2M4-OS-07a-4

クラウドソーシングによって得られた順序データの統合

Crowdordering

松井 都志子[*1] 馬場 雪乃[*1] 神嶌 敏弘[*2] 鹿島 久嗣[*1*3]
Toshiko Matsui Yukino Baba Toshihiro Kamishima Hisashi Kashima

[*1]東京大学 [*2]産業技術総合研究所
The University of Tokyo AIST

[*3]JST さきがけ「知の創生と情報社会」研究領域
JST PRESTO

Crowdsourcing is a promising solution to problems that are difficult for computers, but relatively easy for humans. One of the biggest challenges in crowdsourcing is *quality control*, which is to expect high quality results from crowd workers who are not necessarily very capable nor motivated. There have been proposed several statistical crowdsourcing quality control methods for binary and multinomial questions by collecting multiple responses for each question and aggregating them to obtain accurate answers. In this paper, we extend the existing framework to address ordering questions, and give a new methods for obtaining reliable orderings from data collected by using crowdsourcing.

1. はじめに

不特定多数の人間に対して業務を依頼するクラウドソーシングは，Amazon Mechanical Turk[*1]をはじめとする各種プラットフォームの登場などを背景として，急速に利用が拡大している．クラウドソーシングを用いることによって，音声の書き起こしや，文章の執筆，翻訳，プログラム開発，デザインなど様々な業務の遂行を比較的低コストで手軽に外注することができる．計算機科学分野においてもデータ収集や画像の注釈づけ，Web コンテンツの分類をはじめとする計算機には困難なタスクを安価かつ大量に処理できるクラウドソーシングは大きな注目を浴びており，自然言語処理やコンピュータビジョン，HCI など様々な分野で盛んに利用されている．

クラウドソーシングにおける最大の課題の一つは成果物の品質を保証である．全てのワーカがタスクの遂行に十分な能力を持っていることが保証されないことに加え，手軽に報酬を得ることを目的として低品質の成果物を大量に納める不誠実なワーカも少なからず存在するため，クラウドソーシングの「品質管理」はその信頼性に関わる問題である．多くの商用プラットフォームでは依頼者が成果物を確認して低品質のものには報酬の支払いを行わない等のオプションが用意されてはいるものの，大量の作業結果すべてに対してチェックを行うのは容易ではない．

品質管理のアプローチとしては，正解の分かっているタスクをワーカ性能のベンチマークとして用いる方法がある．この方式はいくつかの商用プラットフォーム上で実用化されているものの，正解の準備コストがかかることや，そもそも正解を一つに決めることのできないタスクも数多く存在するため，その適用範囲は限定的である．一方で，いわゆる「冗長化」によって品質向上を行う仕組みは，正解を必要としないためより適用範囲が広い．これは一つのタスクを複数のワーカに割り当て，得られた複数の回答に対して多数決等の統計的手法を適用することによって統合し最終的な答えを得ようという考え方である [Sheng 08]．より高度な統計的手法としては各ワーカの能力やタスクの難易度といった特性を取り入れた確率モデルを利用したものが提案されている [Dawid 79, Whitehill 09, Welinder 10]．

現在提案されている統計的品質管理手法の対象となるタスクは，はい／いいえで答えられるようなものや，択一式の選択を行うようなものがほとんどであり，これをより一般的なタスクへと拡張する試みがなされている（例えば [Lin 12]）．本論文ではこの流れに従い，整序問題，すなわち与えられた項目の並べ替えを行うタスクへの拡張を行う．順序データの確率生成モデル [Marden 95]，その中でも特に距離ベースのモデルにおいて Spearman 距離を採用したもの [Mallows 57] に対してワーカ毎に異なる能力パラメータを導入することによって，クラウドソーシングにおける順序統合に適用できるようにする．距離として Spearman 距離を採用することで効率のよい推定アルゴリズムを導くことができる．なお，本研究と同様に整序問題を扱った研究としては Chen らによるもの [Chen 13] があるが，彼らはただ一つの順位付けタスクを対象としているのに対し，我々は複数の順序付けタスクを一度に並び替えることを目的としており用途が異なることを強調しておく．また，我々が実験で扱ったような整序問題では，各項目の順序の決定には全体の中での項目の位置が重要であるため，彼らが用いているような 2 項目の比較モデルは適さない．

文書整序と英文整序の 2 種類の問題に対してクラウドソーシングを用いて収集した実データを用いた実験では，提案手法はワーカの能力差を仮定しないモデルと比較して高い精度の回答が得られることを確認した．

2. 順序統合問題

はじめに本論文で扱う順序統合問題を定義する．順序統合問題は複数の人（ワーカと呼ぶ）に並べ替えタスクの回答を依頼した結果を統合し，正しい答えを導く問題である．個々の並べ替えタスクは，与えられた項目を正しい順序に並び替える問題であり，検索エンジンの検索結果の並び替えや，作業項目の順序付けの問題などがこれにあたる．

I 個の並べ替えタスクがあり，そのうちの i 番目のタスク

連絡先: 鹿島 久嗣，東京大学大学院情報理工学系研究科，kashima@mist.i.u-tokyo.ac.jp

*1 https://www.mturk.com/mturk/welcome

は M_i 個の項目を持つものとする．真の順序を順位ベクトル $\boldsymbol{\pi}_i = (\pi_{i,1}, \pi_{i,2}, \ldots, \pi_{i,M_i})$ として表現する．ここで $\pi_{i,j}$ は i 番目のタスクにおいて項目を正しい順に並び替えたときに，j 番目の項目に与えられる順序とする．すなわち $\boldsymbol{\pi}_i$ は $(1, 2, \ldots, M_i)$ の置換となる．

各タスク i の M_i 個の項目を並び替えるためにクラウドソーシングサービス上でタスク実行の依頼を行う．総勢で K 人のワーカが取り組んだものとし，そのうち k 番目のワーカが取り組んだタスクのインデクス集合を $\mathcal{I}^{(k)}$ とする．一方，i 番目のタスクに取り組んだワーカのインデクス集合を $\mathcal{J}_i$ とする．また，i 番目のタスクに対して k 番目のワーカが与えた順位ベクトルを $\boldsymbol{\pi}_i^{(k)} = (\pi_{i,1}^{(k)}, \pi_{i,2}^{(k)}, \ldots, \pi_{i,M_i}^{(k)})$ とする．

クラウドソーシングを用いて収集された順序データ $\{\boldsymbol{\pi}_i^{(k)}\}_{k\in\{1,2,\ldots,K\}, i\in\mathcal{I}^{(k)}}$ から真の順序 $\{\boldsymbol{\pi}_i\}_{i\in\{1,2,\ldots,I\}}$ を推定することが我々の目的である．

3. クラウドソーシングによる順序生成モデル

クラウドソーシングを用いた順序統合問題を解くために，我々はクラウドソーシングワーカによる回答の生成過程を確率モデル化し，これをデータから統計的に推定することにより真の答えを得るというアプローチをとる．

3.1 距離ベース順序生成モデル

クラウドソーシングワーカによる回答の生成過程をモデル化するにあたり，まずは順序の確率的なモデルが必要となる．順序の確率モデルには様々あるが，中でも我々は距離ベースのモデルに注目する [Marden 95]．距離ベースのモデルでは，モード順序 $\boldsymbol{\pi}$ およびモードへの集中度 λ をパラメータとして，順序 $\tilde{\boldsymbol{\pi}}$ が生成される確率を以下の条件付き確率によって定義する：

$$\Pr[\tilde{\boldsymbol{\pi}} \mid \boldsymbol{\pi}, \lambda] = \frac{1}{Z(\lambda)} \exp\left(-\lambda d(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi})\right)$$

ここで，$d(\cdot,\cdot)$ は 2 つの順位ベクトルの間の距離，$Z(\lambda)$ は以下で定義される正規化定数である：

$$Z(\lambda) = \sum_{\tilde{\boldsymbol{\pi}}} \exp\left(-\lambda d(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi})\right)$$

距離 $d(\cdot,\cdot)$ の定義には様々あるが，パラメータ推定の容易さからユークリッド距離（順序モデルの文脈では特に Spearman 距離と呼ばれる）を用いる．このモデルは特に Mallows θ モデルと呼ばれる [Mallows 57].

3.2 クラウドソーシングにおける順序生成モデル

クラウドソーシングにおいては，ワーカ間の能力差等によってより正しい順序を与えるワーカとそうでないワーカが混在している．この属人的な個性を捉えるために，ワーカごとに異なる集中度パラメータ λ をもつ距離ベース順序生成モデルを用いることにする．つまり，ワーカ k が集中度パラメータ $\lambda^{(k)}$ を持つものとすることで，k 番目のワーカによる順序生成のモデルを以下のように定義する：

$$\Pr[\tilde{\boldsymbol{\pi}} \mid \boldsymbol{\pi}, \lambda^{(k)}] = \frac{1}{Z(\lambda^{(k)})} \exp\left(-\lambda^{(k)} d(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi})\right)$$

大きな $\lambda^{(k)}$ の値をもつワーカほど，モード順序 $\boldsymbol{\pi}$ である真の順序と距離の近い順序に与える確率が高くなるため，従って正解により近い回答を行うことができる．このことから $\lambda^{(k)}$ はワーカの能力値として解釈することができる．

4. 推定方法

クラウドソーシングを用いて収集された順序データから真の順序およびワーカの能力パラメータを推定するために最尤推定を行う．最尤推定は真の順序と能力パラメータを交互に最適化することで効率的に行うことができる．

4.1 目的関数

クラウドソーシングを用いて収集された順序データ $\{\boldsymbol{\pi}_i^{(k)}\}_{i,k}$ から真の順序 $\{\boldsymbol{\pi}_i\}_i$ およびワーカの能力パラメータ $\{\lambda^{(k)}\}_k$ を推定するために最尤推定を行う．解くべき最適化問題の目的関数は以下で与えられる対数尤度 J である：

$$\begin{aligned} &J(\{\lambda^{(k)}\}_k, \{\boldsymbol{\pi}_i\}_i) \\ &= \sum_k \sum_{i\in\mathcal{I}^{(k)}} \log \frac{1}{Z(\lambda^{(k)})} \exp\left(-\lambda^{(k)} d(\boldsymbol{\pi}_i^{(k)}, \boldsymbol{\pi}_i)\right) \\ &= -\sum_k \sum_{i\in\mathcal{I}^{(k)}} \left\{ \lambda^{(k)} d(\boldsymbol{\pi}_i^{(k)}, \boldsymbol{\pi}_i) + \log \sum_{\tilde{\boldsymbol{\pi}}} \exp\left(-\lambda^{(k)} d(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi}_i)\right) \right\} \end{aligned} \tag{1}$$

これを最大化することによって，$\{\lambda^{(k)}\}_k$ および $\{\boldsymbol{\pi}_i\}_i$ の最尤推定量を得る．

4.2 最適化アルゴリズム

式 (1) で与えられる目的関数 $J(\{\lambda^{(k)}\}_k, \{\boldsymbol{\pi}_i\}_i)$ を $\{\lambda^{(k)}\}_k$ と $\{\boldsymbol{\pi}_i\}_i$ について最大化するにあたり，両変数集合についての最適化を同時に行うことは困難である．そこで両者を交互に最適化するアプローチをとることにする．

なお目的関数 J は凸関数ではないため，得られる解は初期値に依存する．そこで全ワーカの能力が一定とした状態（すなわち任意の k と ℓ について $\lambda^{(k)} = \lambda^{(\ell)}$）から始めることにする．

4.2.1 真の順序 $\{\boldsymbol{\pi}_i\}_i$ についての最適化

全ワーカの能力 $\{\lambda^{(k)}\}_k$ を固定したとき，目的関数 (1) の第一項目についてのみ考えれば十分である．$\{\boldsymbol{\pi}_i\}_i$ についての最適化は離散最適化となるためこれを解くことは一般に困難であるが，距離として Spearman 距離を用いたときにはこれが容易になる．

i 番目のタスクに対する真の順位ベクトル $\boldsymbol{\pi}_i$ の推定は次のように行うことができる．まず各項目 $m(=1,\ldots,M_i)$ に対して，次で定義されるワーカの能力で重みづけられた重み付き順序 $w_{i,m}$ を求める：

$$w_{i,m} = \frac{1}{|\mathcal{J}_i|} \sum_{k\in\mathcal{J}_i} \lambda^{(k)} \pi_{i,m}^{(k)}$$

そして，$w_{i,1}, w_{i,2}, \ldots, w_{i,M_i}$ を小さいものから並べたときの順序が真の順序の最尤推定量となる．

4.2.2 ワーカの能力 $\lambda^{(k)}$ についての最適化

次に真の順序 $\{\boldsymbol{\pi}_i\}_i$ を固定した場合のワーカの能力 $\lambda^{(k)}$ についての最適化を考える．目的関数 (1) は以下で定義される各ワーカ k ごとの目的関数 $J^{(k)}$ の和に分解できる：

$$J^{(k)}(\lambda^{(k)}) = \lambda^{(k)} d(\boldsymbol{\pi}_i^{(k)}, \boldsymbol{\pi}_i) + \log \sum_{\tilde{\boldsymbol{\pi}}} \exp\left(-\lambda^{(k)} d(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi}_i)\right)$$

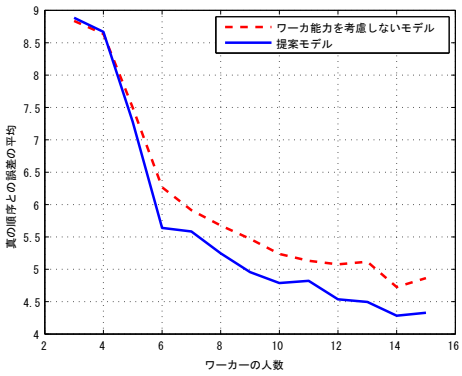
つまり $\lambda^{(k)}$ の最適化には一変数関数 $J^{(k)}(\lambda^{(k)})$ のみを考えればよいため，比較的容易に最適解が得られる．

5. 実験

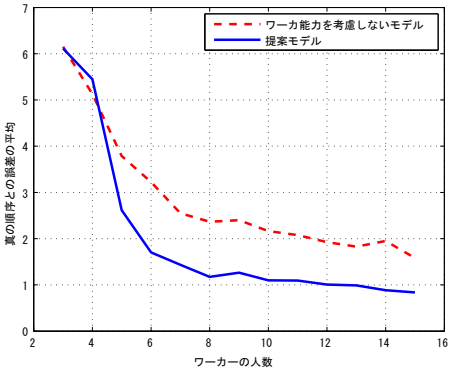
提案手法の有効性を検証するため，文書整序問題と英文整序問題の 2 種類の問題を題材に，実際にクラウドソーシングを用いて収集したデータを用いた検証を行った．データの収集にはクラウドソーシングサービス Lancers[*2] を用いた．

文書整序問題は，日本語の文が 5〜6 文与えられたものを，意味が通るように並び替えるというものである．15 人のワーカに対し 13 問の文書整序問題を解くことを依頼し，195 件の総回答数を得た．英文整序問題は与えられた英単語を正しい英語の文になるように並び替えるというものである．15 人のワーカに対し 20 問の文書整序問題を解くことを依頼し，300 件の総回答数を得た．いずれのデータセットも問題の正解が分かっているため，性能評価が可能である．

評価のベースラインとしてワーカの能力を考慮しないモデル，すなわち全てのワーカの能力が等しいとしたモデルを用いた．図 1 は，ワーカ数を変えた時（3〜15 人）の推定された順序と真の順序との Spearman 距離を示したものである．両方のデータセットにおいて，ワーカ数が 6〜7 人以上の場合には提案手法はベースライン手法よりも精度の高い推定が行えていることがわかる．これはワーカの能力差をモデルに導入したことが現実の状況をより正しく反映していることを示唆している．



(a) 文書整序



(b) 英文整序

図 1: 回答の精度評価：提案手法とベースライン手法（ワーカの能力を考慮しないモデル）によって統合した回答の精度比較． 異なるワーカ数における真の順序との Spearman 距離を示した．提案手法がベースライン手法よりも高い推定精度を示していることがわかる．

6. おわりに

本研究では，回答が順序データとなるタスクにおいて，クラウドソーシングを用いて回答を収集した場合の品質管理問題を考えた．順序の確率的生成モデルの中でも距離ベースモデルを基礎に，これをワーカーの能力を考慮したモデルに拡張することで，ワーカの能力差がある状況においてワーカの回答を統合してより精度の高い解が得られる方法を与えた．また，距離として Spearman 距離を用いることで効率の良い推定手法を提案した．実データを用いた実験により提案手法はベースライン手法であるワーカ能力を考慮しない方法と比較して高い品質の解を得られることを示した．

参考文献

[Chen 13] Chen, X., Bennett, P. N., Collins-Thompson, K., and Horvitz, E.: Pairwise Ranking Aggregation in a Crowdsourced Setting, in *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM)* (2013)

[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)

[Lin 12] Lin, C., Mausam, M., and Weld, D.: Crowdsourcing Control: Moving Beyond Multiple Choice, in *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)* (2012)

[Mallows 57] Mallows, C. L.: Non-Null Ranking Models. I, *Biometrika*, Vol. 44, pp. 114–130 (1957)

[Marden 95] Marden, J. I.: *Analyzing and Modeling Rank Data*, Vol. 64, CRC Press (1995)

[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)

[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)

*2 `http://lancers.jp`