソーシングに限らず，学習や教育に機械学習的な手法を用いた研究として，1-1) 様々な学習の要素が生徒にどのように影響するかを推測する研究，1-2) 生徒の状態から要因を推定する研究，1-3) 生徒を分類して最適な学習プランを検討する研究などがある．

1-1) に関連する研究として，生徒に対して実施したテストや手法がどのような効果があるかを推測する研究 [Xenos 04]，生徒の学習スタイルが最終的にどのように成果に影響しているかを推測する研究 [Garcia 07]，複数の教育手法が生徒にどのような影響があるかを推測し，図示する研究 [Fernandez 11] がある．

1-2) に関連する研究として，生徒の学習状況から社会的経済指標を計算する研究 [May 06]，生徒の家庭環境や収入から生徒の生活背景がどのようなものかを推測する研究 [Hoogerheide 12] がある．

1-3) に関連する研究として生徒をスキル別にグループ分けする研究 [Almond 09, Pardos 10] がある．

我々の研究はワーカーを生徒とみなした場合，どのような要因が生徒の能力向上に影響するかを推測する研究であるため 1-2) のグループに属している．

また，クラウドソーシングに機械学習的な手法を用いた研究として，2-1) ワーカーを処理結果を解析することで分類する研究，2-2) 一つの作業を複数のワーカーに処理させる過程で結果のマージを行う研究，2-3) 一つの作業を複数のワーカーに処理させて得られた結果をグループ分けする研究，2-4) 得られた結果から出題タスクの難易度や品質を推測する研究などがある．

2-1) に関連する研究として，ワーカーを精度に応じてグループ分けする研究 [Nushi 15, Shaw 11, Venanzi 15, Wauthier 11]，作業結果の精度に応じてワーカーの精度を判定し，排除するべきワーカーを判定する研究 [Wais 11]，作業結果の精度に応じてワーカーのスコアリングやランキング付けを行う研究 [Burnap 13, Raykar 14, Shaw 11]，作業結果の精度に応じてワーカーの最適な報酬を推測するための研究 [Xie 15] がある．

2-2) に関する研究として，一つのタスクに対して複数のワーカーから得られた結果からマージされた最適な答えを取得することを目的とした研究 [Carpenter 11, Kamar 12, Sun 12, Tang 11]，得られた文章やツイートにおける一致率を計算し，それに応じて結果をマージする研究 [Simpson 15], SNS やテキストなどへのラベリングデータをマージする研究 [Simpson 15] がある．

2-3) に関する研究として，一つのタスクに対して複数のワーカーから得られた結果を複数のグループに分類する研究 [Bragg 14, Hutton 12, Tang 11] がある．

2-4) に関する研究として，回答したワーカーのスキル，正解率などからタスクの難易度をモデル化する研究 [Bachrach 12]，ワーカーの正解率とワーカーのエラーレートからタスクの難易度をモデル化する研究 [Lin 12] などがある．

このようにワーカーの分類や結果の解析でクラウドソーシングの精度を向上させる研究は行われているが，我々のようにワーカーの行動履歴をベイジアンネットワークなどの機械学習的なアプローチで解析することでワーカーの品質を向上させる研究は行われていない．これは低品質なワーカーは排除することが一般的であることが原因であると考えられる．しかし，前述のように将来的にクラウドソーシングが就労形態として一般的になることを考えた場合，安易な排除は問題になることが予想される．そのため，ワーカーの学習に基づく精度改善による労働環境改善は重要である．

3. ワーカーの精度向上のための学習方法

3·1 段階的学習法の提案

学習とは一般的に一定場面でのある経験が，その後同一，または類似の場面での行動に良い変容をもたらすこと，ということができる．これを様々な分野の作業に当てはめた場合，特定の作業で良い結果を得るためには同一，または類似の作業での経験を積まなければならないということができる．しかし，ある作業において作業者が作業内容を学習するにあたっては，最初から対象となる作業と同等の難易度の作業で学習を行うのではなく，

目的の作業に関連する難易度の低い作業から開始して訓練し，段階的に難易度を上げていくことで作業者の能力を向上させていく段階的な学習方法をとることが一般的である．この手法は学校教育の仕組みと同じであり有効であることは示されている．しかし，学校教育は先生という熟達した管理者によって，様々な学習内容の目的や内容に応じた様々な科目への振り分け，今までの教育経験に裏付けされた学習カリキュラムの設計などがなされ，膨大な学生が実際に定められた学習方法を行い，その結果をフィードバックすることで高い効果を保証しているという点がある．このように段階的学習方法は非常に効果的であるが，段階的学習方法をクラウドソーシングのタスク処理に適用した場合，先生役の不在が問題となる．マイクロタスク型のクラウドソーシングは対象となるタスクが多岐にわたっており，ワーカーも不特定多数であるため，これらに対して学校教育のように明確な科目分け，学習カリキュラムの構築を手動で行う先生役を負担することはリクエスタにとってもシステム管理者にとってもコストの面で現実的ではない．そのため現状多くの場合は目的のタスクを説明するための単純な練習画面を手動で作成するにとどまっている．

我々はこの問題を解決するために従来のワーカーの行動履歴を解析して，学習内容の科目への振り分けと同様に，タスクの目的や内容に応じた様々なカテゴリへの振り分け，学習カリキュラムの設計と同様に，ワーカーへの最適な学習タスクカテゴリの割り当てを自動で行う手法を提案する．また，学習タスクカテゴリを手動で作成するのはコストが高いため既存のタスクカテゴリを再利用して学習タスクカテゴリとして使用する方法を提案する．

前述の学習の定義に従えば，タスクの処理結果の精度を向上させるためには同一，または類似のタスクでの経験を積まなければならない．つまり，タスク A を実施してからタスク B を実施した場合と，タスク A を実施せずにタスク B を実施した場合で，多くのワーカーが前者のケースでタスク B の処理結果が向上していた場合，タスク A はタスク B の練習用タスクとして扱うことができる．この考えに従って，今までのワーカーの行動履歴を解析し，タスク B に対する最適な練習用タスク A を自動で見つけ出すことが我々の提案である．この方法は大量のワーカーの行動履歴とタスクが必要となるため難易度が高かったが，PCSS では多くの運用実績を持っており，これらの大規模データを利用することで実現が可能となった．

人間の教師を必要とせずに学習者へと最適なカスタマイズを行い，かつ学習者から効果的なフィードバックを得ることを目的として ITS が提案されている [Ueno 00]. ITS は段階的学習手法を用いるにあたって非常に効果的な手法であるとみなされており，プログラミング学習など様々な分野で活用されている [Butz 06]．従来，ITS では述語論理表現に基づく知識表現が一般的であったが，

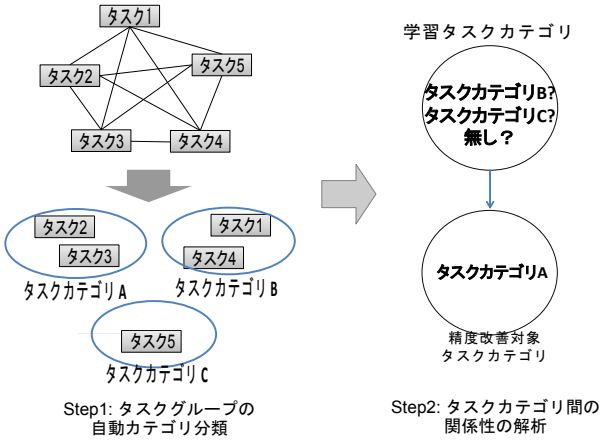


図 1 学習タスクカテゴリ導出のためのステップ

述語論理表現では，ルールの例外に対する処置が難しく，学習者の矛盾した反応に対する柔軟な対応が難しいなどの問題があった．例えば，タスク「単語かどうかの判定」に正解できることがタスク「漢字の読みの入力」に正解できるという必要条件であるというルールが存在した場合，実際にはタスク「単語かどうかの判定」を間違えるが，タスク「漢字の読みの入力」に正解するというケースがケアレスミスや当て推量によってありうる．これはワーカーが不特定多数であり品質が一定していないマイクロタスク型のクラウドソーシングにおいては顕著である．しかし，述語表現によるルール表現ではこのようなケースを処理することは非常に難しい．そこで，ベイジアンネットワークによる確率的アプローチによって，このようなケアレスミスや当て推量を確率に組み込み，矛盾したデータについて合理的な推論を行う．ベイジアンネットワークは因果的な特徴を有向グラフとして表す，現象の因果性，連関性を計算的に推論する理論・技術であり，PCSS では 3·2 節で得られるタスクカテゴリ間の関係性を解析するために用いる．

ITS では様々な学習要素をベイジアンネットワークのノードとして定義し，学習要素の関連性を解析している．ここで述べる学習要素とは「四則演算」や「一次方程式」などの項目であり，手動で定義されている．しかし，PCSS で扱うタスクは多種多様であり，同様にベイジアンネットワークのノードとしてタスクを用いた場合は計算量の面で現実的ではない．そのため，我々は大量のタスクグループを内容ごとに自動でカテゴリに分類し，その後得られたタスクカテゴリ間の関連性の解析を行うことで，学習タスクカテゴリを導出する．このステップを図 1 に示す．

3·2 STEP1: タスクグループの自動カテゴリ分類

PCSS では同内容のタスクを出題する際にはリクエスタが同内容のタスクをまとめてタスクグループとし，タ

スクグループのタイトル，説明文を付与して出題する．タイトル，説明文はシステム管理者がチェックを行い，不適切と思われる内容はリクエスタに再考を依頼している．異なる内容のタスクは別タスクグループとして出題される．本研究ではタスクグループをカテゴリ化するために，タスクグループのタイトルと説明文を形態素解析し，得られた単語を元に各タスクグループの TFIDF 値を計算する．その後に得られた TFIDF 値を用いて各タスクグループ間の類似度を計算して類似度の高いタスクグループ同士をカテゴリ分類する．タスクグループ t における単語 i の出現回数を $W_{t,i}$，タスクグループ t におけるすべての単語の出現回数の和を $W_{t,all}$，すべてのタスクグループ数を T_{all}， 単語 i の出現するタスクグループ数を T_i とした場合，タスクグループ t における単語 i の TFIDF 値 $TFIDF_{t,i}$ は式 (1) のように計算することができる．

$$TFIDF_{t,i} = \frac{W_{t,i}}{W_{t,all}} \log \frac{T_{all}}{T_i} \tag{1}$$

得られた各タスクグループにおける各語彙の TFIDF 値を用いて，各タスクグループ間における類似度の計算を行った．類似度の計算にはコサイン類似度を用いており，タスクグループ t における単語 i の TFIDF 値を $TFIDF_{t,i}$，全単語の集合を W とした場合，タスクグループ $t1$ とタスクグループ $t2$ 間のコサイン類似度 $cos(t1,t2)$ は式 (2) のように計算することができる．

$$cos(t1,t2) = \sum_{i \in W} TFIDF_{t1,i} \cdot TFIDF_{t2,i} \tag{2}$$

その後，得られた類似度を用いてタスクのカテゴリ分類を行う．カテゴリ分類のアルゴリズムは 1) 分類対象となるタスクグループを各カテゴリ所属のタスクグループ全てと比較し，最も類似しているタスクグループが所属するカテゴリに分類，2) 閾値を定め，どのカテゴリのどのタスクグループとも類似度が閾値以下なら新カテゴリを割り当てる，の繰り返しである．

現状の PCSS における 1853 万個のタスク，4153 個のタスクグループに本手法を適用した．全タスクグループ間の類似度を得るために，4153 タスクグループ間の対の全組み合わせ 1724 万通りに対して類似度の計算を行い，得られた類似度を用いてタスクのカテゴリ分類を行うことで 138 個のタスクカテゴリに分類することができた．この適用ではコサイン類似度 0.4 を閾値としている．これは，複数の閾値を用いて本手法を適用し，閾値ごとの結果に対して 100 件のタスクグループの所属するカテゴリを手動で確認したところ，0.4 以上の場合は別カテゴリに所属すると思われるタスクグループは存在しなかったためこの値を用いた．

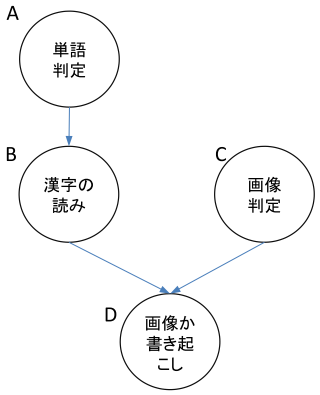


図 2 ベイジアンネットワークをクラウドソーシングに用いた例

3·3 STEP2:タスクカテゴリ間の関係性の解析

クラウドソーシングにおけるタスク処理結果にベイジアンネットワークを用いるにあたって，タスク A における処理結果が高精度である確率を $P(A)$，タスク B における処理結果が高精度であった場合にタスク A における処理結果が高精度である条件付確率を $P(A \mid B)$ のように表すと，$P(A \mid B)$ が高確率ということは「タスク B の処理精度が高かったときにタスク A の処理精度が高い確率」が高確率で発生するということであるため，タスク B を学習タスクとして扱うことができると仮定している．$P(A \mid B)$ は式 (3) のように計算することができる．

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{3}$$

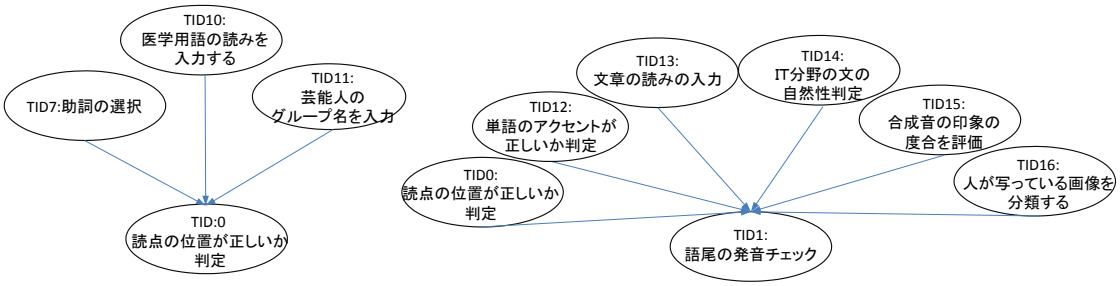
このようにクラウドソーシングワーカーの行動履歴をベイズの定理を用いて解析することにより，精度向上させたいタスクのための学習タスクを自動生成する．ベイジアンネットワークの操作は大別して 1) ベイジアンネットワークの学習，2) ベイジアンネットワークを用いた推論の 2 つで行われる．

ベイジアンネットワークをクラウドソーシングに適用した具体的な例をあげる．クラウドソーシングにおけるタスクとワーカーの行動履歴からベイジアンネットワークの学習を行い，図 2 のような有向グラフが得られたと仮定する．タスク A はタスク B に影響し，タスク B，C はタスク D に影響することがわかる．タスク B，C を処理した後にタスク D を処理したタスク処理結果精度と，タスク B，C を処理せずにタスク D を処理したタスク処理結果精度を比較した場合，前者の方がタスク処理結果の精度が高かった場合，タスク B，C をタスク D の学習タスクとして取り扱うことで結果精度の向上を狙う．

3·2 節で得られた 138 個のタスクカテゴリに対して本手法を適用して有向グラフを作成するにあたり，各タスクカテゴリにおけるワーカーの平均精度を用いた．ワーカーの行動履歴から各タスクカテゴリの平均正解率が 90%以上である確率を計算し，任意のタスクカテゴリ X の平均正解率が 90%以上であった場合にタスクカテゴリ A にお

表 1 精度改善タスクカテゴリ一覧

TID(Task category ID)	タスクカテゴリ名	平均精度 (%)
0	読点の位置が正しいか判定	73.8
1	語尾の発音チェック	82.6
2	対話パターン作成	83.4
3	有名人，芸能人の読み仮名を入力する	85.5
4	キーワードを分類	85.8
5	漢字の読み方の正誤判定	86.2
6	人名の音程の高低を入力する	87.5
7	助詞の選択	87.8
8	英単語の読みを入力する	88.7
9	単語の品詞を選択する	88.9



TID7	TID10	TID11	P(TID0 | TID7,TID10,TID11) >= 90	P(TID0 | TID7,TID10,TID11)< 90
T	T	T	0.75	0.25

TID0	TID12	TID13	TID14	TID15	TID16	P(TID1 | TID0,TID12, TID13,TID14, TID15,TID16) >= 90	P(TID1 | TID0,TID12, TID13,TID14, TID15,TID16) <90
T	T	T	T	T	T	0.528	0.472

図 3 精度改善対象タスクカテゴリにおける有向グラフの例

けるタスクの平均正解率が90%以上である確率 $P(A \mid X)$ を式 (4) のように計算している．

$$P(A \mid X) = \frac{P(X \mid A)P(A)}{P(X)} \tag{4}$$

つまり，式 (4) を用いて得られた有向グラフではタスクカテゴリ X はタスクカテゴリ A の学習タスクカテゴリとして扱うことができることが示される．

3·2 節で得られた 138 個のタスクカテゴリのうち，精度向上させたいタスクカテゴリとして平均精度が低い順に表 1 に示す．結果として，表 1 における精度改善対象となるタスクカテゴリ全てに対して有向グラフを得ることができた．その有向グラフの一例を図 3 に示す．この図は「Task category ID(TID)0: 読点の位置が正しいか判定」と「TID1: 語尾の発音チェック」，それぞれのタスクカテゴリにおける有向グラフである．

図 3 の左の有向グラフは精度改善対象となるタスクカテゴリ A を「TID0: 読点の位置が正しいか判定」として (4) の式を適用した場合，タスクカテゴリ TID0 に関連性があるタスクカテゴリ X はそれぞれ「TID7: 助詞の選択」,「TID10: 医学用語の読みを入力する」,「TID11: 芸能人のグループ名を入力」であることを示している．

ベイジアンネットワークを用いた場合は得られる有向グラフは複数層であるが，本論文では精度改善対象となるタスクカテゴリに直接影響を及ぼしているタスクカテゴリとの関係性に関してのみ述べる．ここで得られた直接影響を及ぼしているタスクカテゴリを学習タスクカテゴリとして扱う．

有向グラフを作成するにあたって，学習タスクが生成できる有向グラフが得られること，学習タスクが多くなりすぎると学習のためのコストが大きくなるため学習タスクが多くなりすぎないことの 2 点を前提にパラメタの設定を行った．アルゴリズムは simulated annealing を用い，Markov Blanket correction は用いず，10 分割交差検証を行った．simulated annealing のパラメタは，Tstart は 10.0，delta は 0.999，Markov Blanket Classifier は false, runs は 10000，scopeType は BAYES，Seed は 1 とした．また，このパラメタの調整にあたっては「TID 0:読点の位置が正しいか判定」をサンプルとして用いている．パラメタの調整においては Markov Branket Classifier の変更では有向グラフに変化はなく，TStart は小さすぎると学習タスクの作成に失敗する傾向があり，大きすぎると大量に学習タスクが発生する傾向があった．また，runs や seed は小さくしすぎると失敗する傾向があった．最終的に前述のパラメタで表 2 に示す学習タスクが得られ，表 3 で示す学習効果を得ることができたため，その他の

表 2 精度改善タスクカテゴリと対応する学習タスクカテゴリ

タスクカテゴリ名	ベイジアンネットワークから得られた 学習タスクカテゴリ	決定木から得られた 学習タスクカテゴリ
TID 0:読点の位置が正しいか判定	TID 7:助詞の選択 TID 10:医学用語の読みを入力する TID 11:芸能人のグループ名を入力	TID 24:人名の「苗字」と「名前」を区切る
TID 1:語尾の発音チェック	TID 0:読点の位置が正しいか判定 TID 12:単語のアクセントが正しいか判定 TID 13:文章の読みの入力 TID 14:ＩＴ分野の文の自然性判定 TID 15:合成音の印象の度合を評価 TID 16:人が写っている画像を分類する	TID 11:芸能人のグループ名を入力
TID 2:対話パターン作成	なし	TID 9:単語の品詞を選択する TID 25:読み仮名と発音を組み合わせる
TID 3:有名人，芸能人の読み仮名を入力する	TID 4:キーワードを分類 TID 11:芸能人のグループ名を入力 TID 14 ＩＴ分野の文の自然性判定	決定木作成に失敗
TID 4:キーワードを分類	TID 17:熟語のアクセントが正しいか判定	TID 5:漢字の読み方の正誤判定 TID 26:単語の読みを入力
TID 5:漢字の読み方の正誤判定	TID 10:医学用語の読みを入力する TID 15:合成音の印象の度合を評価 TID 18:正しい文節の切れ目を選択 TID 19:Wikipedia の単語の読みを入力する	TID 2:対話パターン作成 TID 21:言葉の共通語アクセントを選ぶ TID 27:単語の読み方の正誤判定
TID 6:人名の音程の高低を入力する	TID 3:有名人，芸能人の読み仮名を入力する TID 7:助詞の選択 TID 11:芸能人のグループ名を入力 TID 14:ＩＴ分野の文の自然性判定 TID 18:正しい文節の切れ目を選択 TID 20:文の言い換え文作成	TID 21:言葉の共通語アクセントを選ぶ
TID 7:助詞の選択	TID 17:熟語のアクセントが正しいか判定 TID 21:言葉の共通語アクセントを選ぶ	TID 26:単語の読みを入力
TID 8:英単語の読みを入力する	TID 7:助詞の選択 TID 10:医学用語の読みを入力する	TID 26:単語の読みを入力
TID 9:単語の品詞を選択する	TID 2:対話パターン作成 TID 10:医学用語の読みを入力する TID 17:熟語のアクセントが正しいか判定 TID 22:文章の自然性判定	TID 12:単語のアクセントが正しいか判定 TID 26:単語の読みを入力

タスクカテゴリに関しても同様のパラメタを用いた．また，本適用において各タスクカテゴリにおける作業タスク数が 50 以下のワーカーは作業量が少ないため平均精度としては用いず，そのタスクカテゴリにおけるタスクは処理していないものとして扱った．

また比較対象として，ベイジアンネットワーク以外に決定木を用いてグラフを作成し，同様の実験を行った．決定木は最も影響のある要素を見つけ出しデータを分割していく手法である．決定木を用いて学習タスクカテゴリを作成するにあたって，PCSS は最初に精度向上対象タスクカテゴリ A における決定木を作成する．作成された決定木の例を図 4 に示す．決定木アルゴリズムは対象となるタスクカテゴリの精度に影響のあるタスクカテゴリを分岐ノードとして扱う．図 4 では，タスクカテゴリ B の結果精度が 90%以上であった場合タスクカテゴリ A を高精度で処理できることを示している．すなわち図 4 のような決定木が得られた場合，タスクカテゴリ B はタスクカテゴリ A の学習タスクカテゴリとして扱うことができる．

決定木を求めるためのアルゴリズムとして J48 を用いている．二分割法は用いず，枝刈りの閾値は 0.25，葉の最小数は 2 を用いた．階層の制限は行わなかった．

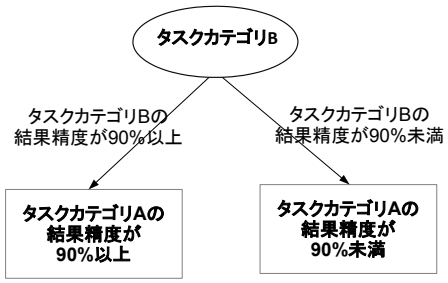


図 4 例：タスクカテゴリ A における決定木

各有向グラフから得られた精度改善対象となるタスクカテゴリと，それぞれのタスクカテゴリに対するベイジアンネットワークから得られた学習タスクカテゴリと，決定木から得られた学習タスクカテゴリは表 2 のようになる．

表 3 学習タスクカテゴリ実施の有無によるタスク改善効果

		ベイジアンネットワーク				決定木			
対象タスクカテゴリ	テストタイプ	テスト実施人数	対象人数	精度向上人数	平均精度向上値(point)	テスト実施人数	対象人数	精度向上人数	平均精度向上値(point)
TID 0:読点の位置が正しいか判定	練習タスクカテゴリ(ワーカーグループ 1)	7	3	3	11.2	8	8	4	-3.2
	同一タスクカテゴリ(ワーカーグループ 2)	8	3	2	3.4	31	17	6	1.3
	関係ないタスクカテゴリ(ワーカーグループ 3)	7	4	2	3.5	6	0	0	0
TID 1:語尾の発音チェック	練習タスクカテゴリ(ワーカーグループ 1)	24	8	6	6.7	12	5	2	-1.0
	同一タスクカテゴリ(ワーカーグループ 2)	13	4	1	-3.0	13	6	3	1.1
	関係ないタスクカテゴリ(ワーカーグループ 3)	13	7	4	1.6	11	5	3	1.2
TID 2:対話パターン作成	練習タスクカテゴリ(ワーカーグループ 1)	学習タスクカテゴリなし				13	7	3	-0.5
	同一タスクカテゴリ(ワーカーグループ 2)	学習タスクカテゴリなし				10	9	5	1.7
	関係ないタスクカテゴリ(ワーカーグループ 3)	学習タスクカテゴリなし				13	10	6	0.8
TID 3:有名人の読み仮名を入力する	練習タスクカテゴリ(ワーカーグループ 1)	8	7	6	8.2	決定木作成できず			
	同一タスクカテゴリ(ワーカーグループ 2)	8	7	3	-0.5	決定木作成できず			
	関係ないタスクカテゴリ(ワーカーグループ 3)	8	6	1	-0.3	決定木作成できず			
TID 4:キーワードを分類	練習タスクカテゴリ(ワーカーグループ 1)	12	10	8	5.2	14	8	4	2.4
	同一タスクカテゴリ(ワーカーグループ 2)	7	6	4	3.3	12	6	3	2.3
	関係ないタスクカテゴリ(ワーカーグループ 3)	8	6	2	0.6	11	3	2	0.6

4. 段階的学習手法の評価及および考察

4·1 実 験 方 法

3 章で得られたタスクカテゴリの相関関係の有効性を確認するために以下のような実験を行った．実験対象となるタスクカテゴリは実験コストの問題から，精度向上対象タスクカテゴリの上位 5 つである，「TID0:読点の位置が正しいか判定」，「TID1:語尾の発音チェック」，「TID2:対話パターン作成」，「TID3:有名人，芸能人の読み仮名を入力する」，「TID4:キーワードを分類」を対象とした．

(1) 初期状態の確認のために対象タスクカテゴリのタスクを 50 問処理させて精度チェックを行う．PCSS では該当カテゴリのタスクにおける精度が 60%以下になったワーカーはその作業をさせないという対応をしている [Ashikawa 14]．これは 70%以下のワーカーを排除対象とすると悪意の無いワーカーを多く排除してしまい，50% 以下のワーカーを対象とすると，悪意のあるワーカーを排除しきれなかったというシステム運用上の経験からの数値である．また，90% 以上のワーカーは既に優秀であるため学習の必要がないと判断し，この初期状態のチェックで正解率が 60%以上 90%以下のワーカーを以降の実験の対象とする．

(2) (1) の条件を満たしたワーカーを以下の 3 グループに分け，それぞれ別の作業を行わせる．

a (ワーカーグループ 1)
有向グラフから得られた学習タスクカテゴリのタスクそれぞれを 10 問ずつ実施させる．

b (ワーカーグループ 2)
学習タスクカテゴリではなく，精度向上対象タスクカテゴリから「10 × ワーカーグループ 1 の学習タスクカテゴリの数」問作業させる．

c (ワーカーグループ 3)
学習タスクカテゴリではなく，全く関係の無いタスクカテゴリから「10 × ワーカーグループ 1 の学習タスクカテゴリの数」問作業させる．

(3) (2) を実施後，(2) を処理した全員のワーカーにもう一度 (1) と同じタスクカテゴリの問題を 50 問処理させて精度改善効果をチェックする．

(4) (2)，(3) を 3 回繰り返す．

(5) 一番最初に行った (1) の結果精度と一番最後に行った (3) の結果精度を比較して改善効果を測定する．

この実験で実施した学習タスクカテゴリは公平を期すために，通常のタスク処理と同等の報酬をワーカーに支払っている．また，学習タスクはタスクの出題をコントロー

ルすることで，強制的に実施させている．

4·2 精 度 評 価

この実験によって得られた効果を表 3 に示す．各ワーカーグループには最初にワーカー全体からランダムに 40 人の固定人数を割り当て，実際に作業した人数をテスト実施人数とし，実際に作業した人数のうち，正解率が 60%以上 90%以下のワーカーの人数を対象人数としている．また，ベイジアンネットワーク，決定木それぞれにおけるワーカーグループ 1，2，3 は比較のために同じタイミングで他の作業をさせずに実験しているが，ベイジアンネットワークの実験と決定木の実験には時間的にずれがある．つまり，ベイジアンネットワークのワーカーグループ 2 と決定木のワーカーグループ 2，ベイジアンネットワークのワーカーグループ 3 と決定木のワーカーグループ 3 は同時には作業していない．これはシステム運用上，他のタスクも存在しており，すべてのワーカーを実験に注力することができないという点に起因している．そのため，ワーカーグループ 2，3 はベイジアンネットワーク，決定木それぞれに関して表に記載した．学習タスクカテゴリを実施したワーカーグループ 1 は，対象タスクカテゴリ TID0 では 11.2 ポイント，対象タスクカテゴリ TID1 では 6.7 ポイント，対象タスクカテゴリ TID3 では 8.2 ポイント，対象タスクカテゴリ TID4 では 5.2 ポイントと，平均 7.8 ポイントの精度向上が確認できた．対象タスクカテゴリ TID2 に関しては有向グラフは得ることができたが，精度向上対象タスクカテゴリに対して影響を及ぼしているタスクカテゴリ (学習タスクカテゴリ) が存在しなかった．これらの結果から，ベイジアンネットワークの学習から導出された学習タスクカテゴリを実施することによって，精度向上対象タスクカテゴリの処理結果精度が大きく向上していることが確認できる．

また，学習タスクカテゴリを実施せずに同じタスクカテゴリを実施したワーカーグループ 2 はわずかながらの改善が見られるが，これは同一のタスクカテゴリ処理を継続することで作業に慣れた結果であると推測している．

また，比較のために決定木を用いて得られた学習タスクカテゴリを実施したワーカーグループ 1 の精度向上効果は平均 −0.6 ポイント，同じタスクカテゴリを実施したワーカーグループ 2 の精度向上効果は平均 1.6 ポイント，関係のないタスクカテゴリを実施したワーカーグループ 3 の精度向上効果は平均 0.6 ポイントとあまり改善効果は見られなかった．また TID3 における決定木は得られなかった．これらの結果から決定木で導出された学習タスクカテゴリをワーカーが実施しても精度改善効果は大きくなく，同一タスクカテゴリの実施，関係のないタスクカテゴリを実施した場合の作業の慣れによる精度向上とほぼ変わらないことが確認できた．

また，得られた表 3 のベイジアンネットワークの結果から学習タスクを行って結果精度が向上した人数の合計:23 人 (表 3 におけるワーカーグループ 1 の「精度向上人数」の合計)，学習タスクを行ったが結果精度が向上しなかった人数の合計:5 人 (表 3 におけるワーカーグループ 1 の「対象人数」−「精度向上人数」の合計)，学習タスクを行わず結果精度が向上した人数の合計:19 人 (表 3 におけるワーカーグループ 2，3 の「精度向上人数」の合計)，学習タスクを行わず結果精度も向上しなかった人数の合計:24 人（表 3 におけるワーカーグループ 2，3 の「対象人数」−「精度向上人数」の合計）を用いて 2×2 のテーブルを作成し，カイ二乗検定を行ったところ得られた P 値は 0.0014 であった．この検証結果により有意性があると判断することができる．

4·3 考 察

実験結果から，マイクロタスク型クラウドソーシングにおけるワーカーの品質改善を行うにあたって，ベイジアンネットワークを用いてワーカーの行動履歴から学習タスクを導出する段階的学習手法は有効であるが決定木を用いた手法はあまり効果が得られないことがわかった．

タスクカテゴリ A を解析するにあたって，タスクカテゴリ A におけるワーカーの処理結果が高精度である確率 $P(A)$ を目的変数とし，A 以外のタスクカテゴリ $X_1, X_2, X_3 \cdots X_n$ におけるワーカーの処理結果が高精度である確率 $P(X_1, X_2, X_3 \cdots X_n)$ を説明変数としている．決定木を用いた解析では, 決定木の各ノードには分類する属性が対応付けられ, ノードを結ぶリンクには属性値が対応付けられる. 決定木を用いてタスクカテゴリ A を解析するにあたって, 属性を A 以外のタスクカテゴリ $X_1, X_2, X_3 \cdots X_n$, 属性値を処理結果が高精度であるか否かとして決定木を作成した. 決定木は影響の大きい要素を優先的に選択して作成されており, 本研究ではこの優先的に選択されている属性 (タスク) を学習タスクとして用いている (図 4). このように決定木の解析では説明変数と目的変数の関係だけを考えて解析を行っているが，実際のタスクカテゴリ同士は 3 章で述べたように，タスクカテゴリ X_i がタスクカテゴリ X_j の学習タスクカテゴリとなりうる可能性があるため，タスクカテゴリ X_i とタスクカテゴリ X_j は相互的に影響がないとは言いがたい．説明変数と目的変数の関係だけを考える決定木に対して, ベイジアンネットワークでは説明変数間の関係を学習しているため [岡本 08],「タスクカテゴリ X_i を高精度で処理することができたのでタスクカテゴリ X_j を高精度で処理することができた」という因果関係を含んだ $P(X_i \mid X_j)$ を学習することができており, 決定木よりも精度向上効果のある学習タスクを算出できたものと推測している．

本実験における意味のある因果効果とは，あるワーカーに今までと違う処理条件 (=学習タスク) を与えることで，そのワーカーの反応に望ましい変化が現れるという現象である．この因果効果を測定するために，学習タスク X_j

を実施させたワーカーグループでのタスク X_i における効果 Y_j と，学習タスク以外のタスク X_k を実施させたワーカーグループのタスク X_i における効果 Y_k とすると，それぞれの差，つまり学習タスクを実施することによる効果の期待値 E

$$E[Y_j - Y_k] = E[Y_j] - E[Y_k] \tag{5}$$

を集団での因果効果と定めることができる [宮川 04]．さらにワーカーに学習タスクを実施させるか (学習タスク X_j を実施)，させないか (学習タスク以外のタスク X_k を実施)，を示す変数を考え，これを確率変数 W として定式化する．この W も 2 値変数で，$W = 1$ は学習タスク X_j の実施を，$W = 2$ は学習タスク以外のタスク X_k の実施を意味する．すると，実験によって得られた結果である表 3 は $W = 1$ のワーカーにおける精度向上効果と Y_j と $W = 2$ のワーカーにおける精度向上効果 Y_k となる．よって

$$E[Y_j \mid W = 1] - E[Y_k \mid W = 2] \tag{6}$$

は計算することが可能となる．一般的には式 (5) と式 (6) は異なるが，本実験ではワーカーに学習タスクを実施させるか ($W = 1$) させないか ($W = 2$) は無作為に割りつけを行っているため，W と（Y_j，Y_k）は統計的に独立になる．このとき

$$E[Y_j \mid W = 1] = E[Y_j \mid W = 2] = E[Y_j] \tag{7}$$

$$E[Y_k \mid W = 1] = E[Y_k \mid W = 2] = E[Y_k] \tag{8}$$

が成り立つため [宮川 04]，式 (6) の条件付き期待値と式 (5) の期待値は等しくなる．このように，本実験は無作為実験であるため，集団の因果的効果を偏りなく推定できていると考えることができる．

さらに，決定木は影響の大きい要素を優先して解析していくため，どの順で解析したかが決定木の作成に大きく影響する．そのため，対象となるデータに外れ値や偏りが多く存在していた場合は優先度に影響を与えてしまう可能性が大きい．決定木の有効性が低い理由として，今回の解析対象となるクラウドソーシングのタスク処理はワーカーが不特定多数であり，品質が一定していないデータであるため，それらのデータも精度改善効果が得られない一因であると推測している．

本実験では，精度改善対象タスクカテゴリに対して有向グラフ上で直接影響を与えていると解析されたタスクカテゴリのみを学習タスクカテゴリとして用いている．例えば図 2 では，精度改善対象をタスクカテゴリ D とした場合，タスクカテゴリ B とタスクカテゴリ C のみを学習タスクカテゴリとし，タスクカテゴリ A は学習タスクカテゴリとして扱っていない．これは，タスクカテゴリ A はタスクカテゴリ B と C と比較してタスクカテゴリ D に対する影響力が少ないため計算量および実験コストを削減するために省略した．すべての影響あるタスクカテゴリを学習タスクカテゴリとして扱った場合の実験を低コストで行う方法は今後の課題である．また，精度向上タスクが学習タスクとなってしまった場合は，有向グラフにおいて 1 階層の相互に影響し合うループが発生する．この場合は精度向上対象タスクカテゴリを繰り返し実施する，すなわち 4 章の実験におけるワーカーグループ 2 のケースで精度が向上すると予測している．しかし，現時点ではこのようなケースは発生しておらず未検証であるため，検証は今後の課題である．

精度改善対象となるタスクカテゴリと, 得られた学習タスクカテゴリの間には一見関連性がないように見えるものも存在する. しかし，タスクカテゴリの内容的に関連性が少なくても，ベースとなる知識やユーザインタフェースなどの点で共通する点があるものと推測される．

従来の教育実践および教育システムにおいてはボトムアップ方式の教授方法が有効である [岡本 08]．クラウドソーシング環境においてボトムアップ方式の教授方法を用いる場合，学習タスクは精度向上対象のタスクよりも簡単であることが理想的である．しかし，クラウドソーシングにおいてタスク A がタスク B より「簡単である」とは「タスク B を処理するために必要な知識よりも少ない知識でタスク A の処理が可能」と考えた場合，学校教育のようにタスク A を処理するために必要な知識の部分知識のみで処理が可能な学習タスク B を設計し，タスク B でワーカーを学習させることは非常に困難である．クラウドソーシングでは大量のタスクが存在し，また教師という熟達した管理者も存在しないため，学習タスク B を作成することがシステム管理者，リクエスタにとっては非常に高コストであるためである．本研究は，システム管理者およびリクエスタに負担をかけることなくワーカーに学習させるために,「タスク A を処理するのに必要な知識の部分知識のみ」で構成された学習タスク B ではなく,「タスク A を処理するのに必要な知識の部分知識を含んだ」既存のタスク C で学習させることはできないかという仮説を立て，効果があることを実証している．

例えば，タスク A を処理するために必要な知識が知識 a であり，タスク C を処理するために必要な知識が知識 a の部分知識 a’ と知識 c だった場合，タスク C を処理するにはタスク A よりも多くの知識が必要である場合がある．しかし，そのようなケースでもワーカーが知識 c を既にもっていた場合はタスク C はタスク A の学習タスクとなることができる．精度向上タスクカテゴリ「TID9 : 単語の品詞を選択する」と学習タスクカテゴリ「TID2 : 対話パターン作成」の場合，TID9 を処理するための必要な知識は「日本語文法の知識 (a)」だが，TID2 を処理

するためには「一般的な日本語の知識 (a’)」と「対話文章の作文能力 (c)」が必要になる．ワーカーによっては正確ではない文法で日本語会話を行っている可能性があり (「一般的な日本語の知識 (a’)」と「対話文章の作文能力 (c)」の知識は持っているが「日本語文法の知識 (a)」の知識に乏しいケース)，その場合でも TID2 を処理する過程で様々な文章を作文していくうちに日本語の文法に慣れ親しんでいくことで TID9 の処理で必要な「日本語文法の知識 (a)」，すなわち「単語が人名であるかどうか」「単語が地名であるかどうか」「単語が名詞であるかどうか」などを学習していると考えている．

また，精度向上対象タスクカテゴリ「TID 4:キーワードを分類」と学習タスクカテゴリ「TID 17:熟語のアクセントが正しいか判定」の場合，TID4 を処理するために必要な知識は「日本語の単語に関する知識 (a)」だが，TID17 を処理するためには「一般的な日本語の知識 (a’)」と「アクセントに関する知識 (c)」が必要になる．ワーカーによっては文法や単語を意識せずに一般的な発音で会話しているケースが存在する (「一般的な日本語の知識 (a’)」と「アクセントに関する知識 (c)」は持っているが，「単語」の概念など「日本語の単語に関する知識 (a)」の知識に乏しいケース)．その場合でも TID17 で文章の中の熟語のアクセントを処理する過程で文章のどの部分が熟語となっているかなどをを確認し，「日本語の単語に関する知識 (a)」，すなわち「単語は文章のどこで切るのか」などを学習していると考えている．

同様に「TID 3:有名人，芸能人の読み仮名を入力する」において必要な知識は「日本語に読み仮名を入力する知識 (a1)」，「芸能人に関する知識 (a2)」，「最新の情報をチェックする能力 (a3)」と考えることが可能であり，「TID 4:キーワードを分類」で必要な知識は「日本語の単語に関する知識 (a1’)」，「TID 11:芸能人のグループ名を入力」で必要な知識は「芸能人に関する知識 (a2)」，「TID 14：ＩＴ分野の文の自然性判定」で必要な知識は「最新の情報をチェックする能力 (a3)」，「ＩＴ関連の知識 (c3)」と考えることが可能である．この場合，精度向上タスクカテゴリ TID 3 に対して，学習タスクカテゴリ TID4 の知識 a1’ は TID3 で必要な知識 a1 に，学習タスクカテゴリ TID11 の知識 a2 は TID3 で必要な知識 a2 に，学習タスクカテゴリ TID14 の知識 a3 は TID3 で必要な知識 a3 に，それぞれ影響を及ぼしていると考えている．

さらに，ユーザインタフェースの設計から学習効果を考察する．精度向上対象タスクカテゴリ「TID 0:読点の位置が正しいか判定」における学習タスクカテゴリ「TID 7:助詞の選択」は選択式のユーザインタフェースで日本語文章の特定の点に関して回答するという点，精度向上対象タスクカテゴリ「TID 1:語尾の発音チェック」における学習タスクカテゴリ「TID 12:単語のアクセントが正しいか判定」は与えられた複数音声データを再生し，正しい音声を選択するというユーザインターフェースという点，精度向上タスクカテゴリ「TID 3:有名人の読み仮名を入力する」における学習タスクカテゴリ「TID 11:芸能人のグループ名を入力」は人名をテキストで入力するユーザインタフェースという点，精度向上タスクカテゴリ「TID 4:キーワードを分類」における学習タスクカテゴリ「TID 17:熟語のアクセントが正しいか判定」は選択式のユーザインタフェースという点でそれぞれ共通点を持つと考えられる．しかし，PCSS はマイクロタスク型のクラウドソーシングなので，ユーザインタフェースは非常に簡易なものが中心となっており，ユーザインタフェースを原因とした精度低下が発生しているとは考えにくい．そのため，学習タスクの導出にはユーザインタフェースなどの類似性ではなく，必要としている知識の類似性が重要であると考えている．

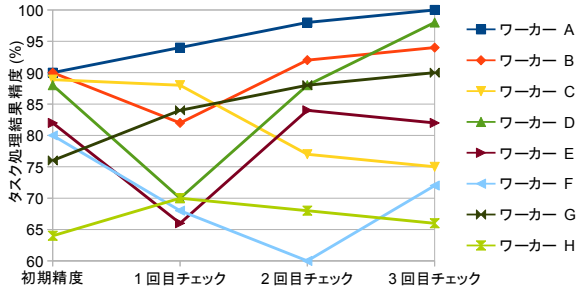


図 5 精度改善対象タスクカテゴリ TID1 におけるワーカーの成長パターン

このようにクラウドソーシング環境では発見が困難な学習タスクを自動的に解析，抽出することが可能になったという点が，本研究の貢献点であると考えている．

また，ワーカーのモチベーション低下はタスクの処理結果精度の低下につながり，ワーカーのモチベーションは報酬やタスクの面白さに影響されることがわかっている [Kittur 08]．今回実施した実験では対象となるワーカーは割り当てられた学習タスクカテゴリ以外のタスクを行うことができなかったため，モチベーションの低下の原因となり，最終的に精度が低下するケースや，最後まで作業を完了させずに途中でやめてしまうケースにつながってしまった可能性がある．より正確に段階的学習手法の効果を確認するためにモチベーションの低下しないテスト方法の構築も今後の課題としたい．

今回の実験で得られたワーカーの成長パターンは様々であった．TID1 におけるワーカーの成長パターンを図 5 に示す．ワーカーの成長パターンを，(1) 継続的に精度が向上していくパターン (ワーカー A，G)，(2) 精度は上下するが最終的に向上するパターン (ワーカー B，D，

表 4 悪影響を与える可能性のあるタスクカテゴリ

精度向上対象タスクカテゴリ	悪影響を与えるタスクカテゴリ候補
TID 0:読点の位置が正しいか判定	TID12:単語のアクセントが正しいか判定 TID14:ＩＴ分野の文の自然性判定 TID15:合成音の印象の度合を評価
TID 1:語尾の発音チェック	TID4:キーワードを分類 TID11:芸能人のグループ名を入力
TID 2:対話パターン作成	TID10:医学用語の読みを入力する TID14:ＩＴ分野の文の自然性判定 TID28:2 文字単語の読みの分割 TID29:映画名の読み付け TID30:会社の読み仮名付け

表 5 悪影響を与える可能性のあるタスクカテゴリによる影響

精度向上対象タスクカテゴリ	テスト実施人数	対象人数	精度低下人数	平均精度低価値 (point)
TID0:読点の位置が正しいか判定	19	11	6	0.8
TID1:語尾の発音チェック	20	11	3	－ 1.8
TID2:対話パターン作成	35	10	7	－ 1

E，H)，(3) 精度は上下するが最終的に低下するパターン (ワーカー F)，(4) 精度が継続的に低下するパターン (ワーカー C)，の 4 パターンに分類した．今回の実験ではパターン (1) と (2) のワーカー数の和がパターン (3) と (4) のワーカー数の和よりも多いため，段階的学習手法は効果があると判断している．例えば TID1 においてはパターン (1) とパターン (2) のワーカー数の和は 6 人，パターン (3) と (4) のワーカー数の和は 2 人である．パターン (3) と (4) のような最終的に低下してしまうケースはワーカーの個人情報などが影響しているのではないかと推測しており，個人情報を加味した段階的学習手法の提案を検討したい．

今回の実験では 5 つの精度改善対象タスクカテゴリすべてにおいて有向グラフを得ることができた．しかし，TID2 に関しては有向グラフを得ることができたが学習タスクカテゴリを得ることができていない．これは，精度改善対象タスクカテゴリが有向グラフのトップに配置されてしまったためである．我々の以前の実験 [Ashikawa 14] では 5 つの精度改善対象タスクカテゴリに対して 2 つの有向グラフしか得ることができていない．この差は解析対象となったワーカーの処理したタスクカテゴリの数によるものと推測される．以前の実験では 700 万タスクから得られた 50 タスクカテゴリを用いたのに対し，今回の実験では 1853 万タスクから得られた 138 タスクカテゴリに解析対象が拡大している．この結果から，ワーカーの行動履歴の量は学習タスクの導出に影響を与えていることがわかる．

これまで述べてきたように，学習タスクはワーカーの行動履歴から導出される．しかしワーカーの行動履歴は常に増加し，さらに段階的学習手法によって変化していく．そのため効果的な学習タスクを導出するには適宜ワーカーの最新の行動履歴を解析しなくてはならないが，計算量が大きく非常に高コストである．ワーカーの成長に応じて効果的に段階的学習を行わせるためには，最適な解析スケジュールの設定が急務である．

本研究における段階的学習法は 3 章で述べたように，タスク A を実施してからタスク B を実施した場合と，タスク A を実施せずにタスク B を実施した場合で，多くのワーカーが前者のケースでタスク B の処理結果が向上していた場合，タスク A をタスク B の練習用タスクとして扱うという手法を提案している．一方で，タスク A を実施してからタスク B を実施した場合と，タスク A を実施せずにタスク B を実施した場合で，多くのワーカーが前者のケースでタスク B の処理結果が低下していた場合，タスク A はタスク B の処理結果に悪影響を与えるのではないかと考えることもできる．この仮説を検証するために，表 1 の精度改善対象タスクカテゴリの TID0，TID1，TID2 に関してベイジアンネットワークを用いて負の影響を考慮した有向グラフを作成した．有向グラフを作成するにあたって $P(A)$ をタスクカテゴリ A の平均正解率が 80%以下の確率，$P(A \mid B)$ を「タスクカテゴリ B の平均正解率が 80%以下であった場合にタスクカテゴリ A の平均正解率が 80%以下である確率」とした．閾値として定めた 80%は，PCSS では該当カテゴリのタスクにおける精度が 60%以下になったワーカーはその作業をさせないという対応をしているため，60%では対象となるワーカーがおらず，70%では対象となるワーカー数が少ないという運用上の経験からの数値である．

精度改善対象タスクカテゴリの TID0，TID1，TID2 に対して表 4 のような有向グラフを得ることができた．

これらの得られた悪影響を与えるタスクカテゴリ候補を用いて，4 章の実験条件におけるワーカーグループ 1 と同様の実験を行ったところ，表 5 のような結果となった．

表 5 の結果から，精度向上対象タスクカテゴリに悪影響を与えると仮定したタスクカテゴリでは，実際にはワーカーに悪影響を及ぼすことはないことがわかる．これは

タスクを処理することによって何らかの知識を得ることはできても，知識を失うことはないことなどが原因であると予想している．

5. まとめと今後の課題

本研究ではマイクロタスク型における精度向上手法を導入したプライベートクラウドソーシングシステムにおいて，ワーカーの能力を向上させるための段階的な学習法を提案した．これまでの運用履歴からタスクを自動でカテゴリ化し，各タスクカテゴリにおけるワーカーの行動履歴をベイジアンネットワークの学習に用い，得られた有向グラフからタスクカテゴリ間の関係性を学習タスクの作成に用いることで，有効な学習タスクの選択を自動的に行うことができた．また，段階的学習手法の効果を確認するため，結果精度の低いタスクカテゴリに対してベイジアンネットワークで有向グラフの構築を行ったところ，平均 7.8 ポイントの精度向上効果が得られることを確認した．

ベイジアンネットワークで得られた学習タスクを用いて段階的学習手法を行うことで精度向上効果があることは確認できたが，あまり効果が見られないワーカーも存在した．これはタスクの難易度に起因するものではなく，ワーカーのモチベーションや集中力に起因するものであると推測している．ワーカーのモチベーションのコントロールはゲーミフィケーション的なアプローチが効果があるため [Ahn 08]，段階的学習手法とゲーミフィケーションを組み合わせたアプローチは今後の課題である．

また，今回の実験により，ワーカーのプロフィールを視野に入れた学習タスクの生成や，モチベーション維持が可能なテスト方式などさらなる精度向上の可能性があることがわかった．今後，これらの可能性の精査と現状の段階的学習手法への組み込みを行うことによって，さらなるワーカーの精度向上を実現していきたい．

本論文に掲載のサービス等の名称は，それぞれ各社が商標として使用している場合があります．

◇ 参 考 文 献 ◇

[Ahn 08] Ahn，L.，Dabbish，L. : “Designing games with a purpose”，Communications of the Association for Computing Machinery，pp.58-67，(2008).

[Almond 09] Almond，R.，et al. : “Bayesian networks: A teacher 2019s view”，International Journal of Approximate Reasoning 50.3，pp.450-460，(2009).

[Ashikawa 14] Ashikawa，M.，Kawamura，T. and Ohsuga，A. : “Speech synthesis data collection for visually impaired person”，Third AAAI Conference on Human Computation and Crowdsourcing. 2014，(2014).

[Bachrach 12] Bachrach，Y.，et al. : “How to grade a test without knowing the answers—A Bayesian graphical model for adaptive crowdsourcing and aptitude testing”，arXiv preprint arXiv: 1206.6386，(2012).

[Bragg 14] Bragg，J.，Weld，DS. : “Crowdsourcing multi-label classification for taxonomy creation”，First AAAI Conference on Human Computation and Crowdsourcing，(2013).

[Burnap 13] Burnap，A.，et al. : “A simulation based estimation of crowd ability and its influence on crowdsourced evaluation of design concepts”，ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers，pp.V03BT03A004-V03BT03A004，(2013).

[Butz 06] Butz，C.，Hua，S. and Maguire，B. : “A Web-based Bayesian Intelligent Tutoring System for Computer Programming”，Web Intelligence and Agent Systems，Vol.4，No.1，pp.77-97，IOS Press，(2006).

[Carpenter 11] Carpenter，B. : “A hierarchical Bayesian model of crowdsourced relevance coding”，TREC，(2011).

[Fernandez 11] Fernandez，A.，et al. : “A system for relevance analysis of performance indicators in higher education using Bayesian networks”，Knowledge and information systems 27.3，pp.327-344，(2011).

[Garcia 07] Garcia，P.，et al. : “Evaluating Bayesian networks’ precision for detecting students’ learning styles”，Computers & Education，49.3， pp.794-808，(2007).

[Hoogerheide 12] Hoogerheide，L.，Block，JH. and Thurik，R. : “Family background variables as instruments for education in income regressions: A Bayesian analysis”，Economics of Education Review， 31.5，pp.515-523，(2012).

[Hutton 12] Hutton，A.，Liu，A. and Martin，C.E. :“Crowdsourcing evaluations of classifier interpretability”，AAAI Spring Symposium: Wisdom of the Crowd，(2012).

[Kamar 12] Kamar，E.，Hacker，S. and Horvitz，E. : “Combining human and machine intelligence in large-scale crowdsourcing”，Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1. International Foundation for Autonomous Agents and Multiagent Systems，pp.467-474，(2012).

[Kittur 08] Kittur，A.，Chi，E. and Suh，B. : “Crowdsourcing user studies with mechanical turk”，Human Computation & Crowdsourcing，pp.453-456，(2008).

[Lin 12] Lin，CH. and Weld，D. : “Crowdsourcing control: Moving beyond multiple choice”，arXiv preprint arXiv: 1210.4870，(2012).

[May 06] May，H. : “A multilevel Bayesian item response theory method for scaling socioeconomic status in international studies of education”，Journal of Educational and Behavioral Statistics，31.1，pp.63-79，(2006).

[宮川 04] 宮川 雅巳：統計的因果推論，朝倉書店，(2004).

[Nushi 15] Nushi，B.，et al. : “Crowd Access Path Optimization: Diversity Matters”，Third AAAI Conference on Human Computation and Crowdsourcing，(2015).

[岡本 08] 岡本 敏雄，香山 瑞恵：人工知能と教育工学,，オーム社，(2008).

[Pardos 10] Pardos，ZA.，et al. : “Using fine-grained skill models to fit student performance with Bayesian networks”，Handbook of educational data mining pp.417-425，(2010).

[Raykar 14] Raykar，VC. and Agrawal，P. : “Sequential crowdsourced labeling as an epsilon-greedy exploration in a Markov decision process”，AISTATS，pp.832-840,(2014).

[Shaw 11] Shaw，AD.，Horton，JJ. and Chen，DL. : “Designing incentives for inexpert human raters”，Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work， ACM，pp.275-284，(2011).

[Simpson 15] Simpson，E. and Roberts，S. : “Bayesian methods for intelligent task assignment in crowdsourcing systems”，Decision Making: Uncertainty，Imperfection，Deliberation and Scalability，pp.1-32，Springer International Publishing，(2015).

[Sun 12] Sun，Y. and Dance，C. : “When majority voting fails: Comparing quality assurance methods for noisy human computation environment”，arXiv preprint arXiv: 1204.3516，(2012).

[Tang 11] Tang，W. and Lease，M. : “Semi-supervised consensus labeling for crowdsourcing”，SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval (CIR)，pp.1-6，(2011).

[Ueno 00] Ueno，M. :“Intelligent tutoring system based on belief net-

works”，International Workshop on Advanced Learning Technologies，pp141-142，(2000).
[Venanzi 15] Venanzi，M.，et al.：“The ActiveCrowdToolkit: An open-source tool for benchmarking active learning algorithms for crowdsourcing research”，Third AAAI Conference on Human Computation and Crowdsourcing，(2015).
[Wais 11] Wais，P.，et al.：“Towards large-scale processing of simple tasks with mechanical turk”，Third AAAI Conference on Human Computation and Crowdsourcing,(2011).
[Wauthier 11] Wauthier，F. L. and Jordan，M. I.：“Bayesian bias mitigation for crowdsourcing”，Advances in Neural Information Processing Systems，pp.1800-1808，(2011).
[Xenos 04] Xenos，M.：“Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks”，Computers & Education 43.4，pp.345-359，(2004).
[Xie 15] Xie，H.，Lui，J.C.S. and Towsley，D.：“Incentive and reputation mechanisms for online crowdsourcing systems”，Quality of Service (IWQoS)，2015 IEEE 23rd International Symposium on. IEEE，pp207-212，(2015).

〔担当委員：奥 健太〕

2016 年 8 月 25 日　受理

著　者　紹　介

芦川　将之（正会員）

1999 年早稲田大学理工学部情報学科卒業．　2001 年同大学院理工学研究科修士課程修了．現在，株式会社東芝インダストリアル ICT ソリューション社および研究開発センターにてクラウドソーシング，大規模データ処理の研究・開発に従事．

川村　隆浩（正会員）

1994 年早稲田大学大学院 理工学研究科電気工学専攻 修士課程修了．同年，（株）東芝入社．2001-2002 年米国カーネギー・メロン大学 ロボット工学研究所 客員研究員 兼任．2003 年より電気通信大学大学院 情報システム学研究科 客員准教授 兼任．2007 年より大阪大学大学院 工学研究科 非常勤講師 兼任．2015 年より科学技術振興機構 情報分析室 主任調査員．現在に至る．工学博士 (早稲田大学)．2012 年 ISWC 10-Year Award 受賞．2013 年本学会研究会優秀賞受賞．本学会理事，代議員等を歴任．主として知識抽出，セマンティック Web に関する研究・開発に従事．

大須賀　昭彦（正会員）

1958 年生．1981 年上智大学理工学部数学科卒．同年 (株) 東芝入社．同社研究開発センター，ソフトウェア技術センター等に所属．1985-1989 年 (財) 新世代コンピュータ技術開発機構 (ICOT) 出向．2007 年より電気通信大学大学院情報システム学研究科教授．現在，同大学情報理工学研究科教授．2012 年より国立情報学研究所客員教授兼任．工学博士 (早稲田大学)．ソフトウェア工学，エージェント等の研究に従事．1986 年度情報処理学会論文賞，2013 年度人工知能学会研究会優秀賞受賞．IEEE Computer Society Japan Chapter Chair，人工知能学会理事，日本ソフトウェア科学会理事，同学会監事等を歴任．人工知能学会，情報処理学会，電子情報通信学会，日本ソフトウェア科学会，電気学会，IEEE Computer Society 各会員．