THE UNIVERSITY OF TOKYO

群衆からの学習
～クラウドソーシング＋機械学習～

The Multidimensional Wisdom of Crowds. Welinder, P. et al., In NIPS, 2010 他

鹿島 久嗣

東京大学

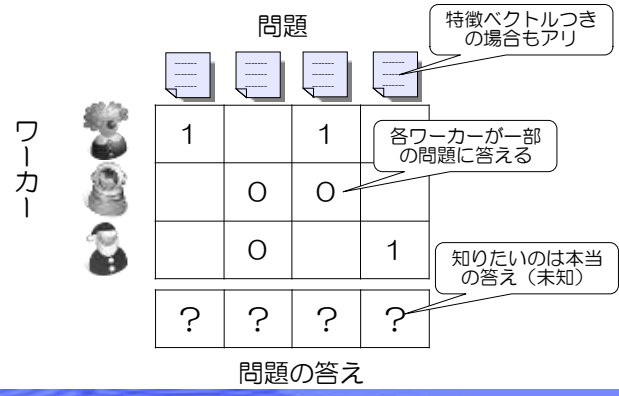
DEPARTMENT OF MATHEMATICAL INFORMATICS

サマリー：なんかここもうちょっとイケそうな感じ！

- 1970年代後半に複数の医者の診断を統合する文脈で始まった「専門家の意見の統合」問題
  - 単純な多数決よりも良い判断を下したい
- 2000年代に入ってからインターネット経由で安価な労働力を調達できるクラウドソーシング（特にAmazon MechanicalTurk）が教師つき学習のラベルづけを安価に行う方法として用いられている
  - 自然言語処理、画像処理における利用が盛ん
- 多数の素人から集められたラベルの信頼度を上げるための手段として、意見統合のワザが再注目されている
  - 「専門家の意見の統合」から「群衆の意見の統合」へ
- そして今、さらに「群衆の意見からの学習（複数の教師からの学習）」へと形をかえ、ひそかに盛り上がりつつある
  - 特徴ベクトルの導入、問題ごとの難易度の導入、オンライン化
- 表題の論文は、上ほどのブレークスルーではないが、問題ごとの難易度考慮を進めた感じ

典型的な問題設定： 各ワーカーがいくつかの問題に対して答えを提示 ⇒ 各問題の真の答えを当てたい

- 入力：各ワーカーがいくつかの問題に対し答える
- 出力：各問題の真の答え（決して観測されない）が知りたい
- 前提：多数決＝答えではない（選挙などと異なる）
- 協調フィルタリングに似ているが穴埋めが目的ではない



群衆からの学習の歴史 I（B.M.: MechanicalTurk出現以前）

- 1979 (B.M.26): 専門家（医者）の意見統合の文脈で議論される
  - Maximum Likelihood Estimation of Observer Error-rated using the EM Algorithm , Dawid, A.P., & Skene, A.M., Applied Statistics, 28(1), pp.20-28, 1979.
  - 専門家のラベルづけ能力をモデルに入れる
- 1999 (B.M.6): 自然言語処理への適用
  - Development and Use of a Gold-Standard Data Set for Subjectivity classifications, Wiebe, J.M., Bruce, R.F., & O'Hara, T.P., In ACL, 1999.
  - Dawid&Skene に良く似たモデルを文が主観的か客観的かを判定するタスクに応用
- 2004 (B.M.1): 画像（惑星の表面写真）処理への適用
  - Inferring Ground Truths from Subjective Labeling of Venus Image, Smyth, P., Fayyad, U., Burl, M., Perona, P., & Baldi, P., In NIPS, 2004.
  - Dawid&Skeneの方法を金星のレーダー画像から火山を見つけるタスクに応用

群衆からの学習の歴史 II （A.M.: MechanicalTurk出現以降）

- 2008(A.M.3): 自然言語処理におけるMTurk利用
  - Cheap and Fast—But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks., Snow, R. et al., In EMNLP, 2008.
  - テキストに対して感情を振る、単語が似ているかどうか、など５つのタスク
- 2008(A.M.3): コンピュータビジョンにおけるMTurk利用
  - Utility data annotation with Amazon Mechanical Turk, Sorokin, A. and Forsyth, D. In CVPR workshop on Internet Vision, 2008.
  - 領域分割や身体のマーカー付けなどの４タスク

群衆からの学習の歴史 III （A.M.: MechanicalTurk出現以降）

- 2009 (A.M.4): テクニカルなブレークスルー
  - 特徴ベクトルの利用：「汎化」可能に
    - Vox Populi: Collecting High-Quality Labels from a Crowd, Dekel, O. and Shamir, O. In COLT, 2009.
    - Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit., Raykar, V.C et al., In ICML, 2009 (Learning from Crowds, JMLR, 2010)
  - オンライン化：コストの概念
    - Efficiently Learning the Accuracy of Labeling Sources for Selective Sampling. Donmez, P. et al., In KDD, 2009.
  - インスタンスの難しさの導入
    - Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise., Whitehill, J. et al., In NIPS, 2009.
- 2010 (A.M.5)：The multi-dimensional wisdom of crowds ← 今ここ
  - インスタンスの、ワーカーに対する難しさを考慮

Mechanical Turkは、2005年にAmazonが開始したクラウドソーシングのプラットフォーム

- 世界中にいるワーカー（Turker）に簡単な作業を、安価で依頼できる
  - このWebサイトの感想をください
  - この画像に鳥は写っていますか
- 自然言語処理、コンピュータビジョンなどのアノテーションづけに盛んに利用されている
- 現在、（発注側は）US内のみ

http://ja.wikipedia.org/wiki/チェス

Dawid&Skeneの先駆的な研究：各ワーカーの能力をモデル化し、EMアルゴリズムで推定する

- この話題では、おそらくほぼ最古の先駆的研究
  - Maximum Likelihood Estimation of Observer Error-rated using the EM Algorithm , Dawid, A.P., Skene, A.M., Applied Statistics, 28(1), pp.20-28, 1979
- データ：45人の患者の麻酔適合度を5人の麻酔医が4段階で診断したもの
- 目的：真実の答え（麻酔適合度）を当てること
- 問題が１つしかないのであれば、多数決以上の結果は望みにくい。一方で、複数の問題があるなら、つねに多数派に入っているワーカーは信頼度が高そうだ
- モデル：
  - π_{jl}^{k}: kさんが真実の答えjの問いに対して答えが l と答える確率
  - 真実の答えを潜在変数としてEMアルゴリズムを適用
    - 結局、多数派に入っているワーカーが正解率が高いとするモデルが選ばれる

３つのテクニカルブレークスルー：特徴ベクトル、オンライン、インスタンスモデル

- 2009 (A.M.4)に３つのテクニカルなブレークスルーがあった
  - 特徴ベクトルの利用：「汎化」可能に
    - Vox Populi: Collecting High-Quality Labels from a Crowd, Dekel, O. and Shamir, O. In COLT, 2009.
    - Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit., Raykar, V.C et al., In ICML, 2009 (Learning from Crowds, JMLR, 2010)
  - オンライン化：コストの概念
    - Efficiently Learning the Accuracy of Labeling Sources for Selective Sampling. Donmez, P. et al., In KDD, 2009.
    - 2010には時間変化も考える
  - インスタンスの難しさの導入
    - Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise., Whitehill, J. et al., In NIPS, 2009.

ブレークスルー1：オンライン化によって（逐次的な）アウトソーシングのコストが意識されるようになった

- Efficiently Learning the Accuracy of Labeling Sources for Selective Sampling. Donmez, P. et al., In KDD, 2009.
- 繰り返し仕事を発注するような状況を想定する
- モチベーション：能力の高いワーカーを素早く見つけ、彼らに仕事をまかせたい（発注コスト／質を安くしたい）
- モデル：
  - 各ワーカーの正解率の信頼区間を適当に見積もり、上位の何パーセントかに対してラベルづけを依頼する
  - 彼らの答えのなかで最も多かったものを正解として、各ワーカーの正解率を更新する
- データ：
  - UCIなどのベンチマークを仮想的にクラウド化
  - Snowのデータのうちの2値分類タスク

補足：ちなみに、ワーカーの能力のモデルにも微妙な粒度の違いがある

- ちなみに、ワーカーの能力を
  - Dawid&Skene流に「真実の答え j の問いに対し l と答える確率」
  - Donmez et al.のように「正解率」

  とするのかは微妙な違いにみえる
- しかし：
  - 前者は細かいモデル化ができるが、複数の問題の答えのフォーマットが同一であることを要求
    - 1=病気、0=健康
  - 後者は複数の答えの値域が一致していなくてもよい
    - 3択問題「答えは次の内どれ？」
    - なんなら問いごとにタスクが異なってもよい

ブレークスルー2：特徴ベクトルの導入により「汎化」と「直接的な予測モデル構築」が可能になった

- Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit., Raykar, V.C et al., In ICML, 2009 (Learning from Crowds, JMLR, 2010)
- 問題に特徴ベクトルがついたことで汎化が可能になった
  - ワーカーがラベルづけしないデータに対しても予測が可能
    ← まさに機械学習
- 手法の面でも、最終的に解くべき問題をより直接的に解いている
  - 以前：機械学習を使い信頼できるラベルをつけた後、 そのラベルを用いて（ここで特徴ベクトル投入）予測モデルをつくる
  - 以降：（ラベルの信頼度を考慮に入れつつ）最初から予測モデルをつくる（目的を一気に果たす意味でVapnik流）
    - 事例の真のラベルよりも、最終的によい予測器をつくることに興味がある

手法としては、Dawid&Skeneの特徴ベクトル入り版

- モデル：
  - α^j: jさんが正クラスを正しく当てる確率
  - β^j: jさんが負クラスを正しく当てる確率
  - ロジスティック回帰（特徴ベクトルをベースにした全ラベラー共通のモデル）
- 推定：（一般化）EM
  - Dawid&Skeneの特徴ベクトルあり版
- データ：Siemensだけにいろいろ
  - マンモグラフィの画像が悪性かどうかを診断する
    - 正例：497、負例：1618、ground truthsあり
    - 27次元の特徴ベクトル
    - 仮想的に2人の技師が診断
  - 胸のMRI診断（75例、技師4人）
  - Snowらのデータ

補足：同じような話が同時期にもうひとつ。こちらは学習理論的な考察つき

- Vox Populi: Collecting High-Quality Labels from a Crowd, Dekel, O. and Shamir, O. In COLT, 2009.
- こちらも特徴ベクトルあり
- モデル：
  - 各人の誤り率を推定して、役に立たない人を除く
  - データを全部使ってつくった予測器（ここで特徴ベクトルが必須）をground truthとみなしてワーカーの誤り率を推定する
- 見どころ：学習理論的な解析
- データ：
  - クエリとURLのペアがマッチするか（MTurk使用）

ブレークスルー3：問題ごとの難しさを考える

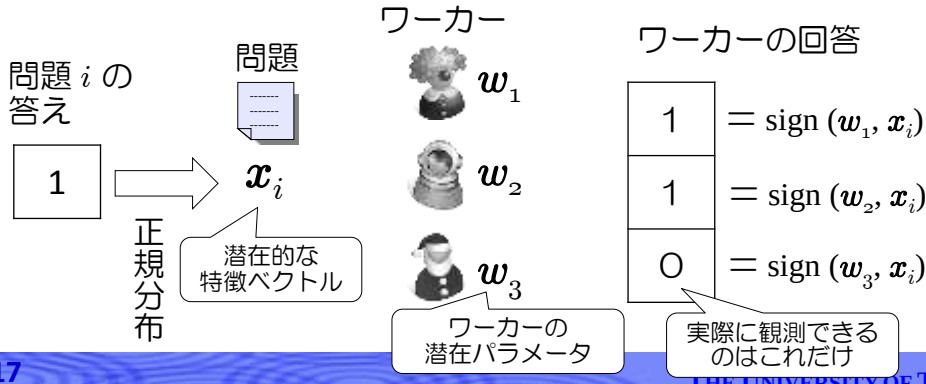
- Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise., Whitehill, J. et al., In NIPS, 2009.
- 各ワーカーの能力だけでなく、事例ごとの難しさを考慮する
- モデル：
  - $-\infty < \alpha_i < \infty$: i さんの能力 （$\alpha_i=\infty$で絶対正解、$\alpha_i=-\infty$で絶対不正解、$\alpha_i=0$でランダム予測）
  - $\beta_j \geqq 0$: 問 j の簡単さ（$1/\beta_j$が難しさ）
  - i さんが問jに正しく答える確率は、ロジスティック関数 $(1/\{1+\exp(-\alpha_i\beta_i)\})$できまる
  - EMで推定
- データ：
  - 架空の生物の雄雌を判定する（10人、100事例；MTurk）
  - 本気笑いと愛想笑いを判定する（20人、160事例；MTurk）

今回の論文：各ワーカーに対する問題の難しさが入った

- The Multidimensional Wisdom of Crowds. Welinder, P. et al., In NIPS, 2010.
- Whitehill, J. et al.のモデルを、もう少し細かく考えてみた（後述）
  - 真のラベルから（潜在）特徴ベクトルがつくられる
  - 各ワーカーは（潜在）線形識別器を持っている
- ワーカーと問題の相性（難しさ）を考慮できる
- 特徴ベクトルは無し
  - 潜在特徴ベクトルは使う ← ここが「multi-dimensional」？
  - 潜在変数はすべて勾配法で推定
- タスク：
  - 人工データ
  - ある鳥かどうか（40人、100事例；MTurk）

特徴ベクトルとワーカーの予測モデルを潜在変数とする

- 生成モデル：
  1. 真のラベルz_iを生成
  2. ラベルから（潜在的な）特徴ベクトル $\boldsymbol{x}_i$ を生成
  3. ワーカー jは自身の線形識別器sign{ $(\boldsymbol{w}_j, \boldsymbol{x}_i)$ }でラベルl_{ij} を生成
- ワーカー j にとって難しい例は特徴ベクトルが境界面に近い



感想：この論文は「着実な進歩」くらい？ 個人的には、2009年の3つの論文が面白い

- この論文自体の貢献がどのくらい大きいのかはよくわからない
  - （精度では勝っているけど）個人的には一年前のNIPSのインスタンスごとの難易度を導入したほうがエライと思う
- 歴史的には、特徴ベクトルの導入が、機械学習的な（Vapnik信者的な）意味で一番重要だと思う
  - 「群衆のための学習」から「群衆からの学習」へとシフト
- トピック自体の印象は「テクニックよりはモデル勝負」
- 他のトピックとの関連
  - マルチタスク学習：ワーカーごとに別々の予測モデルを学習
    ⇔ ひとつの予測モデルを学習
  - オンライン学習：正解がフィードバックされる
    ⇔ 最後まで真実はわからない

結論：イケる！まだイケるよココ！

- 某エライ先生「読むべき論文が５０以内だったらまだイケる」
  - ちなみにこのトピックはせいぜいまだ２０～３０本
- ちょっと考えただけでも色々できるよ！：
  - 特徴ベクトル＋事例難易度
  - マルチタスク
  - プライバシ保護
  - ワーカーのグループ考慮
  - 無限ラベル（タグ、自由記述、構造出力）
- ただし、データ入手が壁
  - MTurkはUS在住でないと使用できない

読んでいない文献

- Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers., Sheng, V.S et al., In KDD, 2008.
- Veritas: Combining Expert Opinions without Labeled Data, Cholleti, S.R. et al., In ICTAI, 2008.
- Learning with an Unreliable Teacher, Lugosi, G., Pattern Recognition, Vol. 25, No. l, pp.79-87, 1992.
- Online crowdsourcing: rating annotators and obtaining cost-effective labels. Welinder, P. and Perona, P., In IEEE Conference on Computer Vision and Pattern Recognition Workshops , 2010.
- Some objects are more equal than others: measuring and predicting importance. Spain, M. and Perona, P., In ECCV, 2008.
- Statistical Modeling of Expert Ratings on Medical Treatment Appropriateness., Uebersax, J.S., Journal of the American Statistical Association, Vol. 88, No. 422, 1993. （Dawid&Skeneとほぼ同じ）