“Of All Things the Measure is Man”
Automatic Classification of Emotions and Inter-Labeler Consistency

Stefan Steidl, Michael Levit, Anton Batliner, Elmar Nöth, and Heinrich Niemann
Chair for Pattern Recognition, University Erlangen-Nuremberg, Germany

Introduction

For evaluating a classifier, it is necessary to know which class a given sample belongs to. The membership to a certain class is not always well-defined. In emotion recognition, e. g., human labelers do often not agree on one common emotion class. In our case, the reason is that we deal with realistic emotions that are very weak in contrast to full-blown emotions of acted speech. Due to different human activation levels, confusions with *neutral* are very common. There is a number of different measures to evaluate the inter-labeler consistency. In this paper, we pursue a different approach: we propose to incorporate the systematic confusions of the human labelers into the evaluation of the machine classifier by using a new entropy-based measure. At least in our scenario, recognition “errors” that also occur in human labeling are not as severe as if two classes are mixed up that are never confused by humans. We can show that a classifier which achieves a recognition rate of “only” about 60 % on a four class problem performs as well as our five human labelers on average.

Figure 1: Children playing with the Sony robot AIBO.

Our Scenario

Our Aibo-Emotion-Corpus consists of natural German speech of 51 children at the age of 10 to 13 years. The children were asked to direct the Aibo along a given route and to certain objects (Figure 1). To elicit emotions, we proceeded as follows:

- Aibo was operated by remote control and misbehaved at predefined positions.
- The children were told to address Aibo like a normal dog, especially to reprimand or to laud it.
- We put up some danger spots where Aibo was not allowed to go under any circumstances.
- The children were pressed slightly for time.

The corpus was annotated at word level by five experienced graduate labelers. Before labeling, the labelers agreed on a common set of eleven emotions: angry, touchy, reprimanding, joyful, motherese, emphatic, surprised, bored, hesitated, neutral, and remaining. Due to sparse data, we mapped these emotions onto basically the four cover classes anger, motherese, emphatic, and neutral.

Our Problem

Our experiments are based on a subset of our data, consisting of 1557 words which the majority labeled as **Anger (A)**, 1224 words labeled as **Motherese (M)**, and 1645 words each for **Emphatic (E)** and **Neutral (N)**. Still, the inter-labeler agreement is very low:

- All five labeler agree in only 14 % of all cases.
- In 54 % of all cases, only three of five labelers agree.
- Multi rater kappa: 0.36
- Weighted multi rater kappa: 0.48

There are a number of reasons why the inter-labeler consistency is low:

- The emotions are relatively weak. Consequently, confusions with *Neutral* are frequent because of different activation levels of the labelers.
- *Emphatic* is a sort of pre-stage for *Anger*. Therefore, *Emphatic* is not only often confused with *Neutral* but also with *Anger*.

Taking the majority voting of our five labelers does not fully reflect the ground truth: Deciding for a different class does not necessarily mean that this decision is (totally) wrong.

Entropy-Based Evaluation of Decoders

Convert the hard decisions of your reference labelers into one soft label l_{ref}. Omit one labeler who can be used as a decoder.

labeler	class
1	A
2	E
3	A
4	N
5	A
6	E
7	A
8	A
9	N
10	E

→

A	M	E	N
0.5	0.0	0.3	0.2

Table 1: Conversion of the hard decisions of the reference labelers into one soft label .

Add the decision of the decoder l_{dec}:

$$l(s) = \frac{1}{2} \cdot l_{\text{ref}}(s) + \frac{1}{2} \cdot l_{\text{dec}}(s)$$

Calculate the entropy for a given sample s:

$$H(s) = -\sum_{k=1}^{K} l_k(s) \cdot \log_2(l_k(s))$$

Implicit weighting of classification “errors”

The more the reference labelers agree on one class the lower the entropy will be. If the decoder is added, the entropy will change in the given example of Table 1:

- It will decrease if the decoder decides for *Anger* as this is the result of the majority voting.
- It will slightly increase if the decoder decides for *Emphatic* since at least 30 % of the human labelers also decided for this. The same for *Neutral*.
- It will highly increase if the decoder decides for *Motherese* as no human labeler decided for this class.

Average the entropy over series of samples and plot histograms or calculate the mean entropy for the whole data set (S being the number of samples):

$$H = \frac{1}{S} \sum_{s=1}^{S} H(s)$$

Experimental Results

In our experiments, we compare different classifiers with a human labeler on average.
Figure 2 shows that the average human labeler clearly outperforms a random choice classifier which randomly chooses one of the four classes. It also outperforms a naive classifier which always decides for *Neutral* or *Motherese*.

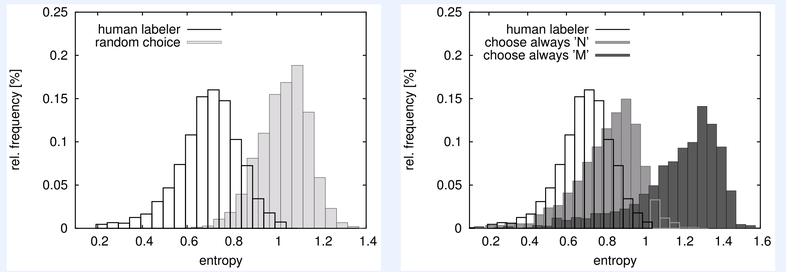

Figure 2: Entropy histograms for different naive decoders in comparison to the human labelers

Figure 3 (left) shows the comparison of an average human labeler and our machine classifier which is based on 95 prosodic features and 30 part-of-speech features. The average recognition rate per class is 58.1 % for this four class problem. Both decoders are almost identical; our machine classifier performs as well as one of our labelers on average. The entropy measure has its minimum if a decoder always chooses what the majority of reference labelers decides for (right part of Figure 3).

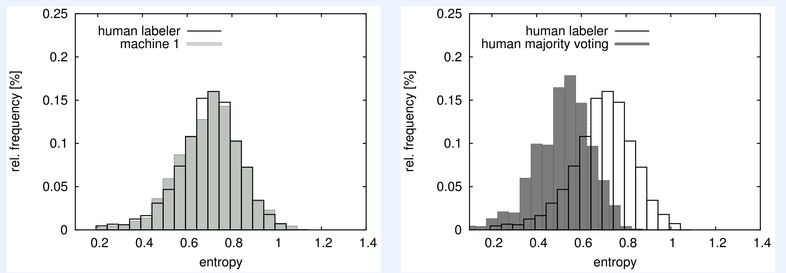

Figure 3: Left: Entropy histograms for our machine classifier in comparison to the human labelers. Right: Comparison of the majority voting and the average human labeler.