“Of all things the measure is man” - automatic classification of emotions and inter-labeler consistency

Stefan Steidl, Michael Levit, Anton Batliner, Elmar Nöth, Heinrich Niemann

Angaben zur Veröffentlichung / Publication details:

Steidl, Stefan, Michael Levit, Anton Batliner, Elmar Nöth, and Heinrich Niemann. 2005. “‘Of all things the measure is man’ - automatic classification of emotions and inter-labeler consistency.” In *International Conference on Acoustics, Speech, and Signal Processing [ICASSP 2005], 23 March 2005, Philadelphia, PA, USA*, 317–20. Piscataway, NJ: IEEE. https://doi.org/10.1109/ICASSP.2005.1415114.

“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

Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg
Martensstraße 3, 91058 Erlangen, Germany
steidl@informatik.uni-erlangen.de

Published at ICASSP 2005, Philadelphia, March 18-23, 2005, U.S.A., ISBN: 0-7803-9076-8 (DVD-ROM)

ABSTRACT

In traditional classification problems, the reference needed for training a classifier is given and considered to be absolutely correct. However, this does not apply to all tasks. In emotion recognition in non-acted speech, for instance, one often does not know which emotion was really intended by the speaker. Hence, the data is annotated by a group of human labelers who do not agree on one common class in most cases. Often, similar classes are confused systematically. We propose a new entropy-based method to evaluate classification results taking into account these systematic confusions. 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.

1. INTRODUCTION

An essential aspect of pattern recognition is the classification of patterns. Besides the search for applicable features, a lot of work has also been done to develop new and to improve existing automatic classification techniques. Well known are, for instance, artificial neural networks, or support vector machines which became very popular in the last few years. In the case of supervised learning, the classifiers are trained to map a set of features onto a given reference class. The standard method to evaluate an automatic classifier is to calculate the recognition rate which is the percentage of correctly recognized samples. The basic assumption is that this reference class is given and that it is non-ambiguous.

In our work on the recognition of emotions on the basis of emotional speech, we face the problem that it is not clear at all which emotions the people expressed when they were recorded. The corpus on which the experiments were done in this paper consists of children playing with the Sony robot Aibo. The kids were asked to direct the Aibo along a given route; they were not asked to express any emotions. Nevertheless emotional behavior can be observed in these recordings. As these emotions are not acted by professional actors, but are emotions as they appear in daily life, they are called “realistic”. From the application developers’ point of view, it is very important to deal with such realistic behavior. However, one side effect is that one has to cope with relatively weak emotions in contrast to full-blown emotions of acted speech. As the recorded persons do not have to play a given emotion and due to the fact that it is often not feasible to ask them afterwards what kind of emotion they felt during the recordings, one employs human labelers to label the data set. Normally, only in a few cases, all available labelers agree on one common label. In our corpus, in most cases, only three out of five labelers agreed. Yet this is not a problem of bad labeling but rather a consequence of the fact that we are dealing with a realistic classification problem which also raises real difficulties for humans. Accordingly, the expectations of the automatic classifier measured in recognition rates must be lowered.

In order to be able to calculate recognition rates at all, hard decisions are needed for the reference as well as for the classifier’s output. If a metric can be imposed on the label space, the labels of all labelers can be averaged. This is well possible, for example, if the tiredness of persons is labeled on a scale from 1 to 10. If two labelers judge someone with ’8’ as very tired and one labeler says only ’5’, the reference would be ’7’. But this does not work for categorical emotion labels like *anger*, *bored*, etc. as the mean of *anger* and *bored* is not defined. In those cases, the state-of-the-art is to use a majority voting to create the reference. Proceeding this way, we achieve recognition rates of about 60 % on our corpus with four emotion classes which is a state-of-the-art result for a task set-up like that. Nonetheless, the assessment of the emotion classification success should not be done without considering how well humans would perform this task. Depending on the number and type of classes, human labelers confuse certain classes with each other more than other classes. In general, the more similar classes are, the more they are confused. This confusion should be considered in the evaluation of a classifier. If the automatic classifier makes the same “mistakes” as many humans do, then this fault cannot be as severe as if the classifier mixes up two classes that are never confused by humans. Instead, the question is if such systematic confusions are faults at all since “of all things the measure is man” as already Protagoras said more than 2400 years ago.

In this paper, we would like to propose a new entropy-based measure to judge a classifier’s output taking systematic confusions made by humans into account.

2. ENTROPY-BASED MEASURE

According to Shannon’s information theory [1], the entropy is a measure for the information content. We propose to use the entropy to measure the unanimity of the labelers. If all reference labelers agree on one class, the entropy will be zero. Otherwise, the more the labelers disagree, the higher the entropy will be. In the following, we assume to have N labelers L_n who have labeled a data set of S samples X_s. For each sample, each of our labelers has to decide in favor of one of K classes C_k. However, the

This work was partially funded by the EU projects PF-STAR (IST-2001-37599) and HUMAINE (IST-2002-507422). The responsibility for the content lies with the authors.

approach is also easily portable to soft labels where all classes get scores from a continuous range of values and all scores for a sample sum up to one. The hard decisions of any number of labelers can be converted into one soft reference label as it is depicted in Fig. 1 for a four-class-problem ($K = 4$) with ten labelers. The more the labelers disagree the flatter is the distribution of the soft label.

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

Fig. 1. Conversion of the hard decisions of ten labelers into a soft reference label $\boldsymbol{l}_{\text{ref}}$. The four classes are '**A**nger', '**M**otherese', '**E**mphatic', and '**N**eutral'

Our suggestion is to leave out each labeler (we can also use a more general term "decoder") in succession. If labeler n is left out, then the resulting soft reference label for sample X_s is denoted $\boldsymbol{l}_{\text{ref}}(\bar{n}, s)$, with $\bar{n}$ indicating the omitted labeler.

Now, we add another decoder. This can be an automatic classifier, but also the remaining human labeler who was omitted in the reference, so that direct comparisons between a classifier and a human labeler are possible. In order to avoid dependency on the number of labelers, the new decoder is not considered in the same manner as the other reference labelers. Instead, the hard decision of the new decoder for sample X_s (also converted into a soft label $\boldsymbol{l}_{\text{dec}}(s)$) is weighted $1 : 1$ with the reference label $\boldsymbol{l}_{\text{ref}}(\bar{n}, s)$:

$$\boldsymbol{l}(\bar{n}, s) = 0.5 \cdot \boldsymbol{l}_{\text{ref}}(\bar{n}, s) + 0.5 \cdot \boldsymbol{l}_{\text{dec}}(s) \tag{1}$$

Then, the entropy can be calculated for the given sample X_s:

$$H(\bar{n}, s) = -\sum_{k=1}^{K} l_k(\bar{n}, s) \cdot \log_2(l_k(\bar{n}, s)) \tag{2}$$

Taking the example of Fig. 1, the entropy will decrease compared to the reference labels if the decoder decides in favor of '**A**nger' as '**A**nger' is what the majority of labelers said. Otherwise, if the decoder chooses '**E**mphatic', the entropy will increase but not as much as if the decoder decides in favor of '**N**eutral' since 30 % of the labelers agree that this sample is '**E**mphatic' and only 20 % said the sample is '**N**eutral'. As none of the labelers decided for '**M**otherese', choosing this class yields the highest entropy. This makes sense since '**M**otherese' seems to be definitely wrong in this case. Note that if using hard decisions, '**A**nger' would be the only correct class although 50 % of the labelers disagree.

The next step is to average the computed entropy value for X_s over the left-out labelers:

$$H(s) = \frac{1}{N} \sum_{n=1}^{N} H(\bar{n}, s) \tag{3}$$

We say that our classifier performs not worse than an average human labeler on sample X_s, if the entropy from Eq. 3 with our classifier as the new decoder does not exceed the entropy where the additional decoders were always humans. By plotting two corresponding histograms of $H(s)$ for the entire corpus, we obtain a visual means for the assessment of the performance of the classifier on this corpus: the closer the histogram for the machine classifier is to the histogram for the human labelers, the better the classifier is. In general, nothing is known about the distributions approximated by these histograms. However, if instead of plotting entropy values of individual samples we average them over series of several samples, then, according to the central limit theorem, the resulting distributions will be approximately normal, and thus, describable in terms of its means and variances. In our experiments we used series of 20 samples.

The overall entropy mean itself can be used for comparison and is computed by averaging $H(s)$ over all samples of the data set:

$$H = \frac{1}{S} \sum_{s=1}^{S} H(s) \tag{4}$$

3. THE AIBO-EMOTION-CORPUS

This entropy-based measure is useful in all those cases where a large discrepancy amongst the human reference labelers exist. In this paper, we demonstrate the evaluation of different decoders considering the example of emotion recognition in speech of children. All experiments are done on a subset of our Aibo-Emotion-Corpus which consists of 51 children at the age of 10 to 13 years. The children were asked to direct the Aibo robot along a given route and to certain objects. To elicit emotions, the Aibo was operated by remote control and misbehaved at predefined positions. In addition, the children were told to address Aibo like a normal dog, especially to reprimand or to laud it. Besides that, we pressed the children slightly for time and put up some danger spots where Aibo was not allowed to go under any circumstances. Nevertheless, the recorded emotions are relatively weak, especially in contrast to full-blown emotions of acted speech. The corpus consists mainly of the four emotions '**A**nger', '**M**otherese', '**E**mphatic', and '**N**eutral' which were annotated at word level by five experienced graduate labelers. Before labeling, the labelers agreed on a common set of discrete emotions. For a more detailed description of the corpus, please refer to [2]. As '**N**eutral' is the most frequent "emotion" by far, we downsampled the data until all four classes were equally present according to the majority voting of our five labelers. At least three labelers had to agree. Cases where less than three labelers agreed were omitted as well as those cases where other than the four basic classes were labeled. In the final data set, 1557 words for '**A**nger', 1224 words for '**M**otherese', and 1645 words each for '**E**mphatic' and for '**N**eutral' are used. The inter-labeler consistency can be measured using the multi rater kappa statistic. The formula is given e. g. in [3]. For our subset, the kappa value is only 0.36 which expresses the large disagreement of our five labelers. It is generally agreed that kappa scores greater than 0.6 indicate good agreement. There exists also a weighted version of the multi rater kappa statistic which weights confusions according to a given distance measure between the two confused classes [4]. This approach demands that the four classes are arranged on a linear scale. We assigned 2 to 'Neutral' to put 'Neutral' in the center, 1 to 'Motherese', 4 to 'Anger', and '3' to 'Emphatic' as we consider 'Emphatic' as a sort of pre-stage of 'Anger'. Doing

Published at ICASSP 2005, Philadelphia, March 18-23, 2005, U.S.A., ISBN: 0-7803-9076-8 (DVD-ROM)

this, we get a weighted kappa value of 0.48. The divers versions of kappa are not the only methods to evaluate the inter-labeler agreement. For an overview, please see [5]. As mentioned above, our low kappa value is not due to bad labeling. Rather, we are dealing with a difficult classification problem where even human labelers disagree about certain classes. On the one hand, our emotions are relatively weak what makes it hard to decide whether a given word is emotional or not. The consequence is a high confusion rate of the three emotion classes with 'Neutral'. On the other hand, 'Emphatic' as a pre-stage of 'Anger' is not only hard to distinguish from 'Neutral' but also from 'Anger'.

4. MACHINE CLASSIFICATION OF EMOTIONS

The experiments described in the following are all conducted with artificial neural networks. Because of the small data set, we do "Leave-One-Speaker-Out" experiments: each of the 51 speakers is used once for testing, while 40 of the remaining speakers are used for training, and the other 10 speakers for validation of the neural networks. As features we use our set of 95 prosodic features and 30 part-of-speech features. Details of these features can be found in [6, 7]. The total number of features is reduced to 95 using principal component analysis (PCA). Two different machine classifiers are trained: *machine 1* is trained with soft labels, *machine 2* with hard labels. The results in terms of traditional recognition rates are given Tab. 1 and Tab. 2 together with a confusion matrix of the classes. Note that in both cases, the output of the classifiers are hard decisions in order to be able to compute recognition rates. The majority voting of all five human labelers serves as hard reference. The average recognition rate per class is with 59.7 % slightly higher for *machine 2* which is trained with hard labels than for *machine 1* which achieves 58.1 %.

	A	M	E	N	Σ	RR
A	791	47	261	458	1557	50.8 %
M	56	559	27	582	1224	45.7 %
E	214	23	947	461	1645	57.6 %
N	100	94	161	1290	1645	78.4 %
∅						58.1 %

Table 1. Machine decoder 1 (trained with soft labels): confusion matrix and recognition rates (RR) evaluated using hard decisions for the classes '**A**nger', '**M**otherese', '**E**mphatic', and '**N**eutral'

	A	M	E	N	Σ	RR
A	899	90	303	265	1557	57.7 %
M	110	697	68	349	1224	56.9 %
E	273	43	1076	253	1645	65.4 %
N	215	201	266	963	1645	58.5 %
∅						59.7 %

Table 2. Machine decoder 2 (trained with hard labels): confusion matrix and recognition rates (RR) evaluated using hard decisions for the classes '**A**nger', '**M**otherese', '**E**mphatic', and '**N**eutral'

The intention of this paper is to compare these two machine classifiers with an average human labeler as described in Sec. 2. But prior to this, we present results for different naive classifiers. In Fig. 2 (left), entropy histograms for an average human labeler and a random choice classifier, which randomly chooses one of four classes, are shown. As expected, the mean entropy H from Eq. 4 for the simple classifier (1.050, Tab. 3) is much higher than for the average human labeler (0.722). Accordingly, the histogram of the random choice classifier is shifted to the right. This naive classifier clearly performs worse than one of our labelers on average. On the right side of Fig. 2, the histograms of two other naive decoders are shown. One classifier decides always in favor of '**N**eutral', the other one always for '**M**otherese'. Analyzing the data set, it is obvious that human labelers are often not sure whether they should label a word as emotional or as neutral due to the weak emotions we are dealing with. Consequently, deciding for '**N**eutral' conforms more to the human labeling behavior than deciding for a certain emotion class. This fact is reflected in our entropy values as well. The mean entropy for the classifier that always chooses '**N**eutral' is 0.843 which is better than random choice. In contrast, always deciding for '**M**otherese' is even worse (1.196).

decoder	entropy measure
human majority voting	0.542
human labeler	**0.721**
machine 1	**0.722**
machine 2	0.758
choose always 'N'	0.843
choose always 'E'	1.049
random choice	1.050
choose always 'A'	1.127
choose always 'M'	1.196

Table 3. Different decoders and their classification results w. r. t. our entropy measure

In the comparison with the two machine classifiers, the entropy measure H shows that the decoder *machine 1* performs as well as an average human labeler, albeit it yields an average recognition rate per class of "only" 58.1 %. The mean entropy is with 0.722 almost identical with the value attained by the human labelers (0.721). Our second machine decoder *machine 2*, even though it is slightly superior to *machine 1* in terms of recognition rates, performs a little worse than *machine 1* in terms of the mean entropy (0.758). The reason becomes obvious if one looks at the confusion matrices in Tab. 1 and Tab. 2. Both neural networks are trained in such a way that all four classes should be recognized comparably well. This works better if hard labels are used for training as in the case of *machine 2*. In contrast, *machine 1* tends to favor '**N**eutral', and this is exactly what humans do in our data set. This is why the entropy measure, being a rather intuitive one, prefers *machine 1* over *machine 2*, even though its overall recognition rate is slightly lower.

The reference for calculating recognition rates is the majority voting of all five labelers. This majority voting can also be interpreted as decoder. In Fig. 3 (right), this decoder is plotted in comparison with an average human labeler. The mean entropy of 0.542 specifies the minimum entropy which can be achieved by a machine decoder. Thus, a machine classifier can very well be better than a single human on average. The results show that we are as good as one of our human labelers on average, but that there is also enough room for further improvements.

Published at ICASSP 2005, Philadelphia, March 18-23, 2005, U.S.A., ISBN: 0-7803-9076-8 (DVD-ROM)

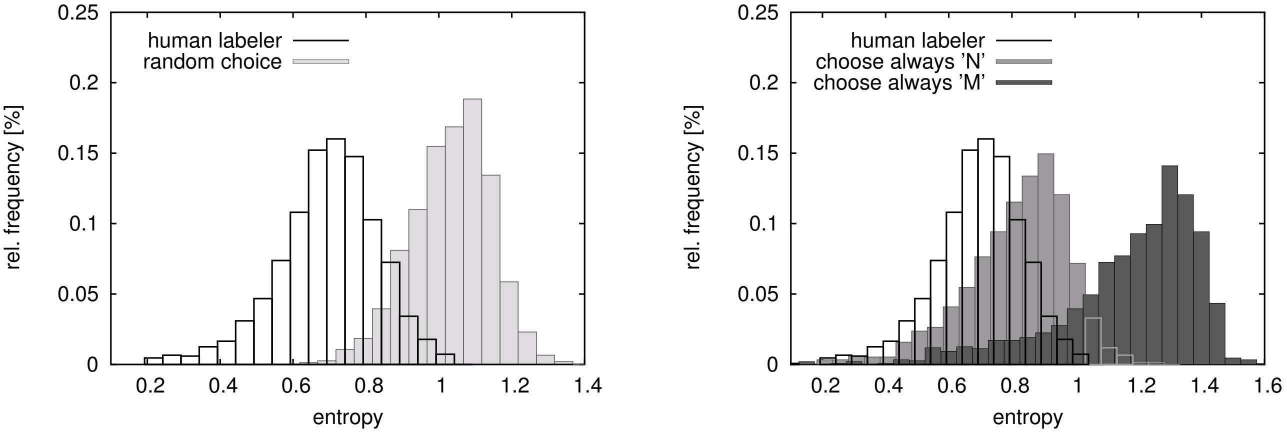

Fig. 2. Comparison between an average human labeler and three naive classifiers: a decoder which selects randomly one of the four classes (left) and two decoders which always choose '**N**eutral' and '**M**otherese' respectively (right)

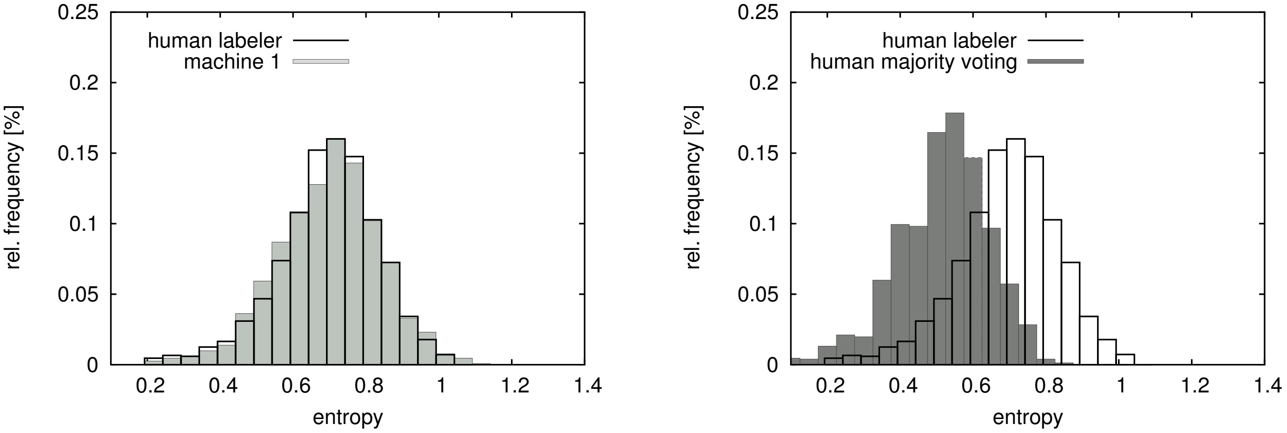

Fig. 3. Comparison between an average human labeler and our machine decoder 1 (left) and the majority voting of our five human labelers respectively (right)

5. CONCLUSION

We proposed a new entropy-based measure which makes a comparison between human labelers and machine classifiers possible. Even more important for the evaluation is the fact that systematic confusions of human reference labelers are taken into account as in most of our cases the reference is far from being non-ambiguous. For instance, slight forms of '**A**nger' are often confused with '**E**mphatic' or with '**N**eutral' since it is very hard to distinguish among these emotions – even for humans. From the application's point of view, deciding for a very similar class cannot be that wrong in those cases. Our measure automatically punishes those classification faults that also occur in human classification less than those faults that are never done by humans. Traditional recognition rates are not capable of this distinction.

6. REFERENCES

[1] C. E. Shannon, "A Mathematical Theory of Communication," in *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656. 1948, reprint available at http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html (08/17/2004).

[2] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russel, and M. Wong, "'You stupid tin box' - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus," in *Proc. of the 4th Int. Conference on Language Resources and Evaluation (LREC)*, 2004, vol. 1, pp. 171–174.

[3] R. Sproat, W. Black A. S. Chen, S. Kumar, M. Ostendorf, and C. Richards, "Normalization of non-standard words," in *Computer Speech and Language*, vol. 15, pp. 287–333. 2001.

[4] F. Krummenauer, "Erweiterungen von Cohen's kappa-Maß für Multi-Rater-Studien: Eine Übersicht," *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*, vol. 30, pp. 3–20, 1999.

[5] M. Reyelt, *Experimentelle Untersuchungen zur Festlegung und Konsistenz suprasegmentaler Einheiten fr die automatische Spracherkennung*, Shaker, Aachen, 1998.

[6] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to Find Trouble in Communication," *Speech Communication*, vol. 40, pp. 117–143, 2003.

[7] J. Buckow, *Multilingual Prosody in Automatic Speech Understanding*, Logos, Berlin, 2003.

Published at ICASSP 2005, Philadelphia, March 18-23, 2005, U.S.A., ISBN: 0-7803-9076-8 (DVD-ROM)