Crowdsourcing

Markus Rokicki

L3S Research Center

29.05.2018
Example: Product Categorization

Choose the Best Category

Instructions:
For each item, read its title, product description and attributes.

1. If you think any listed category can describe the item correctly, click that category.
2. If you think none of the category can describe the item correctly, click "None of the Above".
3. If you think the provided information is not sufficient to make a decision, also click "None of the Above".

Figure: Product categorization task on Amazon Mechanical Turk

1 https://www.mturk.com
Example: Damage Assessment during Disasters

Crowdsourcing definition

Crowdsourcing, according to a literature survey\(^3\):

- Participative online activity
- an individual, institution, non-profit, or company proposes voluntary undertaking of a task
- in a flexible open call to a group of individuals
  - varying knowledge, heterogeneity, and number
- Users receive satisfaction of a given type of need
  - economic, social recognition, self-esteem, learning, …
- always entails mutual benefit

Crowdsourcing Platforms

Some crowdsourcing platforms:
- Amazon Mechanical Turk\(^4\)
  - Paid “microtasks”
- figure eight (formerly CrowdFlower)\(^5\)
  - Paid “microtasks”
- Topcoder\(^6\)
  - Programming competitions
- Threadless\(^7\)
  - Propose and vote on t-shirt designs
- Kickstarter\(^8\)
  - Fund products

\(^4\) https://www.mturk.com
\(^5\) https://www.figure-eight.com/
\(^6\) https://www.topcoder.com
\(^7\) https://www.threadless.com
\(^8\) https://www.kickstarter.com
Paid Microtask Crowdsourcing

Paid crowdsourcing scheme\textsuperscript{9}.

\textsuperscript{9}Figure source: http://dx.doi.org/10.1155/2014/135641
Paid Microtask Crowdsourcing

Paid crowdsourcing scheme\(^9\).

\(^9\)Figure source: http://dx.doi.org/10.1155/2014/135641
Paid Microtask Crowdsourcing

Paid crowdsourcing scheme\(^9\).

Figure source: http://dx.doi.org/10.1155/2014/135641
Task design

Depending on application:

- (how) should you break up the task?
- more or less answering options? Free-form answers?
- how to present the problem

E.g.: Highlighting keywords in search results¹⁰:

Task design issues: Cognitive Biases

Anchoring effects

▶ “Humans start with a first approximation (anchor) and then make adjustments to that number based on additional information.”

▶ Observed in crowdsourcing experiments

▶ Group A Q1: More or less than 65 African countries in UN?
▶ Group B Q1: More or less than 12 African countries in UN?
▶ Q2: How many countries in Africa?

12 Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. “Running experiments on Amazon Mechanical Turk”. In: Judgment and Decision Making 5.5 (2010).
13 Adapted from https://www.slideshare.net/ipeirotis/managing-crowdsourced-human-computation
Task design issues: Cognitive Biases

Anchoring effects

▶ “Humans start with a first approximation (anchor) and then make adjustments to that number based on additional information.”

▶ Observed in crowdsourcing experiments

▶ Group A Q1: More or less than 65 African countries in UN?
▶ Group B Q1: More or less than 12 African countries in UN?

▶ Q2: How many countries in Africa?
▶ Group A mean: 42.6
▶ Group B mean: 18.5

Also: Order effects and many more


12 Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. “Running experiments on Amazon Mechanical Turk”. In: Judgment and Decision Making 5.5 (2010).

13 Adapted from https://www.slideshare.net/ipeirotis/managing-crowdsourced-human-computation
Quality: Worker Access and Qualification

Crowdsourcing platforms offer means to restrict access to tasks based on:

▶ Trust levels based on overall past accuracy

14 https://www.figure-eight.com/crowdflower-now-offering-twelve-language-skill-groups
Quality: Worker Access and Qualification

Crowdsourcing platforms offer means to restrict access to tasks based on:

- Trust levels based on overall past accuracy

Figure: Individual Performance on IQ-Tests depending on past worker accuracy.


[14] https://www.figure-eight.com/crowdfower-now-offering-twelve-language-skill-groups
Quality: Worker Access and Qualification

Crowdsourcing platforms offer means to restrict access to tasks based on:

- Trust levels based on overall past accuracy
- Geography
- Language skills
  - Certified based on language tests
  - Classified by the platform based on geolocation, browser data and user history
- Qualification for specific kinds of tasks gained through qualification/training tasks

14 https://www.figure-eight.com/crowdfower-now-offering-twelve-language-skill-groups
Qualification Tasks

Qualification tests also provide feedback about the judgments expected by the worker and thus influence quality\textsuperscript{15}.

Figure on the right: Accuracy depending on class imbalance in training questions.

Classes: ‘Matching’, ‘Not Matching’

Measuring Worker Accuracy: Gold Standard

Would like to know how reliable answers are for the task
  ▶ Requesters may want to reject the work of unreliable workers
    ▶ In particular for ‘spammers’ who do not try to solve the task

Gold standard / honey-pot tasks:
  ▶ Add tasks for which the ground truth is already known at random positions
  ▶ Compare worker input with the ground truth
  ▶ Feedback on correctness is given to the workers
Using Gold Standard

Challenges:

- Need enough gold standard tasks or limit #tasks per worker
  - Workers who recognize gold standard will tend to ‘spam’
- Data composition:
  - Class balance
  - Educational examples addressing likely errors
- Need to be indistinguishable from regular tasks. But: need to be unambiguous

---

Using Gold Standard

Challenges:

▶ Need enough gold standard tasks or limit #tasks per worker
  ▶ Workers who recognize gold standard will tend to ‘spam’
▶ Data composition:
  ▶ Class balance
  ▶ Educational examples addressing likely errors
▶ Need to be indistinguishable from regular tasks. But: need to be unambiguous

Standard workflow:

▶ Iterate ground truth creation
  ▶ 1. Start with hand crafted small ground truth

---

Using Gold Standard

Challenges:

- Need enough gold standard tasks or limit #tasks per worker
  - Workers who recognize gold standard will tend to ‘spam’
- Data composition:
  - Class balance
  - Educational examples addressing likely errors
- Need to be indistinguishable from regular tasks. But: need to be unambiguous

Standard workflow:

- Iterate ground truth creation
  - 1. Start with hand crafted small ground truth
  - 2.1 Use feedback to refine/fix existing gold standard data

---

Using Gold Standard

Challenges:
- Need enough gold standard tasks or limit #tasks per worker
  - Workers who recognize gold standard will tend to ‘spam’
- Data composition:
  - Class balance
  - Educational examples addressing likely errors
- Need to be indistinguishable from regular tasks. But: need to be unambiguous

Standard workflow:
- Iterate ground truth creation
  - 1. Start with hand crafted small ground truth
  - 2.1 Use feedback to refine/fix existing gold standard data
  - 2.2 Use (redundant) annotations to create additional gold standard data

---

Redundancy: Wisdom of Crowds

So far: ensuring individual annotation quality. However: even best effort human annotations are not perfect most of the time.
Redundancy: Wisdom of Crowds

Source: https://www.domo.com/learn/the-wisdom-of-crowds
Redundancy: Wisdom of Crowds

Source: https://www.domo.com/learn/the-wisdom-of-crowds
Redundant Annotations and Aggregation of Results

Redundancy:
- Each task is annotated by multiple workers
- Quality can be estimated based on inter-annotator agreement (e.g. Fleiss’ kappa)
  - If agreement threshold is not reached: obtain more annotations
- Aggregate answers by majority vote (in categorical case)
  - Improves accuracy if workers are better than random
- Introduces additional cost
Majority vote accuracy depending on redundancy and individual accuracy, assuming equal accuracy (p) for all workers.\(^\text{17}\)

\(^{17}\)Figure source: https://www.slideshare.net/ipeirotis/managing-crowdsourced-human-computation

In reality, worker accuracy varies. Solution: estimate worker reliability and take it into account for computing the labels.

Sketch of the approach:

▶ Treat each worker as a classifier characterized by
  ▶ Probability of correctly classifying each class
▶ Estimate iteratively using expectation maximization:
  ▶ E-step: Estimate hidden true labels of the data given current estimated worker reliability
  ▶ M-step: Estimate worker reliability given current estimated labels

Aggregation: Estimating Worker Reliability and Task Difficulty

Figure 1. Factor graph for the joint difficulty-ability-response estimation (DARE) model.

Figure source: Bachrach et al.20

Incentives: Influence of Payment 1

Setting: Reorder a list of images taken from traffic cameras chronologically.

611 participants sorted 36,000 images sets of varying size, for varying payments.

Findings

Figure: Accuracy (left) and number of completed tasks (right) in relation to payment
Findings

Figure: Post-hoc survey of perceived value
Influence of Payment 2

Setting: Find planets orbiting distant stars

---

Findings

Experiments with 356 workers and 17,000 annotated light curves. Simulated transits were added to real light curves (with noisy brightness values).

Figure: Accuracy of results depending on task difficulty.
No Payment: Games With A Purpose

Figure: The ESP Game

No Payment: Games With A Purpose

Figure: The ESP Game

Player 1 guesses: purse
Player 1 guesses: bag
Player 1 guesses: brown
Success! Agreement on “purse”

Player 2 guesses: handbag
Player 2 guesses: purse
Success! Agreement on “purse”

Monetary incentives
Between Games and Payment: Paid Competitions

Monetary incentives

Competitions

Microtask Crowdsourcing Competitions

Competitive setting on external application:

- Workers compete during limited time period
- Participants obtain scores based on their performance (e.g. no. of tasks fulfilled)
  - Worker Ranking
  - Distribute rewards according to the rank
- Participants are shown scores of their immediate competitors
Competition among Top-10 Workers

(a) Baseline

(b) Competition

R=1  R=4  R=7  R=10
R=2  R=5  R=8
R=3  R=6  R=9

Markus Rokicki | rokicki@L3S.de | 26/38
Between Games and Payment: Competitions and Teamwork

Monetary incentives

Competitions

---

Between Games and Payment: Competitions and Teamwork

Monetary incentives

Competitions

Between Games and Payment: Competitions and Teamwork

Competitions and Teamwork

Rewards:
- Non-linear distribution among teams
- Individual share proportional to contribution

Communication
- Team chats
Top-10 Workers

(a) individual comp.  
(b) team competition  
(c) combined
Paper 1: Dynamics and task types on MTurk

Data-driven analysis of microtask crowdsourcing platform. Based on scraping information:

<table>
<thead>
<tr>
<th>Label objects in image</th>
<th>View a HIT in this group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requester: Princeton Vision</td>
<td>HIT Expiration Date: Feb 4, 2015 (2 days 20 hours)</td>
</tr>
<tr>
<td>Time Allotted: 60 minutes</td>
<td>HITs Available: 17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LabelMe: Label objects in this image - Qualification Hit</th>
<th>View a HIT in this group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requester: John Kozar</td>
<td>HIT Expiration Date: Feb 9, 2015 (1 week 1 day)</td>
</tr>
<tr>
<td>Time Allotted: 30 minutes</td>
<td>HITs Available: 3</td>
</tr>
<tr>
<td>Description: Please look at this image, outline all the objects present in the image, and enter their names. These hits</td>
<td></td>
</tr>
<tr>
<td>Keywords: LabelMe, annotation, image, computer, vision, object, recognition, label</td>
<td></td>
</tr>
<tr>
<td>Qualifications Required:</td>
<td></td>
</tr>
<tr>
<td>HIT approval rate (%) is greater than 95</td>
<td></td>
</tr>
<tr>
<td>Location is US</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annotate retinal images to highlight lesions and abnormalities</th>
<th>View a HIT in this group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requester: Mr. Danny Mitry</td>
<td>HIT Expiration Date: Feb 7, 2015 (5 days 11 hours)</td>
</tr>
<tr>
<td>Time Allotted: 4 days</td>
<td>HITs Available: 1</td>
</tr>
</tbody>
</table>

Paper1: Dynamics and task types on MTurk

Micro Reward (USD)

Count

Year

2009 2010 2011 2012 2013

0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08

Figure taken from paper.
Paper1: Dynamics and task types on MTurk

Figure taken from paper.
Paper 2: Postprocessing results for quality

- How to learn from noisy human (expert and crowd) labels?
- Estimate worker reliability and true labels using EM
- Scenarios:
  - Binary classification
  - Multi-class classification
  - Ordinal regression
  - Regression

---

Paper 2: Postprocessing results for quality

Figure taken from paper.
Paper 2: Postprocessing results for quality

Figure taken from paper.
Paper 3: Influence of compensation and payment on quality

- Empirical study on influences on overall judgment quality
- Varying
  - payment
  - worker qualifications
  - required effort
- Scenario: relevance assessments for information retrieval

---

Paper 3: Influence of compensation and payment on quality

Figures taken from paper.
Paper 4: Collaborative workflows

Word processing interface that integrates workers to edit parts of documents on demand.

Find-Fix-Verify: split tasks into generation and review stages.

Figure taken from paper.