A Stepwise Analysis of Aggregated Crowdsourced Labels Describing Multimodal Emotional Behaviors

Alec Burmania and Carlos Busso

Multimodal Signal Processing (MSP) lab
The University of Texas at Dallas
Erik Jonsson School of Engineering and Computer Science
Labels from Expressive Speech

- Emotional databases rely on labels for classification
  - Usually obtained via perceptual evaluations

Lab Setting

  + Allows researcher close control over subjects
  - Expensive
  - Small demographic distribution
  - Smaller corpus size

Crowdsourcing

  + Can solve some of the above issues
  + Widely tested and used in perceptual evaluations
  - Raises issues with rater reliability

Amazon Mechanical Turk

CrowdFlower
Labels from Expressive Speech

- How do we balance quality and quantity in perceptual evaluations?
- How many labels is enough?
- Crowdsourcing makes these decisions important

Many Evaluators & Low Quality

or

Few Evaluators & High Quality

- What is the value of an extra evaluator?
Previous Work

- Burmania et al. (2016) explores tradeoff between quality and quantity of emotional annotations on emotion classification
  - Explore the concept of effective reliability proposed by Rosenthal [2008]
    
    \[ R_{SB} = \frac{n\kappa}{1 + (n - 1)\kappa} \]

- It is equivalent to have:
  - 15 annotators with reliability \( \kappa=0.45 \) (\( R_{SB}=92 \))
  - 10 annotators with reliability \( \kappa=0.54 \) (\( R_{SB}=92 \))

- Classification performance may be increase via design of label collection instead of maximizing inter-evaluator agreement

Motivation

- Compare the value of additional evaluators by analyzing consensus labels

\[ N \text{ evaluators} \]

\[ = \]

\[ N \text{ Evaluators} + 1 \text{ new evaluator} \]

- Derive guideline for subjective evaluations
  - Case study: emotional annotations of the MSP-IMPROV corpus
MSP-IMPROV Corpus

- Recordings of 12 subjects improvising scenes in pairs (>9 hours, 8,438 turns) [Busso et al, 2017]
- Actors are assigned context for a scene that they are supposed to act out
- Collected for corpus of fixed lexical content but different emotions

Data Sets
- Target – Recorded Sentences with fixed lexical content (648)
- Improvisation – Scene to produce target
- Interaction – Interactions between scenes

An example scene.

MSP-IMPROV Corpus

Anger
Lazy friend asks you to skip class

Happiness
Accepting job offer

Sadness
Taking extra help when you are failing classes

Neutral
Using coupon at store
Perceptual Evaluation

- Verify if a worker is spamming in real time
- We will focus on a five class problem (angry, sad, neutral, happy, other)
- Reference set includes target sentences (648)

## Rater Quality

**Constant sample size**

<table>
<thead>
<tr>
<th>Δθ</th>
<th># sent</th>
<th>κ</th>
<th># sent</th>
<th>κ</th>
<th># sent</th>
<th>κ</th>
<th># sent</th>
<th>κ</th>
<th># sent</th>
<th>κ</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>638</td>
<td><strong>0.572</strong></td>
<td>525</td>
<td>0.558</td>
<td>246</td>
<td>0.515</td>
<td>52</td>
<td>0.488</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>643</td>
<td>0.532</td>
<td>615</td>
<td>0.522</td>
<td>466</td>
<td>0.501</td>
<td>207</td>
<td>0.459</td>
<td><strong>26</strong></td>
<td><strong>0.455</strong></td>
</tr>
<tr>
<td>15</td>
<td>648</td>
<td>0.501</td>
<td>643</td>
<td>0.495</td>
<td>570</td>
<td>0.483</td>
<td>351</td>
<td>0.443</td>
<td><strong>112</strong></td>
<td><strong>0.402</strong></td>
</tr>
<tr>
<td>20</td>
<td>648</td>
<td>0.469</td>
<td>648</td>
<td>0.471</td>
<td>619</td>
<td>0.463</td>
<td>510</td>
<td>0.451</td>
<td>182</td>
<td>0.414</td>
</tr>
<tr>
<td>25</td>
<td>648</td>
<td>0.452</td>
<td>648</td>
<td>0.450</td>
<td>643</td>
<td><strong>0.450</strong></td>
<td>561</td>
<td>0.440</td>
<td>247</td>
<td>0.416</td>
</tr>
<tr>
<td>30</td>
<td>648</td>
<td>0.438</td>
<td>648</td>
<td>0.433</td>
<td>648</td>
<td>0.436</td>
<td>609</td>
<td>0.431</td>
<td>298</td>
<td>0.410</td>
</tr>
<tr>
<td>35</td>
<td>648</td>
<td>0.425</td>
<td>648</td>
<td>0.433</td>
<td>648</td>
<td>0.426</td>
<td>619</td>
<td>0.424</td>
<td>346</td>
<td>0.403</td>
</tr>
<tr>
<td>40</td>
<td>648</td>
<td>0.420</td>
<td>648</td>
<td>0.427</td>
<td>648</td>
<td>0.425</td>
<td>629</td>
<td>0.423</td>
<td>356</td>
<td>0.402</td>
</tr>
<tr>
<td>90</td>
<td>648</td>
<td><strong>0.422</strong></td>
<td>648</td>
<td>0.419</td>
<td>648</td>
<td>0.422</td>
<td>629</td>
<td><strong>0.419</strong></td>
<td>381</td>
<td>0.409</td>
</tr>
</tbody>
</table>

Increasing agreement due to filter

Decreasing samples meeting size criteria
Label Groups

- We consider two sets of labels based on kappa agreement:
  - High agreement group (n=12)
  - Moderate agreement group (n=20)

<table>
<thead>
<tr>
<th>Δθ</th>
<th>5 Raters</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># sent</td>
<td>κ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>638</td>
<td>0.572</td>
<td>525</td>
<td>0.558</td>
<td>246</td>
<td>0.515</td>
<td>52</td>
</tr>
<tr>
<td>10</td>
<td>643</td>
<td>0.532</td>
<td>615</td>
<td>0.522</td>
<td>466</td>
<td>0.501</td>
<td>207</td>
</tr>
<tr>
<td>15</td>
<td>648</td>
<td>0.501</td>
<td>643</td>
<td>0.495</td>
<td>570</td>
<td>0.483</td>
<td>351</td>
</tr>
<tr>
<td>20</td>
<td>648</td>
<td>0.469</td>
<td>648</td>
<td>0.471</td>
<td>619</td>
<td>0.463</td>
<td>510</td>
</tr>
<tr>
<td>25</td>
<td>648</td>
<td>0.452</td>
<td>648</td>
<td>0.450</td>
<td>643</td>
<td>0.450</td>
<td>561</td>
</tr>
<tr>
<td>30</td>
<td>648</td>
<td>0.438</td>
<td>648</td>
<td>0.433</td>
<td>648</td>
<td>0.436</td>
<td>609</td>
</tr>
<tr>
<td>35</td>
<td>648</td>
<td>0.425</td>
<td>648</td>
<td>0.433</td>
<td>648</td>
<td>0.426</td>
<td>619</td>
</tr>
<tr>
<td>40</td>
<td>648</td>
<td>0.420</td>
<td>648</td>
<td>0.427</td>
<td>648</td>
<td>0.425</td>
<td>629</td>
</tr>
<tr>
<td>90</td>
<td>648</td>
<td>0.422</td>
<td>648</td>
<td>0.419</td>
<td>648</td>
<td>0.422</td>
<td>629</td>
</tr>
</tbody>
</table>

High Agreement Condition

Moderate Agreement Condition
Label Aggregation

- Aggregation of votes is done using majority vote
- Each vote is equally weighted
- Votes are iteratively added chronologically as they were collected
- Due to majority vote, we establish the following transitions:
  - EmoA → EmoA (No Change)
  - EmoA → NA (No Agreement – a tie has been established)
  - NA → EmoA (A tie is broken)
  - NA → NA (tie remains a tie)

We cannot transition from one emotion to another!
Experiments

- Trends in labels will be evaluated iteratively for each added label
- We consider:

  - Label Stability
  - Label Changes
  - Frequency of Change
  - Adding more than one evaluator

Five class problem (angry, sad, neutral, happy, other)!
Label Stability

Percentage of videos with the same aggregated labels before and after adding an additional evaluator

- EmoA → EmoA
- NA → NA

Observations

- After 4 evaluators, labels are stable
- n=6, less than 10% of labels change
- Similar trends for high and moderate agreement conditions
Label Changes

Percentage of the videos in which their labels changed as we add one extra evaluator

- Inverse plots
- NA ➔ EmoA
- EmoA ➔ NA

Observations

- n=2, 40-44% agreement is lost
- n=3, most of the ties are solved
Change Frequency

Percentage of the videos in which their aggregated labels changed $m$ times as we incrementally add evaluators

- Example, ~25% change labels 2 times

- Observations
  - 45% to 50% never change labels
  - Trend on even values of $m$ indicate that ties are usually broken
  - About 75% sentences change labels less than 4 times
  - About 10% of the sentences change labels multiple times
Adding More than One Evaluator

- How different are the aggregated labels when we add more than one evaluator?
  - 3 versus 5, 5 versus 20
- This analysis does not follow the incremental stepwise approach
  - Snapshots different values of $n$
- We consider:
  - 3, 5, 9, and 20 annotators
- We have an additional case:
  - $\text{EmoA} \rightarrow \text{EmoB}$ (from one emotion to another)
Adding More than One Evaluator

Observations:
- Labels are very stable, even 3 versus 20 (76% overlap in labels)
- Only few labels benefit from extra evaluations
- Higher agreement case shows more stability
Discussion

- There is a reduced value in additional annotations
  - It helps about 10% of the labels

- We can save resources by tracking consistency of evaluations
  - Five evaluators per sentence resolve most of the ambiguities
  - We observe this trend for moderate and high inter-evaluator agreement

- Zhang et al. [2015] proposed to stop evaluation when agreement is reached
  - If n=5 and three people agree, stop the evaluation

Discussion

- An important exception is when consensus labels are not the goal
  - Training with soft-margin [Lotfian and Busso, 2017]
  - Study of emotion perception

- Emotion perceptual evaluations are complex cognitive tasks
  - We expect higher label stability for simpler behavioral tasks

R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in International Conference on Affective Computing and Intelligent Interaction (ACII 2017), San Antonio, TX, USA, October 2017.
Limitation and Future Work

- Generalizing the patterns in other databases
  - Larger or small numbers of classes
  - Different corpora
  - Inter-evaluator agreement variability
- Use of other aggregation techniques
  - Entropy based techniques
Questions?

Interested in the MSP-IMPROV database? Come visit us at msp.utdallas.edu and click “Resources”

This work was funded by NSF CAREER award IIS-1453781
References


