Active Learning for Dimensional Speech Emotion Recognition

Wenjing Han*, Haifeng Li*, Huabin Ruan§, Lin Ma*, Jiayin Sun†, Björn Schuller*†

* School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
§ Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
† Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, München, Germany

Abstract

State-of-the-art dimensional speech emotion recognition systems are trained using continuously labelled instances. The data labelling process is labour intensive and time-consuming. In this paper, we propose to apply active learning to reduce according efforts: The unlabelled instances are evaluated automatically, and only the most informative ones are intelligently picked by an informativeness measure function for a human to label. Specifically, we estimate the informativeness of each unlabelled instance based on a binary-classification confidence score for an emotion being predicted to be negative or positive on a given emotional dimension. For verification, we consider a pool-based and a stream-based scenario run on part of the continuous AVEC 2012 task to demonstrate the feasibility of the proposed approach in practice. In the result, our approach requires significantly less human labelled data instances to reach a given performance than passive learning does in both scenarios.

Index Terms: Active Learning, Speech Emotion Recognition, Affective Computing, Continuous Emotion Representation

1. Introduction

Dimensional speech emotion recognition (SER) research has recently gained increasing attention in the SER community [1, 2, 3, 4, 5]. This approach represents emotions as points in a multi-dimensional emotion space and predicts continuous-valued emotions using regression methods. For a supervised dimensional SER system to perform well, an extensive amount of continuously labelled speech instances are required for model learning.

However, although the spontaneous speech source is abundant in the real world and can be gathered conveniently, the continuous annotation process can be challenging: If the labelling is not performed fully continuously such as by tools as FEELTRACE [6] or GTrace [7], but by a discrete method, such as the popular SAMs [8], this requires several human labelers to reach a smooth continuum by averaging over labellers. If it is performed fully continuously, the process will become more tedious, even extremely time consuming. Take the Solid-SAL part of the SEMAINE corpus [9] for example, which consists of a duration of 7:32:54 (h:m:s) of natural human-human conversation recording and was annotated in five dimensions in continuous time and continuous value using FEELTRACE – the annotator moves a ‘slider’ in real-time for one dimension after another. Thus, for five dimensional labelling at least a working week is needed not taking exhaustion into account. Imagine that, with more and more audio being collected, the time consumption on annotation can become considerably huge. In this sense, a strategy to select instances worth for labelling and discarding redundant ones is desired for the purpose of reducing manual annotation effort.

On the other hand, it is well known that redundant or unevenly distributed instances will not only increase training cost, but also cause harmful interference on the performance of supervised learning systems. This is a problem worth attention, especially in the context of spontaneous SER: As experienced in our daily life, most of the time people stay in (or near) a neutral state rather than in pronounced emotional states, which naturally leads to an imbalance that the majority of spontaneous speech collection involves neutral or weak emotions rather strong emotions. So in this case, if the whole collection of speech is used for model training without appropriate filtering, the obtained SER model will be inevitably over-fitting on neutral or weak emotions, and less-fitting on strong emotions. In this sense, it can be helpful to make a spontaneous SER system able to selectively pick instances benefiting the training.

Essentially, the above two demands share the same goal: actively selecting data to be labelled which will have the largest improvement on the performance. This is where the Active Learning (AL) approach can assist. AL is a kind of machine learning strategy proposed in the 1990s and designed for reducing the number of training instances to be labelled by inspecting the unlabelled instances, and selecting the most informative ones with respect to a given informativeness measure function for a human to label [10, 11]. AL has been mainly employed in many classification applications until now, like automatic speech recognition [12, 13] or information extraction [14], and shown distinct advantage. In our previous work [15], we also successfully applied AL to SER in the case of discrete classes. In this paper, we adapt the algorithm for the regression application in a dimensional SER context to address its above demands. To the authors’ best knowledge, this is the first time to import AL into the dimensional SER field.

In this paper, a binary classification confidence score is calculated for the evaluation of the worth of an instance for labelling. Briefly, we construct a binary emotion classifier using a small amount of labelled instances to roughly classify unlabelled instances into negative or positive emotion, and calculate the classifier’s confidence scores to measure how certain
the classifier is with respect to corresponding classification results. The intuition is that there is a reverse correlation between such a binary classification confidence score given by the classifier and the informativeness degree of that instance. We then select those instances with low scores for humans to label, and use them for regression model training. To verify the validation of AL for dimensional SER, we investigate the performance impact on a dimensional SER system in two different problem scenarios: 1) a pool-based scenario, and 2) a stream-based scenario. The former is motivated by the typical real-world learning problem that a large pool of unlabelled data has been gathered in advance, such as by recordings of a call centre over thousands of conversations. By that, before deciding which instances should be selected in each training round, every instance in the pool has to be evaluated. The latter fits another real-world problem that unlabelled instances can be gathered sequentially, and one should decide which instances should be retained and which should be discarded on-line, such as in an on-line training SER system. In contrast to the former, the latter is more appropriate for the situation when memory or processing power may be limited, as with mobile and embedded devices. In this paper, all experiments are implemented on two dimensions of the large AVEC partition of the SEMAINE corpus of naturalistic emotion as in AVEC’s Word-Level Sub-Challenge.

The rest of this paper is organised as follows: Section 2 discusses the related-work. Section 3 presents a general description of our proposed AL approach for dimensional SER. Section 4 describes the used database and acoustic feature extraction. Then, Section 5 presents our experimental layout, results, and discussions before concluding in Section 6.

2. Related Work

In the context of human emotion recognition, AL has been previously introduced for classification in [16, 17] and by part of the authors in [15], [17] focused on the task of arousal classification especially using physiological features [18]. It aimed to construct a good classifier with a small amount of examples by optimally selecting classes to obtain training examples. The inverse approach, which selected examples from the hardest classified class in each iteration round, was verified to be useful. [15] proposed two AL strategies for binary speech emotion classification in the pool-scenario. One selected the speech instances likely to be the easiest ones to be annotated by humans from the candidates in the pool. The other one selected the speech instances according to the classifier confidence scores attached to them. Their effectiveness was investigated on the INTERSPEECH 2009 Emotion Challenge task [19], [20] investigates AL for music emotion. In contrast to the classification research of SER, the topic of dimensional emotion regression is relatively new [21]. We are not aware of works dealing with AL for regression in dimensional SER.

3. Approach

When we represent emotions in a k-dimensional emotion space (e.g., valence-arousal emotion space), each emotional state is associated with k emotion values, correspondingly. Since these dimensions are assumed to be largely independent of each other in the psychological view, a dimensional SER task can be naturally separated into k individual regression tasks. For a simplified depiction, the following description is thus given in the context of one of the k subtasks.

3.1. Problem description of AL for dimensional SER

Let, in training, a small set of labelled instances \(S_l = \{(x_i, y_i, \text{continuous})\}_{i=1}^m\) and a large source of unlabelled instances \(S_u = \{(y_j)\}_{j=1}^n\) be provided, where \(y_j\) is the corresponding label of \(x_i\). Human annotators are now asked to label the instances from \(S_u\) to improve the system performance. The goal next is to achieve high performance and minimise manual cost of annotation by asking for informative unlabelled instances to be labelled by the human annotators. Then, the key problem becomes the evaluation of informativeness of unlabelled instances.

3.2. Informativeness measure

A confidence-based informativeness measure is used in this paper. In the case of dimensional SER, the confidence score attached to a speech instance should be the measure of how much we can trust this instance being correctly predicted to a certain numeric value. However, the technology for calculating a confidence score of a speech classification task is considerably more mature in comparison to confidence value calculation for speech regression tasks. We thus prefer to calculate the confidence a speech instance has when its emotion is binary classified as either falling into a negative interval \([-1, 0]\) or a positive interval \([0, 1]\), rather than calculating its confidence when it is predicted to a certain real number, and refer to it as binary classification confidence score.

Specifically, in order to compute a binary-classification confidence-based informativeness, we first convert the continuous-valued label \(y_i, \text{continuous}\) of any initial labelled instance \(x_i\) from \(S_l\) into a binary label \(y_i, \text{binary}\) with a step function as follows:

\[
\begin{align*}
    y_i, \text{binary} = \begin{cases} 
        -1, & y_i, \text{continuous} \in [-1, 0] \\
        1, & y_i, \text{continuous} \in (0, 1].
    \end{cases}
\end{align*}
\]

(1)

After the above conversion, a binary labelled set \(S_b = \{(x_i, y_i, \text{binary})\}_{i=1}^m\) can be obtained from the original continuous labelled set \(S_l\). Second, we learn a binary classifier \(C\) making use of \(S_b\). Third, with \(C\), we can then binary classify the emotion of any unlabelled instance \(y_j\) from \(S_u\) = \{(y_j)\}_{j=1}^m\) to be negative or positive, and calculate its corresponding binary-classification confidence score \(\text{conf}_{2-class}(y_j)\). The way of its calculation depends on the classification method we use. In our case, we apply support vector machines (SVMs) for classification and compute its confidence scores by fitting a logistic model on its outputs. Formally, for a trained SVM model, \(\text{conf}_{2-class}(y_j)\) can be given as:

\[
\text{conf}_{2-class}(y_j) = \frac{1}{1 + e^{A(y_j) + B}},
\]

(2)

where \(f(y_j)\) is the output prediction of the SVM for the instance \(y_j\), and the parameters \(A\) and \(B\) need to be further estimated. Finally, the informativeness measure \(I\) of \(y_j\) can be obtained by:

\[
I(y_j) = -\left|\text{conf}_{2-class}(y_j) - 0.5\right|.
\]

(3)

Such definition makes the instances with confidence scores nearer to 0.5 to be considered as higher informative, which is more suitable for the binary classification situation in contrast to the widely used least certainty AL strategy.

3.3. Algorithms

We present the detailed approach of AL for dimensional SER in a pool-based and a stream-based sampling scenario in Tables 1
Table 1: AL for dimensional SER in the pool-based scenario.

<table>
<thead>
<tr>
<th>Input</th>
<th>a small set of continuously labelled speech $S_l$, a large pool of unlabelled speech $S_u$, an emotion predictor $M$ trained by $S_l$, selecting size $N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do</td>
<td>Convert $S_l$ to $S_l$ using Eq. (1) and train a binary classification model $C$; Binary classify all instances in $S_u$ using $C$ and calculate their binary-classification confidence scores using Eq. (2); Compute $I$ of each instance in $S_u$ using Eq. (3); Select $N$ instances with the biggest $I$s from $S_u$, and label them. Call the new labelled set $S_{new}$; $S_l = S_l \cup S_{new}$, $S_u = S_u - S_{new}$; Train emotion predictor $M$ using new data $S_l$;</td>
</tr>
<tr>
<td>Until</td>
<td>$S_u = \emptyset$</td>
</tr>
</tbody>
</table>

Table 2: AL for dimensional SER in the stream-based scenario.

<table>
<thead>
<tr>
<th>Input</th>
<th>a small set of continuously labelled speech $S_l$, a large stream of unlabelled speech $S_u$, an emotion predictor $M$ trained by $S_l$, buffer $B$, buffer size $N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do</td>
<td>Save the incoming $N$ instances from $S_u$ into the buffer $B$; Convert $S_l$ to $S_l$ using Eq. (1) and train a binary classification model $C$; Binary classify all instances in buffer $B$ using $C$ and calculate their binary-classification confidence scores using Eq. (2); Compute $I$ of each instance in $B$ using Eq. (3); Select half of the instances with larger $I$s from $B$, and label them. Call the new labelled set $S_{new}$; $S_l = S_l \cup S_{new}$, $S_u = S_u - S_{new}$; Train emotion predictor $M$ using the new $S_l$;</td>
</tr>
</tbody>
</table>
| Until | $S_u$ is interrupted, i.e., no new data comes in and 2, respectively. We can see that in the pool-based scenario, in order to select instances for labelling in each training round, we traverse all the remaining unlabelled instances in the pool. And for the stream-based scenario, we simulate it as a situation where ‘each day’ we obtain a fixed number of instances from a data stream and select only half of them for labelling. This buffer mechanism thus considers AL efficiency in real-world application with constant updates.

4. Database and Acoustic feature extraction

The audio data used in the AVEC 2012 competition [22] is chosen for our experiments. It is a part of the SEMAINE corpus [9] which is freely available for scientific research purposes. It includes the conversations between participants and four emotionally stereotyped characters at 48 kHz with 24 bits per sample. The original data were split into three partitions: a training, development, and test partition, and only the labels of the latter two partitions were published during the ongoing challenge. This is why we finally choose the combination of the former two partitions as our experimental data.

The audio consists of 15 recordings (36,494 words) of 15 different speakers (6 males and 9 females). All recordings together last 5:16:55 (h:m:s) in total. Continuous-valued labels are provided at word level by continuous time and continuous value ratings for the four dimensions arousal, expectancy, power, and valence. In our experiments, the initial training set of labelled instances, which is used to train the initial dimensional SER model consists of 2,000 words. The additional set of instances which are used as candidates for labelling consists of 28,000 words. The test data consists of 6,494 words.

The audio feature set used is the official baseline audio feature set of AVEC 2012 with a total of 1,841 features, extracted by our openSMILE toolkit [23]. This set consists of 25 energy-related and spectral low-level descriptors (LLDs) × 42 functionals, 6 voicing related LLDs × 32 functionals, 25 delta coefficients of the energy/spectral LLDs × 23 functionals, 6 delta coefficients of the voicing related LLDs × 19 functionals, and 10 voiced/unvoiced durational features. The 42 functionals consist of 23 statistical functionals, 4 regression functionals, 9 local minima/maxima related functionals, and other 6 functionals. For details on LLDs and functionals, please refer to [22].

5. Experiments and results

5.1. Experimental layout

In order to fully investigate the performance impact on the dimensional SER system as we actively select instances for model learning, two different experimental strategies were designed corresponding to two scenarios as outlined: pool-based and stream-based. However, they share five basic settings:

- For emotion prediction, support vector regression (SVR) with a radial basis function (RBF) kernel as implemented in the WEKA toolkit [24] is used.
- For binary classification, SVMs with a RBF kernel are used. Moreover, the classification confidence scores are calculated by fitting a logistic model to the output of the SVMs.
- All experiments are performed for valence and arousal.
- Performance is measured by correlation coefficient (CC), which has been the official competition measure of AVEC 2012.
- As the baseline result, ten rounds of passive learning (i.e., randomly selecting instances for annotation) are executed to reduce the impact of statistical effects.

Specifically, for the pool-based sampling scenario, we first trained the initial emotion predictor using the initial labelled set, and then incrementally selected and ‘labelled’ (i.e., unblinded the labels given in the set) instances from the candidate pool for model retraining as described in Table 1 in steps of 2,000 words. To simulate the stream-based scenario, we continuously selected 2,000 words from the transcription candidate set in a random way at a time. Each time, we select and ‘label’ half of them (1,000 words) for model retraining as described in Table 2. Additionally, ten rounds of stream-based AL processing are executed to reduce the impact of statistical effects.

5.2. Results and discussion

Learning curves for CC in the pool-based and stream-based scenarios are presented in Fig. 1, respectively. We plot the baseline passive learning curves and AL curves in terms of mean with a standard deviation bar.

From the top-row of Fig. 1, one can observe that the AL curves lie above the mean curve of ten passive learning rounds on both emotional dimensions in the pool-based scenario. Thus, the proposed AL approach successfully reduced the number of labelled instances required to reach a given CC level when compared to passive learning. Specifically, for valence prediction (cf. Fig. 1.(a)), the best performance with AL (CC = 0.359) is achieved when 93.3% of the candidate instances (28,000 words) are used. However, the best performance with passive learning (CC = 0.335) can only be reached using all instances of the pool (30,000 words). To achieve the same level of performance, 12.0% less data (3,600 words) is needed in the context of AL learning. A similar situation can also be observed for the arousal.
Table 3: Numbers of human-labelled instances needed to reach passive learning’s best performance (CC= .355/.327 for valence and CC=.499/.479 for arousal in the pool/stream-based scenario) by adopting active learning (AL) and passive learning (PL).

<table>
<thead>
<tr>
<th>Learning method</th>
<th>Pool-based scenario</th>
<th>Stream-based scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Valence</td>
<td>Arousal</td>
</tr>
<tr>
<td>AL</td>
<td>26 400 / 88.0 %</td>
<td>14 200 / 88.8 %</td>
</tr>
<tr>
<td>PL</td>
<td>30 000 / 100.0 %</td>
<td>16 000 / 100.0 %</td>
</tr>
</tbody>
</table>

Figure 1: Pool− (top-row) and stream-based (bottom-row) learning curves for CC by valence (left column) and arousal (right column). Abbreviations are AL: Active Learning, PL: Passive Learning.

6. Conclusions

We presented an efficient Active Learning approach to reduce manual efforts of labelling in dimensional SER. It can be applied to many regression tasks where data may be abundant but labels are scarce or expensive to obtain. The core of our proposed AL approach is to treat the instance, which is hardest to be correctly classified binary, to be informative. Then, by selectively labelling the instances with highest informativeness, a given performance index can be reached with less training data. For experimental verification, two scenarios were constructed to simulate different types of learning problems in a real-world scenario:

1) For the pool-based processing, i.e., consideration of all training instances at once, the amount of training data required for a certain performance on a test set was reduced by 12.0% and by 11.7% for valence and arousal prediction, respectively, in comparison to the amount of data required for passive learning. Moreover, the best performance for AL is obtained when 93.3% of the data are used. In brief, while our proposed AL approach not only saves cost and time of labelling, it also improves the overall performance compared to training with all labelled data.

2) In stream-based processing, i.e., new training material ‘comes in’ gradually – such as in the case when updating a server after collecting speech data from several clients – 11.2% and 6.2% less data are needed for AL to achieve the best passive learning performance for valence and arousal prediction, respectively, as compared to a passive learning approach.

Future efforts may consider other forms of quantisation than the chosen binary one followed here. Further, meaningful confidence measures such as the ones introduced in [25, 26] can be used in exchange for the classification-based estimation. Finally, active class selection could be tailored for regression, to indicate emotion of which continuous value to synthesise [27], if synthesised speech is used in training [28].

7. Acknowledgements

The authors would like to thank the National Natural Science Foundation of China (61171186, 61271345), the Key Laboratory Opening Funding of the MOE-Microsoft Key Laboratory of Natural Language Processing and Speech (HIT.KLOF.2011XXX), and the Fundamental Research Funds for the Central Universities (HIT.NSRIF.2012047).
8. References


