特　集

クラウドソーシングと機械学習

Crowdsourcing and Machine Learning

鹿島 久嗣　Hisashi Kashima
梶野 洸　Hiroshi Kajino
東京大学 情報理工学系研究科 数理情報学専攻
Department of Mathematical Informatics,
Graduate School of Information Science and Technology, The University of Tokyo
{kashima, hiroshi_kajino}@mist.i.u-tokyo.ac.jp

keywords: crowdsourcing, machine learning, natural language processing, computer vision, human computation

1. はじめに

人間には可能だが，機械による自動化は極めて難しいとされる自然言語理解や画像理解，音声認識などの領域において，機械学習をはじめとするデータ駆動アプローチが主流となって久しい．そして近年，機械学習に用いるデータを収集する目的で，インターネットを介して世界中の不特定多数の人間に作業を安価で依頼することのできるクラウドソーシングサービスの利用が増加している．しかし，従来の比較的コントロールされた環境下において取得されたデータと比較すると，クラウドソーシングによって得られるそれは，その手軽さと引き換えに品質に大きなばらつきを持つ．この問題を解決するためには，作業者の能力や適正等を適切に見積もり，信頼のおける情報を適切にフィルタリングするなど，データの品質のばらつきに対処する新しいデータ解析技術が必要になる．

本稿では特に機械学習技術とクラウドソーシングの関わりに着目し，クラウドソーシングの利用を明確に意識した機械学習の最近の研究動向を紹介する．まず 2 章ではデータ収集のための新たな手段として台頭してきたクラウドソーシングとその品質管理問題について述べる．3 章では特に教師付き学習の正解データ取得の目的にクラウドソーシングサービスを用いた場合に，成果物として得られたデータの質を向上させるための統計的な手法を紹介する．そして 4 章ではクラウドソーシングによって得られたデータから直接的にモデルを推定するための方法と，機械学習分野におけるクラウドソーシング研究の最近の発展について紹介する．

そして最後に上記の文脈からは若干外れるが，ヒューマンコンピュテーションと呼ばれる，計算過程の中に人間の手による処理が明示的に取り込まれたシステムにおけるクラウドソーシングの利用と，そのための機械学習技術について 5 章で触れる．

2. クラウドソーシング：データ収集のための新たな手段とその品質管理問題

2·1 クラウドソーシングの台頭

ウェブの普及は人間のコミュニケーションや情報アクセスの形と効率を大きく変化させた．そして今，それは人の働き方だけではなく，雇用形態そのものにも影響を与えつつある．クラウドソーシング[*1][Howe 06] とは米 Wired 誌の寄稿編集者ジェフ・ハウ氏によって名づけられた「（インターネットを通じて）不特定多数の人に仕事を依頼すること，もしくはその仕組み」一般を指す比較的新しい言葉である．この名前は業務の一部を外部に委託する「アウトソーシング」を捩ったものであり，通常のアウトソーシングの委託先が素性の知れた特定の相手であるのに対して，クラウドソーシングでは（時には匿名の）不特定多数の相手に仕事を依頼するのが特徴である．クラウドソーシングには報酬や公募形式の有無など様々な形態があり，その目的もビジネスから科学，あるいは社会への貢献など様々である．クラウドソーシングの初期の例としては米 P&G 社が同社の持つ技術的な課題についての研究開発を広く一般に公開し解決策を募る取組みを行ったものが有名である．Wikipedia もまた不特定多数の人間がその編集に関わるという意味でクラウドソーシングの一種であるといえよう．そして近年ではクラウドソーシングをサポートする様々なサービスが広く用いられるようになってきており，例えば米 InnoCentive 社は前述の研究開発の委託を仲介するサービスを提供している．クラウドソーシングサービスの中でも代表的なのは 2005 年に米 Amazon によって開始されたクラウドソーシングのプラットフォームである Amazon Mechanical Turk[*2]（以降 AMT）であろう．AMT では世界中にいるワーカー（AMT では Turker と呼ばれる）に対して比較的単純な作業を安価[*3]で依頼することができる．作業

∗1　同じ「クラウド」でもクラウドコンピューティングの「クラウド」は cloud（雲）であるのに対しクラウドソーシングのそれは crowd（群衆）であることに注意する．
∗2　https://www.mturk.com/mturk/welcome
∗3　Ipeirotis による AMT を使った調査によれば，平均的なワーカーは週に 8 時間以内の労働を AMT 上で行い 10 米ドル程度

内容は翻訳作業や Web サイトの感想の入力，画像データへの注釈作業など多岐にわたる．現時点では AMT における仕事の発注元は US 在住者のみに限定されているが，日本国内においても徐々にではあるが同様のサービスが浸透しつつある状況である*4．

2·2 コンピュータサイエンス研究における クラウドソーシングサービス利用

近年，研究コミュニティにおいて，クラウドソーシングサービスを研究のためのデータ収集目的で利用するという流れが起こっている．とくにその傾向は自然言語処理やコンピュータビジョンなどの，特別なスキルを持っていない一般の人にもある程度は可能であるようなタスクにおいてより顕著である．例えば自然言語処理分野において Snow らは文章に伴う感情の種類や，単語の類似度判定などいくつかのタスクに対して AMT を利用したデータ収集を行い，その可能性を示した [Snow 08]．より最近では，国際会議 NAACL-HLT において開催された自然言語処理におけるクラウドソーシング利用についてのワークショップにおいて，参加グループがそれぞれ 100 米ドルの資金をもとにクラウドソーシングサービスを用いた様々なタスクを行うという試みがある*5 [Callison-Burch 10]．コンピュータビジョン分野では Sorokin と Forsyth がオブジェクトの切り出しや人の身体箇所の指定などのタスクへの適用を試みている [Sorokin 08]．クラウドソーシングサービスを用いて収集されたデータはそのまま利用されることもあれば，機械学習，中でも特に教師つき学習を用いる際に必要な教師データ（正解ラベル）として用いられることが多い．教師データの取得には専門家による検討を必要とすることも多く，ここが機械学習技術の適用における金銭的，時間的なボトルネックとなってしまう．一方，クラウドソーシングサービスを用いることで比較的安価で大量に教師データを取得することが可能となるため，近年その利用への期待が高まっている．

また，ここ数年盛んに研究が行われている分野として，人間の労働力を計算のための資源と捉えることで，コンピュータと人間の両方を適切に組み合わせ問題解決を図るヒューマンコンピュテーションがある．ヒューマンコンピュテーションにおけるアルゴリズムは，そのフローの中で人間に繰り返し作業を依頼することによって計算を行うことになるが，その労働力の供給源としてクラウドソーシングは重要な位置を占めている [Law 11]．

の収入を得ているようである [Ipeirotis 10a]．

*4 例えばランサーズ（http://www.lancers.jp/）など．

*5 ワークショップの Web サイト（http://sites.google.com/site/amtworkshop2010/data-1）では取集されたデータが公開されている．

2·3 クラウドソーシングサービスの品質管理問題

クラウドソーシングにおける重要な課題のひとつがワーカーによる成果物の品質管理問題である．例えば自然言語処理における注釈データの作成は，通常は十分に訓練された人間によって行われる．一方，クラウドソーシングにおいてはワーカーが課題を達成するための能力を十分にもっているということは保証されず，ワーカーには高い能力の者もいれば，そうでないものもいるという玉石混合といった状態にある．さらには報酬を得ることだけを目的としてでたらめなデータを生成する「スパムワーカー」も少なからず存在することが知られている．そこで，信頼度の高いデータを得るための何らかの工夫が必要となる．その方法のひとつとしてはクラウドソーシングサービスのシステムの機能として，仕事の品質を保つための仕組みを取り入れることが挙げられる．例えばワーカーがこれまでに行ったタスクの数や，依頼者による承認率などの指標を用いることでフィルタリングを行ったり，実際にタスクに取り掛かる前に当該タスクに取り掛かる素養があるかどうかをチェックするテストを行うような仕組みが取り入れられている．実際，クラウドソーシングサービスのひとつ CrowdFlower では正解の分かっている問題をいくつか課題の中に潜ませることでワーカーの能力を測る仕組みが提供されている*6．他にも，アフリカの言語の翻訳を行うワーカーはアフリカ在住者の方が向いているであろうといった理由から，場所によるフィルタリング等もまた有効である．より細かいテクニックとしては，翻訳タスクを依頼する際にワーカーが機械翻訳サービス（大抵はテキストデータのコピー&ペーストによって入力を受け付ける）を利用することを阻止するために原文を画像で表示するようにするなどの工夫もある．

3. クラウドソーシングの品質管理のための 機械学習

3·1 ワーカーの能力と真のラベルの推定

2·3 節では多くのクラウドソーシングサービスには結果の品質を上げるためのいくつかの工夫が組み込まれていることを紹介した．一方で，すでに結果が得られたあとの後処理によって結果の信頼度を高めるという方向性も十分に考え得る．例えば，同じタスクを複数のワーカーに依頼した場合に，得られた複数の結果を突き合わせて検証することによって結果の質を上げられることが期待できる．その突き合わせの方法としては単純には多数決によって行うのが良さそうであるが，信頼度の高い結果を得るためには，ひとつの単位タスク（例えば画像の注釈付けであるならば，1 枚の画像）に対して十分な数の結果を得る必要がある．そこで注目したいのは，通常は複数の単位タスクが同時に依頼されている（画像の注釈付けならば，注釈を付けたい画像は複数ある）という点で

*6 http://crowdflower.com/docs/gold

ある．依頼されたタスクが単に画像 1 枚のみについてであるのならば（ワーカーについての事前知識が何もない場合には）多数決以上に良い解はなさそうであるが，実際には 1 人のワーカーが複数の単位タスクについて作業を行っていることが期待できるため，そのような状況では上手く単位タスク間で情報を共有することによって多数決よりもよい解が得られそうである．実はこのように信頼度の低い複数の回答から正解を推定するという問題は，1970 年代後半に統計科学の分野で既に扱われていた．Dawid と Skene は複数の医師の診断結果から適切な診断を下すという医療診断の文脈においてこの問題を考えた [Dawid 79]．彼らはこの状況を，隠された正解に対して各医師が摂動を加えたもの（医師ごとに異なった確率で確率的に診断が変化する）が観測されるという確率的な生成モデルによってモデル化した．

Dawid と Skene の方法が機械学習の分野において適用されたのは，恐らく 1995 年に発表された Smyth らによるものが初めてである [Smyth 95]．彼らは与えられた金星の表面の画像内の火山の有無を判定するという画像認識のタスクにこのモデルを適用している．近年になって機械学習分野では複数の（必ずしも信頼できない）ワーカーによって生成されたデータから真のラベルを推定する問題が多く取り扱われているが，Dawid と Skene の先駆的な研究はこれらの最近の研究の基本となるものであるため，ここで少し詳しく紹介しておくことにする．なお，彼らのもともとの提案手法は多値ラベルの場合を扱っているが，ここでは単純に 2 値のラベルのみを考えることにする．

Dawid と Skene による問題設定を形式的に述べる．N 個の単位タスクがあり，それぞれに対する真の答えを $y_i \in \{0,1\}$ とする（$i = 1,2,\ldots,N$）．一方ワーカーの数は J 人であるとし，i 番目のタスクに対して回答を行ったワーカーの集合を $\mathcal{J}_i \subseteq \{1,2,\ldots,J\}$ とする（$i = 1,2,\ldots,N$）．（全てのワーカーが全ての単位タスクに回答する必要はないことに注意する．）目標はワーカーによって与えられたラベルの集合 $\{y_i^j\}_{j\in\mathcal{J}_i, i\in\{1,2,\ldots,N\}}$ を手掛かりに真のラベル $\{y_i\}_{i\in\{1,2,\ldots,N\}}$ を推定することである．

Dawid と Skene のモデルでは各ワーカーが正しく回答する確率をモデル化する．j 番目のワーカーが真のラベルが 1 であるときに正しく 1 と答える確率を α^j，一方で真のラベルが 0 であるときに正しく 0 と答える確率を β^j とする．つまり $\alpha^j = \beta^j = 1$ のときワーカー j は必ず正しく回答する．真のラベルとワーカーの回答の次の過程に従って生成されるものとする．まずそれぞれの単位タスクに対して確率 p で真のラベルが 1 に（確率 $1-p$ で 0 に）決定される．次に，各ワーカーが真のラベルと自らのパラメータ（α^j と β^j）をもとに回答を行う．真のラベルが与えられたもとでは，ワーカーの解答は互いに独立であると仮定する．

さて，全てのモデルパラメータ $\{\alpha^j\}_{j\in\{1,2,\ldots,J\}}$ と $\{\beta^j\}_{j\in\{1,2,\ldots,J\}}$ および p が分かっていれば，単位タスク i に対する真の答え y_i をワーカーの回答 $\{y_i^j\}_{j\in\mathcal{J}_i}$ から次の式によって推定することができる．

$$
\begin{aligned}
&\Pr[y_i = 1 \mid \{y_i^j\}_{j\in\mathcal{J}_i}] \\
&\propto \prod_{j\in\mathcal{J}_i} \Pr[y_i^j \mid y_i = 1]\Pr[y_i = 1] + \Pr[y_i^j \mid y_i = 0]\Pr[y_i = 0] \\
&= \prod_{j\in\mathcal{J}_i} (\alpha^j)^{y_i^j} \cdot (1-\alpha^j)^{1-y_i^j} \cdot p \\
&\qquad + (1-\beta^j)^{y_i^j} \cdot (\beta^j)^{1-y_i^j} \cdot (1-p)
\end{aligned}
$$

一方で，仮に真の答え $\{y_i\}_{i=1}^N$ の値が分かっているのならばここから p を，また真の答えと各ワーカー j の回答 $\{y_i^j\}_{\{i|j\in\mathcal{J}_i\}}$ を照らし合わせればモデルパラメータ α^j と β^j も簡単に推定することができる．しかし，真の答えもまた未知であり，これこそが我々が推定したいものである．Dawid と Skene は真の答えとワーカーのパラメータの両方を推定するため，片方を既知としてもう片方を推定するということを繰り返す推定方法（真の答えを潜在変数とする EM アルゴリズム）を提案した．これは直感的には，各単位タスクにおける多数決において多数側により入っている回数の多いワーカー（つまり常に多数派であるようなワーカー）の信頼度が高くなるような推定方法となっている．

3·2 タスク難易度のモデル化

Dawid と Skene のモデル [Dawid 79] はワーカーの能力にはばらつきがあるという事実をモデルに取り入れたものであるが，一方でタスクのほうも簡単な単位タスクから難しいものまでその難易度には様々なものがありそうである．そこで Whitehill らはタスクの難易度を明示的に取り入れたモデルを提案している [Whitehill 09]．ワーカー j の能力を ω^j，単位タスク i の簡単さを η_i (≥ 0) とすると，ワーカー j が単位タスク i に正解する確率を

$$
\Pr(y_i^j = y_i) = \frac{1}{1+\exp(-\omega^j\eta_i)}
$$

と定義する．exp の中身が大きいほどその確率は 1 に近づき，小さいほど 0 に近づくことがわかる．つまり $\omega^j = +\infty$ のときワーカー j は必ず正解し，$\omega^j = -\infty$ のときには必ず不正解となることがわかる．またちょうど $\omega^j = 0$ のときには，ちょうど 1/2 の確率で正解することになる．一方，η_i が大きいほどその単位タスクが簡単であり，正解される確率も高くなる．

Whitehill らのモデルは各単位タスクの普遍的な難易度をモデルに取り入れたものであったが，さらにワーカー毎の各単位タスクに対する得手不得手までを考慮したのが Welinder らのモデルである [Welinder 10]．各単位タスクが D 次元の実数値ベクトル $\boldsymbol{x}_i \in \Re^D$ として潜在的

に表現されるものとする*7．一方，各ワーカーは同じく D 次元の決定パラメータ $\boldsymbol{w}^j \in \Re^D$ および閾値パラメータ $\tau^j \in \Re$ を持つものとする．そして，ワーカー j は単位タスク i に対して ${\boldsymbol{w}^j}^\top \boldsymbol{x}_i > \tau^j$ のときラベル 1 を，そうでないときにはラベル 0 を与えるとするのが Welinder らのモデルである．彼らのモデルは Whitehill らのモデルを多次元に拡張したものと考えることができ，D 次元の潜在空間に配置された単位タスクとワーカーの位置関係によって回答が決定されるため，これらの間の相性を捉えることができる．

3·3 低品質ワーカーの排除

2·3 節においてスパムワーカーと呼ばれる低品質ワーカーの存在について触れたが，継続的にクラウドソーシングサービスを用いて高い品質のモデルを得られるようにするためには彼らを特定して排除することが重要である．前述の方法 [Dawid 79] によって得られた各ワーカーのパラメータからワーカーの品質を測る方法を紹介する．

まず考えられるのは，(推定された) 正解と一致した回答を数多く行っているワーカーほど品質が高いとするものである．実際 Dekel と Shamir は同様の考え方に基づく方法を提案している [Dekel 09]．彼らはまず全てのデータを用いて予測モデルを構築したあと，このモデルによる予測と大きく食い違う回答を行っているワーカー（および彼らによって作られたデータ）を排除し，残りのデータで改めてモデルを構築しなおすことによってモデルの予測性能が向上することを示している．

この考え方は一応理にかなっているように思えるが，少しよく考えてみると正解と必ず食い違う回答を出すワーカーは，タスクの意図を正反対に誤解しているか，あるいは（能力は高いが）悪意のあるワーカーである可能性があり，必ずしも能力が低いワーカーとは言えないことに気付く．(特に 2 値ラベルの場合にはラベルを反転させることで必ず正解になる．) 従って，正解率は必ずしもワーカーの能力の指標にはならない恐れがある．Ipeirotis らは，ワーカーの回答に含まれる不確定性は能力とその他の要素（彼らはこれを「バイアス」と呼んだ）の 2 つに分けて考えることを提案し，バイアスを排除することでより正確に能力を測ることを試みた [Ipeirotis 10b]．具体的には，ワーカー j によって i 番目の問題にラベル 1 が与えられた（$y_i^j = 1$）とき，その観測から推定される真のラベルが 1 となる ($y_i = 1$) 確率は，ベイズの定理より

$$\begin{aligned}\Pr[y_i = 1 \mid y_i^j = 1] &= \frac{\Pr[y_i^j = 1 \mid y_i = 1]\Pr[y_i = 1]}{\Pr[y_i^j = 1]} \\ &= \frac{\alpha^j p}{\alpha^j p + (1-\beta^j)(1-p)}\end{aligned}$$

となる．($y_i^j = 0$ の場合も同様に計算できる．) ワーカーによって与えられたラベルの代わりにこの確率値を “柔らかな” ラベルとして用い，その正解率（期待正解数）をもってワーカーの能力とするというのが彼らのアイディアである．

別の考え方が Raykar と Yu によって提案されている [Raykar 11]．スパムワーカーはできるだけ作業を行うことなしに報酬を得ようとするため，結果として全て答えを 1（あるいは 0）などとしたり，あるいはランダムに回答を行うことになる．これはつまりタスクの内容（真の答え）とは関わりなく独立にワーカーの回答が決定されていると考えることができる．これを確率的に考えれば，真の答えが 0 であっても 1 であってもワーカーが 1 と回答する確率（0 と回答する確率）は変わらない，つまり

$$\Pr[y_i^j = 1 \mid y_i = 1] = \Pr[y_i^j = 1 \mid y_i = 0]$$

であるということである．これはすなわち $\alpha^j = 1 - \beta^j$，つまり $\alpha^j + \beta^j = 1$ であり，彼らはワーカーがスパムワーカーである可能性を $\alpha^j + \beta^j$ の値が 1 に近いかどうかで判定することを提案している*8．

4. 群衆からの学習

4·1 群衆からの学習: モデルの直接推定

3 章で紹介した方法は，各々の単位タスクに対する正しい答えを得ることが目的であった．しかし，クラウドソーシングサービス利用の目的が教師付き学習の訓練データを得ることである場合には，将来訪れるであろう単位タスクに対して正しい答えを出力する予測モデルを得ることが本来の目的のはずである．もちろん，まず 3 章で紹介した方法によって正しい答えを推定した後にこれを用いてモデルを推定することも間違いではないが，我々の本来の目的は予測モデルを得ることであるから，これを直接的に実現する方が望ましいと考えることもできる．

教師付き学習の通常の問題設定では，訓練データとして入力の特徴ベクトル $\boldsymbol{x}_i \in \Re^D$（$D$ 次元の実数値ベクトル）とそれに対する正しい出力ラベル $y_i \in \{0,1\}$ の組が N 個与えられ，これをもとに入力から出力を予測するようなモデル $f : \Re^D \to \{0,1\}$ を推定する．一方クラウドソーシングを用いて訓練データを収集した場合には，i 番目の単位タスクに対する正しい出力ラベル y_i は未知のままであり，かわりに複数のワーカーによって（正しくないかもしれない）ラベルが与えられる．i 番目の単位タスクに対して回答を行ったワーカーの集合を $\mathcal{J}_i \subseteq \{1,2,\ldots,J\}$（$J$ はワーカーの人数）とすると，i 番目の単位タスクにはラベル集合 $\{y_i^j\}_{j \in \mathcal{J}_i}$ が与えられることになる．

*7 この $\boldsymbol{x}_i$ はモデルパラメータの一部であり，教師付き学習の入力のように予め与えらえるものではないことに注意する．

*8 なお，彼らの論文中では多値ラベルの場合やランキングの場合も考察されている．

クラウドソーシングを用いて収集された教師データを用いてモデルを推定するための最も素朴な方法としては，各々のラベルがどのワーカーによって与えられたかは考慮せずに，単純に全てのデータを機械学習アルゴリズムに与えてモデルを推定することが考えられる．例えば Sheng らは同じデータに対して繰り返しラベル付けを行うことで，得られる識別器の精度が向上することを示した [Sheng 08]．しかし 3 章でも述べたようにワーカーの能力の差を考慮し，これらを適切に重みづけてモデルを推定する方法が望ましいであろう．Raykar らは前述の Dawid と Skene の枠組みを教師付き学習に拡張し，真のラベルと予測モデルを同時に推定する方法を提案した [Raykar 09, Raykar 10]．推定すべき予測モデルはパラメータ $\boldsymbol{w}$ をもつロジスティック回帰の形で

$$\Pr[y_i = 1|\boldsymbol{x}_i] = \frac{1}{1+\exp(-\boldsymbol{w}^\top \boldsymbol{x}_i)} \tag{1}$$

のように書かれるものとする．彼らは，各ワーカーがラベルを付ける際には，このモデルによって生成された（実際には観測されない）真のラベルに対して，各ワーカーの持つパラメータ α^j と β^j（意味は前述の Dawid と Skene のモデルと同じ）を用いて摂動を加えることで，ワーカーの回答が確率的に生成されるものとした．Raykar らは EM アルゴリズムを用いてモデルを推定する方法を提案している．

Yan らはワーカーの誤り率が単位タスクに依存するモデルに拡張を行っている [Yan 10]．彼らのモデルではワーカーが真のラベルと同じラベルを回答する確率 α^j（および β^j）自身が $\boldsymbol{x}_i$ に依存して決まるようなモデルを提案した．

$$\alpha^j(\boldsymbol{x}_i) = \frac{1}{1+\exp(-\boldsymbol{w}^{j\top} \boldsymbol{x}_i)} \tag{2}$$

ここでは $\boldsymbol{x}_i$ は単位タスクの特徴を明示的に表現したものであるという違いはあるが，前述の Welinder らのモデル [Welinder 10] と同様に，単位タスクの特徴を表す $\boldsymbol{x}_i$ とワーカーの特徴を表す $\boldsymbol{w}^j$ との関係によって回答が決定されるモデルとなっている．

さて，これまでに紹介した多くのモデルでは真のラベルを潜在変数として導入し，潜在変数とワーカーのモデルを同時に推定する方式をとっている．通常ではこれらの定式化では最適化問題の目的関数が凸とならないため，EM アルゴリズムを用いて潜在変数の推定とモデルの推定を交互に行うのが標準的である．一方で Kajino らは真のラベルを潜在変数とする代わりに，各ワーカーがそれぞれ予測モデル (1) と同様の形式をもつラベル付けモデルを有するような形のモデル化を提案している [Kajino 12a]．彼らは予測モデルと個々のワーカーのモデルの関係づけとして，予測モデルのパラメータ $\boldsymbol{w}$ に各ワーカーの個性を表す部分 $\boldsymbol{v}^j$ が加わることで各ワーカーのモデルパラメータ $\boldsymbol{w}^j$ が表現されるとした．

$$\boldsymbol{w}^j = \boldsymbol{w} + \boldsymbol{v}^j$$

これはちょうどマルチタスク学習と呼ばれる複数の関連した学習問題を同時に解くような状況に対するモデル [Evgeniou 04] と同様の構造を持っている．(Evgeniou と Pontil のモデルでは全学習問題の共通部分が $\boldsymbol{w}$ として，個々の学習問題の個性にあたる部分が $\boldsymbol{v}^j$ として表現されている．）潜在変数を用いた前述のモデル群と比較して，Kajino らの定式化では潜在変数としての真のラベルの推定が問題に含まれず，個々のワーカーの与えたラベルのみを用いてモデル推定を行うため，問題を凸最適化問題として定式化することができる．そのため最適化アルゴリズムが局所解に陥ることがなく，結果として頑強なモデル推定を行うことができるのが特長である．

4·2 発展: 能動学習，逐次学習，正解データの利用

クラウドソーシングサービスを用いてモデル推定を行う際には当然のことながら取得するデータの量に応じて金銭的・時間的なコストがかかる．従って一度にまとめて大量のデータを取得するのではなく，必要に応じてクラウドソーシングサービスを呼び出し，できるだけ能力の高いワーカーから必要なデータを取得するような枠組みの方がより効率的かつ経済的であろう．機械学習分野においては，このような問題設定はデータが逐次的に到来するオンライン学習や，次にラベルを付与すべきデータを選択する能動学習として研究されてきたが，クラウドソーシング環境においても同様の試みが行われている．

Donmez らは繰り返し作業を発注するような状況において，能力の高いワーカーをなるべく早く発見するために，各ワーカーの正解率とその信頼区間を適当に見積もってから上位の何パーセントかに対してラベルづけを依頼するという戦略を提案している [Donmez 09]．信頼区間の見積もりに際し実際の正解は得られないため，現時点で信頼度が高いと思っているワーカーから得られた回答の中で多数決をとり，これを正解と見做すことで各ワーカーの正解率と信頼区間を更新するということを行っている．Zheng らはこのモデルをワーカーによってコストが異なるような状況に拡張している [Zheng 10]．また，Donmez らは時間と共にワーカーの能力が変化する設定においても考察を行っている [Donmez 10]．

一方 Yan らは 4 章で紹介した彼らのモデルを能動学習に拡張している [Yan 11]．クラウドソーシングを用いる場合，次に作業を依頼すべき単位タスクの選択と同時にどのワーカーに依頼するかの選択も必要となってくる．彼らは現在の予測モデルにおいて最もラベルが不確定性をもつような単位タスクを選択し，この単位タスクに対して正解率モデル (2) によって最も高い正解率が期待されるワーカーに依頼を行う方法を提案した．

ところで，ここまでに紹介した研究では真実のラベルは未知であることを仮定してきた．これはクラウドソーシング環境におけるその特徴を明確に表したものではあるが，実際には真実に近い情報がある程度は分かっている場合のほうが多いであろう．2･3 節では正解の分かっている問いをタスクの中に潜ませることによってワーカーの能力を判断する仕組みがいくつかのクラウドソーシングサービスにおいては実際に利用されていると述べたが，Tang らはこの状況を前述の Dawid と Skene のモデルを拡張することで統計的に取り扱う方法を提案した [Tang 11]．一方，予測モデルの直接推定の文脈においても同様の拡張が Kajino らによってなされている [Kajino 12b]．

5. ヒューマンコンピュテーション — 人間と機械による協調問題解決

AMT のようにクラウドソーシングサービスの中にはその機能を API として提供しているものがあるが，これはつまり計算機プログラム中からクラウドソーシングサービスを必要に応じて呼び出すことが可能であることを意味している．このことは人間を計算資源の一部として用いるという考えに一般化できる．ヒューマンコンピュテーションとは，計算資源としての人間の労働力を明確に意識しコンピュータと人間がそれぞれの得意領域を生かし一方のみでは解決できないような複雑な問題解決を行うという考え方である [Law 11]．

ヒューマンコンピュテーションの初期の試みは von Ahn による ESP ゲームに遡る [Von Ahn 06]．ESP ゲームは二人の地理的離れたプレイヤーによるインターネットを利用した協力ゲームであり，同一の画像に対して二人のプレイヤーがその画像にふさわしいと思うキーワードを独立に与え，これが一致したときに得点が得られる．これは人間による画像へのタグ付け作業をゲームの形で実現したものであり，このような不特定多数のプレイヤーに対して，ゲームの形式を持ちながら何らかの作業を暗黙的に行わせるゲームは「目的をもったゲーム（GWAP; Game With A Purpose）」とも呼ばれる．画像に対するタグ収集，ひいては収集したタグを利用した画像検索システムを全体としてある種の計算であると捉えることで，ESP ゲームは人間を計算資源の一部として用いる計算機プログラムと考えることができる．

ヒューマンコンピュテーションアルゴリズムはそのフローの中で人間に繰り返し作業を依頼することによって計算を行うことになるが，その具体的手段としては ESP ゲームのようにゲームの形によってこれを実現したり，あるいはクラウドソーシングサービスを（例えば AMT の API を通じて）用いることになる．例えば，Little らによって開発された TurKit は，AMT を用いたヒューマンコンピュテーションのプログラミングモデルである [Little 10]．TurKit の中では AMT のタスク呼び出しをあたかも通常の関数のように扱うことができる．例えば絵画をその芸術性が高い順に並び替えたいとしよう．一人の人間には多数の絵画について並べ替えを行うのは困難である．一方で機械には大量のデータを扱うことはできるが，絵画の芸術性を判断させるのは難しい．そこで，クイックソートのアルゴリズムにおける 2 枚の絵画の比較の部分を AMT を通じて人間に任せるというヒューマンコンピュテーションアルゴリズムがこの問題を解決する．

能力やモチベーションの違いなどに由来する人間の作業の質の問題は前の章でも中心的な課題であったが，ヒューマンコンピュテーションにおいてもやはり同様の問題が起こる．たとえば前述の絵画の比較といった単位タスクにおいてもワーカー毎のばらつきは出てくるであろう．人間の動作は確率的であるのでプログラムのフロー制御もまた人間の曖昧性に応じて確率的に行う必要がある．一つ一つの決定の結果というよりも逐次的な決定がもたらす最終的なプログラムの実行結果をもたらすことを考慮し，これを（部分観測）マルコフ決定過程としてモデル化したのが TurKontrol である [Dai 10, Dai 11]．彼らは文章入力とその記述の改善，改善ループの停止判断をタスクとして用いて TurKontrol による適切な分岐決定や終了判定によって全体のコストを下げられることを示した．TurKontrol はフローの骨子は決まっており，分岐においてどちらへ進むか，ループを何度回すかといった決定を行っていたのに対し，フローの設計自体もある程度人間に任せてしまおうというのが Turkomatic であり，各ワーカーは与えられた問題に対してこれを「自分で解く」か「分割して別のワーカーに依頼するか」を決定することができる [Kulkarni 11]．

6. おわりに

近年，クラウドソーシングを用いることで機械学習システムのボトルネックであるデータ取得（とくに教師付き学習における教師データ取得）を極めて低コストで行うことが可能になってきているが，集めたデータを全てそのままで，もしくは多数決等の素朴な方法によって前処理を行ってから用いる方法よりも，ワーカーの能力や特性を考慮した統計モデルを用いて真実を推定する方法のほうが優れていることが分かってきた．さらに最近ではより直接的にデータから予測モデルを直接推定するといった方法が模索されている．現在のところこれらの試みの多くが対象としているのは分類問題や回帰などの単純な予測タスクであるが，今後は様々なタイプの問題に対して同様のアプローチが提案されていくことであろう．たとえば 3･2 節で紹介した Welinder らのモデル [Welinder 10] は Gomes らによってクラスタリング問題に拡張されている [Gomes 11]．彼らの問題設定では，各ワーカーは 2 つのクラスタリング対象が同じクラスタに属するか否かという質問に答える．その回答は 2 つの対象の表現 $\boldsymbol{x}_i$

と $\boldsymbol{x}_k$ およびワーカーのもつ（行列）パラメータ $\boldsymbol{W}^j$ を用いて $\boldsymbol{x}_i^\top \boldsymbol{W}^j \boldsymbol{x}_k > \tau^j$（$\tau^j$ は決定の閾値）の符号によって決定される．

また，これまでに提案されている手法では，許容されるデータの形式は選択肢等の比較的単純な定型データに限定されているが，今後はこれらに留まることなく，より複雑で非定型なデータに対しても適用可能な方法が発展していくことが予測される [BakIr 07]．自然言語で書かれた自由回答，あるいは画像や音声といった対象に対してもこれらをうまく扱うことのできるができればその適用範囲は大きく広がるであろう．

一方で，クラウドソーシングを機械学習のために用いるのとは対称的に，クラウドソーシングをより効果的に機能させるための機械学習技術もまた重要になっていくであろう．ワーカーのスキルの種類や能力，あるいはモチベーション，地理的条件などを勘案し適切なタスクを適切な人に対して割り当てる仕組みはクラウドソーシングの効率化を図る上で極めて有効であろう．例えば予測モデルを用いてタスクの説明文をもとに各ワーカーに合わせて取り組むべきタスクを推薦するという試みも既に始まっている [Ambati 11]．

5 章で紹介したヒューマンコンピュテーションはクラウドソーシングをある意味で包括する枠組みともいえるが，ここにおいてもやはり人間にまつわる不確実性は重要なファクターとして捉えられている．ヒューマンコンピュテーションの試みはまだその初期の段階にはあるものの，ここ数年で急激な広がりを見せており*9，今後の発展が大きく期待される分野である．人間と機械がそれぞれの得意な領域を生かし，どちらか一方だけではこれまで解くことのできなかった重要で大きな問題を解決をするためには，人間の不確定性を捉えこれを制御するための確率・統計的モデリングや機械学習，そして逐次的な意思決定やプランニングなど複雑な意思決定のための最適化や制御といった様々な技術がヒューマンコンピュテーションの発展のカギを握っているといえる．

謝　　辞

本稿を執筆するに当たり有用なご意見を頂いた東京大学の佐藤一誠氏，日本アイ・ビー・エム東京基礎研究所の坪井祐太氏，北海道大学の小山聡氏に感謝する．

◇ 参 考 文 献 ◇

[Ambati 11] Ambati, V., Vogel, S., and Carbonell, J.: Towards Task Recommendation in Micro-Task Markets, in *Proceedings of the 3rd Human Computation Workshop (HCOMP)* (2011)

[BakIr 07] BakIr, G., Hofmann, T., Schölkopf, B., Smola, A., and Taskar, B.: *Predicting structured data*, The MIT Press (2007)

[Callison-Burch 10] Callison-Burch, C. and Dredze, M.: Creating Speech and Language Data With Amazon 's Mechanical Turk, in *Proceedings of Workshop of Creating Speech and Language Data with Amazon's Mechanical Turk at NAACL HLT 2010* (2010)

[Dai 10] Dai, P., Mausam, , and Weld, D. S.: Decision-theoretic Control of Crowd-sourced Workflows, in *Proceedings of the 24th Conference on Artificial Intelligence (AAAI)* (2010)

[Dai 11] Dai, P., Mausam, , and Weld, D. S.: Artificial Intelligence for Artificial Artificial Intelligence, in *Proceedings of the 25th Conference on Artificial Intelligence (AAAI)* (2011)

[Dawid 79] Dawid, A. P. and Skene, A. M.: Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statics)*, Vol. 28, No. 1, pp. 20–28 (1979)

[Dekel 09] Dekel, O. and Shamir, O.: Vox Populi: Collecting High-Quality Labels from a Crowd, in *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)* (2009)

[Donmez 09] Donmez, P., Carbonell, J. G., and Schneider, J.: Efficiently Learning the Accuracy of Labeling Sources for Selective Sampling, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (2009)

[Donmez 10] Donmez, P., Carbonell, J., and Schneider, J.: A Probabilistic Framework to Learn from Multiple Annotators with Time-Varying Accuracy, in *Proceedings of the SIAM International Conference on Data Mining (SDM)* (2010)

[Evgeniou 04] Evgeniou, T. and Pontil, M.: Regularized multi–task learning, in *Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)* (2004)

[Gomes 11] Gomes, R., Welinder, P., Krause, A., and Perona, P.: Crowdclustering, in *Advances in Neural Information Processing 24* (2011)

[Howe 06] Howe, J.: The Rise of Crowdsourcing, *Wired Magazine* (2006)

[Ipeirotis 10a] Ipeirotis, P. G.: Demographics of Mechanical Turk, Technical Report CeDER-10-01, NYU Center for Digital Economy Research Working Paper (2010)

[Ipeirotis 10b] Ipeirotis, P. G., Provost, F., and Wang, J.: Quality Management on Amazon Mechanical Turk, in *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP)* (2010)

[Kajino 12a] Kajino, H., Tsuboi, Y., and Kashima, H.: A Convex Formulation for Learning from Crowds, in *Proceedings of the 26th Conference on Artificial Intelligence (AAAI)* (2012)

[Kajino 12b] Kajino, H., Tsuboi, Y., Sato, I., and Kashima, H.: Learning from Crowds and Experts, in *Proceedings of the 4th Human Computation Workshop (HCOMP)* (2012)

[Kulkarni 11] Kulkarni, A., Can, M., and Hartmann, B.: Turkomatic: Automatic Recursive Task and Workflow Design for Mechanical Turk, in *Proceedings of the 25th Conference on Artificial Intelligence (AAAI)* (2011)

[Law 11] Law, E. and Von Ahn, L.: *Human Compuation*, Morgan & Claypool Publishers (2011)

[Little 10] Little, G., Chilton, L., Goldman, M., and Miller, R.: Turkit: human computation algorithms on mechanical turk, in *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology (UIST)* (2010)

[Raykar 09] Raykar, V. C., Yu, S., Zhao, L. H., Jerebko, A., Florin, C., Valadez, G. H., Bogoni, L., and Moy, L.: Supervised Learning from Multiple Experts: Whom to Trust

*9 例えばヒューマンコンピュテーションの話題に特化した国際ワークショップである Human Computation Workshop (HCOMP) は第 5 回を迎える 2013 年から国際会議に「格上げ」される．

When Everyone Lies a Bit, in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, ACM (2009)
[Raykar 10] Raykar, V. C., Yu, S., Zhao, L. H., Florin, C., Bogoni, L., and Moy, L.: Learning From Crowds, *Journal of Machine Learning Research*, Vol. 11, pp. 1297–1322 (2010)
[Raykar 11] Raykar, V. C. and Yu, S.: Ranking annotators for crowdsourced labeling tasks, in *Advances in Neural Information Processing 24* (2011)
[Sheng 08] Sheng, V. S., Provost, F., and Ipeirotis, P. G.: Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers, in *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (2008)
[Smyth 95] Smyth, P., Fayyad, U., Burl, M., Perona, P., and Baldi, P.: Inferring Ground Truth from Subjective Labelling of Venus Images, in *Advances in Neural Information Processing Systems 7* (1995)
[Snow 08] Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y.: Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2008)
[Sorokin 08] Sorokin, A. and Forsyth, D.: Utility data annotation with Amazon Mechanical Turk, in *Proceedings of the 1st IEEE Workshop on Internet Vision at CVPR 2008* (2008)
[Tang 11] Tang, W. and Lease, M.: Semi-Supervised Consensus Labeling for Crowdsourcing, in *ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR)* (2011)
[Von Ahn 06] Von Ahn, L.: Games with a Purpose, *Computer*, Vol. 39, No. 6, pp. 92–94 (2006)
[Welinder 10] Welinder, P., Branson, S., Belongie, S., and Perona, P.: The Multidimensional Wisdom of Crowds, in *Advances in Neural Information Processing Systems 23* (2010)
[Whitehill 09] Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., and Movellan, J.: Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise, in *Advances in Neural Information Processing Systems 22* (2009)
[Yan 10] Yan, Y., Rosales, R., Fung, G., Schmidt, M., Hermosillo, G., Bogoni, L., Moy, L., Dy, J., and Malvern, P.: Modeling Annotator Expertise: Learning When Everybody Knows a Bit of Something, in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)* (2010)
[Yan 11] Yan, Y., Rosales, R., Fung, G., and Dy, J. G.: Active Learning from Crowds, in *Proceedings of the 28th International Conference on Machine Learning (ICML)* (2011)
[Zheng 10] Zheng, Y., Scott, S., and Deng, K.: Active Learning from Multiple Noisy Labelers with Varied Costs, in *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM)* (2010)

〔担当委員：××○○〕

19YY 年 MM 月 DD 日　受理

著　者　紹　介

鹿島　久嗣（正会員）

1999 年京都大学大学院工学研究科応用システム修士課程修了．2007 年京都大学大学院情報学研究科知能情報博士課程修了．1999 年から 2009 年まで IBM 東京基礎研究所勤務．2009 年より東京大学大学院情報理工学系研究科数理情報学専攻准教授．機械学習，データマイニングの研究に従事．博士（情報学）．

梶野　洸

2011 年東京大学工学部計数工学科卒業．現在，同大学大学院情報理工学系研究科に在学中．クラウドソーシングを用いた機械学習に興味をもつ．