Aggregating outputs

Jérôme Waldispühl, McGill University
Designing HIT

Input → Task routing → Task design → Task aggregation → Output

Today
Aggregating output

Challenges:
• Output are noisy (lack of expertise)
• Humans are not always reliable (cheating)
• Cultural context may bias the answers

Goal: Automatic procedure to merge HIT results

Assumptions:
• It exists a “True” answer
• Redundancy helps
What is truth?

**Objective truth**
exists freely or independently from a mind (E.g. ideas, feelings)

Medical diagnosis, protein structure, number of birds...

**Cultural truth**
shared beliefs of a group of people, often involving perceptual judgments.

Is the music sad? Is this image pornographic? Is this text offending? ...
Latent class models

Observed: HIT outputs
Latent (hidden): Truth, user experience, task difficulty.

- Often, the matrix is incomplete
- Ground truth may never been known
Majority vote

Observed output

True answer
Majority vote (2)

\[
Y_n = \arg \max_j P(Y_n = j \mid O)
\]

\[
Y_n = \arg \max_j \frac{\prod_{m=1}^{M} P(O_{n,m} = o_{n,m} \mid Y_n = j)P(Y_n = j)}{P(O)}
\]

\[
Y_n \propto \arg \max_j \prod_{m=1}^{M} P(O_{n,m} = o_{n,m} \mid Y_n = j)
\]

\[
Y_n \propto \arg \max_j (1 - \varepsilon) \sum_{m=1}^{M} 1_{(o_{n,m} = j)} \cdot \varepsilon \sum_{m=1}^{M} 1_{(o_{n,m} \neq j)}
\]
Hidden factors

Task Characteristics
• Quality (e.g., blurry pictures)
• Difficulty (e.g., transcription of non-native speech)

Worker Characteristics
• Expertise (e.g., bird identification)
• Bias (e.g., mother vs college students)
• Physical Conditions (e.g., fatigue)
Incorporating worker quality

Observed output  Worker Characteristics

$O_{nm}$  $\pi_{m}$

True answer
Example: Medical classification

**Objective:** Medical diagnosis by doctors

**Model:** Doctors have different rates and types of errors.
- $\pi_{jl}^{(k)}$ defines the probability of doctor $k$ to declare a patient in state $l$ when the true state is $j$,
- $\eta_{il}^{(k)}$ is the number of time the clinician $k$ gets responses $l$ from patient $i$.

**Solution:** Expectation-Maximization (EM) Algorithm.

(Dawid & Skene, 1979)
EM Algorithm in a nutshell

**Goal:** Maximize the likelihood

\[
p(O \text{ on patient } i) \propto \sum_{j=1}^{J} p_j \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{jl}^{(k)}) \eta_{il}^{(k)}
\]
EM Algorithm in a nutshell

Idea:
1. Obtain some initial estimates of the missing data,
2. [Expectation step] Calculate the maximum likelihood estimates for the quantities of interest as if the missing data had been found,
3. [Maximization step] Calculate new estimates of the missing data,
4. Repeat steps 2. and 3. until both the maximum likelihood estimates and the missing data estimates converge.
EM Algorithm in a nutshell

We can calculate the maximum likelihood estimates:

$$\hat{\pi}^{(k)}_{jl} = \frac{\sum_i T_{ij} \eta_{il}^{(k)}}{\sum_l \sum_i T_{ij} \eta_{il}^{(k)}}$$

And estimate the probabilities:

$$\hat{p}_j = \frac{\sum_i T_{ij}}{I}$$

Where $T_{ij}$ is the set of indicators ($T_{ij}=1$ if $j$ is the true response and 0 otherwise).
EM in a nutshell

1. Take initial estimates of the T's.
2. Compute \( \pi \)'s and p’s using previous equations
3. Use estimates of \( \pi \)'s and p’s to compute new T’s s.t.

\[
p(T_{ij} = 1 \mid \text{data}) = \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{jl}^{(k)})^{\eta_{il}^{(k)}} p_j \div \sum_{q=1}^{J} \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{ql}^{(k)})^{\eta_{il}^{(k)}} p_q
\]

4. Iterate 2. and 3. until convergence
Incorporating task difficulty

Task difficulty: $\beta_n$ → Observed output: $O_{nm}$ ← Worker Characteristics: $\Pi_m$

True answer: $Y_n$
Example

HIT: Select images containing at least one “duck”

- competence varies with bird image,
- worker’s bias toward various mistakes,
- difficulty of the image.

(Welinder et al., 2010)
Example

• 200 waterbirds
• 40 pictures w/o birds
• 40 workers

(Welinder et al., 2010)
Learning from imperfect data

So far we assumed that:

• The system can distribute HIT to many unique workers
• The worker performs enough task to estimate their performance rates.

Problem: Does not always hold true…

Dekel & Shamir (2009):
1. Train classifier on unfiltered data
2. Use learned hypothesis to “guess” the truth and use it to remove bad workers.
Learning from Imperfect data

Other source of errors:
• inaccurate labels
• redundant labels
Beyond labeling

Challenges:
• How to decompose the problem?
• How to aggregate the results?
1. Pairwise comparison & creation of a ranking.

2. (When possible) Rank all objects & compute a consensus ranking.
Clustering

Objective: minimizing disagreement.

Technique:
- HIT link or separate object & cluster are computed from these properties.
- HIT link object to predefined sets and consensus is performed.
Prediction markets

**Hypothesis**: Knowledge is distributed and can be accessed by aggregating the belief of many individuals.

**Technique**: Workers report their belief and we estimate the probability that an event happen based on mean & median.

*Source: Data from 489 movies, 2000–2003 ([http://www.hsx.com](http://www.hsx.com)).*
References

Human Computation
Edith Law, Luis von Ahn
Morgan & Claypool Publishers