Practice of Efficient Data Collection via Crowdsourcing: Aggregation, Incremental Relabelling, and Pricing

Alexey Drutsa, Valentina Fedorova, Dmitry Ustalov, Olga Megorskaya, Evfrosiniya Zerminova, Daria Baidakova
Part VIII

Theory on incremental relabelling and pricing

Valentina Fedorova,
Research analyst

Yandex.Toloka is a service of Swiss company Yandex Services AG
Tutorial outline

Part I: 40 min
Main Components

Introduction: 20 min

Part II: 25 min
Brainstorming pipeline

Coffee break: 30 min

Part III: 10 min
Introduction to Crowd Platform

Part IV: 85 min
Set & Run Projects

Coffee break: 30 min

Part V: 35 min
Interface & Quality control

Lunch break: 90 min

Part VI: 25 min
Theory on Aggregation

Part VI: 60 min
Set & Run Projects cont.

Part VII: 20 min
Incremental relabeling and pricing

Part VIII: 10 min
Results & Conclusions
Key components of labelling with crowds

Aggregation → Incremental relabelling → Performance-based pricing
Incremental relabelling
aka dynamic overlap
### Pool settings: dynamic overlap

<table>
<thead>
<tr>
<th>POOL NAME</th>
<th>pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRIVATE DESCRIPTION</td>
<td></td>
</tr>
<tr>
<td>Use project description</td>
<td>✓</td>
</tr>
<tr>
<td>PUBLIC DESCRIPTION</td>
<td>Does all traffic lights are selected (by a rectangle) on the image?</td>
</tr>
<tr>
<td><strong>Price per task suite</strong></td>
<td></td>
</tr>
<tr>
<td>PRICE</td>
<td>0.01</td>
</tr>
<tr>
<td>MARKUP</td>
<td>0.05</td>
</tr>
<tr>
<td>Dynamic pricing</td>
<td>[+ Dynamic pricing]</td>
</tr>
<tr>
<td><strong>Overlap</strong></td>
<td></td>
</tr>
<tr>
<td>OVERLAP</td>
<td>3</td>
</tr>
<tr>
<td>DYNAMIC OVERLAP</td>
<td>Off</td>
</tr>
<tr>
<td><strong>Tasks settings</strong></td>
<td></td>
</tr>
<tr>
<td>TIME ON TASK</td>
<td>600</td>
</tr>
<tr>
<td>EXPIRES</td>
<td>2020-07-12</td>
</tr>
<tr>
<td>CAPTCHA FREQUENCY</td>
<td>None</td>
</tr>
<tr>
<td>TIME TO CLOSE</td>
<td>0</td>
</tr>
</tbody>
</table>
Incremental relabelling problem

Obtain aggregated labels of a desired level of quality using a fewer number of noisy labels.
Incremental relabelling scheme (IRL)

Request 1 label for each object

In real time IRL algorithm receives:
(1) previously accumulated labels
(2) new labels

Decides:
(1) which object are labelled
(2) which objects to relabel

Repeat until all tasks are labelled
Notations

› Consider one object

› \( z \in \{1, \ldots, K\} \) - latent true label

› \( y_w \in \{1, \ldots, K\} \) - observed noisy label from worker \( w \):
Notations

› Noisy label model for worker $w$: $M_w \in [0,1]^{K \times K}$: $\Pr(Y_w = k | Z = c) = M_w[c,k]$

› Prior distribution: $\Pr(Z = k) = p_k$
Posterior distribution

\( \{y_{w_1}, ..., y_{w_n}\} \) - accumulated noisy labels for the object

Using Bayes rule:

\[
Pr(Z = k | \{y_{w_1}, ..., y_{w_n}\}) = \frac{Pr(Z = k) Pr(\{y_{w_1}, ..., y_{w_n}\} | Z = k)}{Pr(\{y_{w_1}, ..., y_{w_n}\})} = \frac{p_k \prod_{i=1}^{n} M_{w_i}[k, y_{w_i}]}{\sum_{t=1}^{K} p_t \prod_{i=1}^{n} M_{w_i}[t, y_{w_i}]}.
\]
Expected accuracy of aggregated labels

Let $A$ be an aggregation model, e.g. MV, DS, GLAD,…

Denote aggregated label $z^A = A(\{y_{w_1}, \ldots, y_{w_n}\})$

Expected accuracy of aggregated labels given noisy labels is

$$E(\delta(z = z^A) | \{y_{w_1}, \ldots, y_{w_n}\}) = Pr(z = z^A | \{y_{w_1}, \ldots, y_{w_n}\})$$

Stop labeling if

$$E(\delta(z = z^A) | \{y_{w_1}, \ldots, y_{w_n}\}) \geq C$$
Incremental relabelling algorithm

Input: $\bigcup_{t=1}^{T-1} Y^t$ - previous labels till step $T$

$Y^T$ - new labels

Output: $R$ - objects to relabel

For each object $j$ with a label in $Y^T$:

- Object with a new label
- Current aggregated label
- Expected accuracy for the current aggregated label

$z_j^M = M(\bigcup_{t=1}^{T} Y^t)$

$c_j = E\left(z_j = z_j^M \mid \bigcup_{t=1}^{T} Y^t\right)$

If $c_j < c$, then $R = R \cup j$

Parameter: $c$ - threshold for expected accuracy
Threshold in IRL: cost – accuracy trade-off

Optimal threshold $c = 0.95$
A higher $c$ does not increase accuracy
Saving $\sim 35\%$ of noisy labels
How to obtain a cost-accuracy plot

Data for the plot:
› Label a pool of objects with a redundant overlap (e.g., 10)
› Obtain ground truth labels for the objects (e.g., expert labels or MV labels)

Simulate IRL with different thresholds using the data:
› For each threshold $c$ from a grid $0 < c_0 < \ldots < c_m \leq 1$:
› Repeat N times:
  1. Shuffle noisy labels and fix the order of labels
  2. Draw labels sequentially and test the IRL condition after each label
  3. Once the IRL condition for an object is met, discard unused labels for the object
  4. When all objects are labelled calculate
     • accuracy of aggregated labels
     • cost as the fraction of used noisy labels
› Average N values of aggregated accuracy and N values of cost for each value of threshold $c$
Key components of labelling with crowds

Aggregation → Incremental relabelling → Performance-based pricing
Performance-based pricing
aka dynamic pricing
Pool settings: dynamic pricing

Price per task suite
- Price: 0.01
- Markup: 0.005
  + Dynamic pricing

Overlap
- Overlap: 3
- Dynamic overlap: Off

Tasks settings
- Time on task: 600
- Captcha frequency: None
- Expires: 2020-07-12
- Time to close: 0
Labelling as a game: notation

Worker $w$

Task

Effort $h \in [0,1]$

Accuracy $a \in [0,1]$

Value $v = v(a)$

Payment $p = p(a)$

Requester

Classify images:
- Cat
- Dog
- Other

$\mathcal{A} = \{\text{Cat}, \text{Dog}, \text{Other}\}$
Labelling as a game: formalization

> Each worker $w$ chooses a level of effort $h$ for labelling object to maximize earnings per unit of spent effort:

\[
p(a_w(h)) \rightarrow \max_{h \geq 0}
\]

> The requester chooses a pricing $p(a)$ to minimize payments per unit of obtained value:

\[
\frac{v(a)}{p(a)} \rightarrow \max_{a \in [0,1]}
\]
Labelling as a game: incentive compatible pricing

Assume \( a_w(h) \) is a linear function of \( h \):

\[
a_w(h) = c_1 h + c_0
\]

Accuracy

Theorem:

The requester and workers maximize their utility simultaneously if the pricing \( p(a) \) for each label is proportional to its accuracy \( a \)
Performance-based pricing in practice\textsuperscript{4}: settings

\begin{itemize}
  \item Price $p$ for the level of accuracy $a_0$: $\Pr(\hat{z} = z) \geq a_0$

    \begin{itemize}
      \item $p = 0.3\$\quad a_0 = 0.99$
    \end{itemize}

  \item $\hat{q}_w = \Pr(y^w = z)$ - estimated quality level of worker $w$, e.g. the fraction of correct labels for golden set (GS):

    \begin{itemize}
      \item 5 correct GS among 10 \quad $\hat{q}_w = 0.5$
      \item 16 correct GS among 20 \quad $\hat{q}_w = 0.8$
      \item 100 correct GS among 100 \quad $\hat{q}_w = 1$
    \end{itemize}
\end{itemize}

\textsuperscript{4} Wang, Ipeirotis, and Provost, Quality-Based Pricing for Crowdsourced Workers, 2013
Performance-based pricing in practice: settings

Aggregation \[ \hat{z}_j^{w_{MV}} = \arg \max_{y=1,\ldots,K} \sum_{w \in W_j} q_w \delta(y = y_j^w) \]

IRL algorithm is based on the expected accuracy of \( \hat{z}_j^{w_{MV}} \)
Performance-based pricing in practice

**Pricing rules**

1. If \( \hat{q}_w \geq a_0 \), then the price is \( p \)

2. Else find \( n \):

\[
\sum_{k=0}^{n/2} \binom{n}{k} \hat{q}_w^{n-k} (1 - \hat{q}_w)^k \geq a_0
\]

Expected accuracy for MV

The price is \( \frac{p}{n} \)

\[
a_0 = 0.99
\]

\[
\hat{q}_w = 1 \quad \Rightarrow \quad n = \infty
\]

\[
\hat{q}_w = 0.8 \quad \Rightarrow \quad n = 15
\]

\[
\hat{q}_w = 0.5 \quad \Rightarrow \quad n = \infty
\]
Key components of labelling with crowds

Aggregation  →  Incremental relabelling  →  Performance-based pricing
Thank you!
Questions?

Valentina Fedorova
Research analysts

valya17@yandex-team.ru

https://research.yandex.com/tutorials/crowd/wsdm-2020