Active Learning by Sparse Instance Tracking and Classifier Confidence in Acoustic Emotion Recognition

Zixing Zhang and Björn Schuller

Institute for Human-Machine Communication, Technische Universität München, Germany
{zixing.zhang|schuller}@tum.de

Abstract

Data scarcity is an ever crucial problem in the field of acoustic emotion recognition. How to get the most informative data from a huge amount of data by least human work and at the same time to obtain the highest performance is quite important. In this paper, we propose and investigate two active learning strategies in acoustic emotion recognition: Based on sparse instances or based on classifier confidence scores. The first strategy focuses on the problem of unbalanced binary or multiple classes. The latter strategy pays more attention on clearing up the boundary confusion between different classes. Our experimental results show that by using active learning aiming at sparse instances or based on classifier confidence, the amount of transcribed data needed is significantly reduced and the unweigted accuracy boosts greatly as well.

Index Terms: Active Learning, Acoustic Emotion Recognition, Sparse Instances, Confidence Scores

1. Introduction

The issue of data scarcity seriously impacts the affect recognition performance and restrains its further practical application. Semi-supervised and unsupervised learning are promising approaches to remedy the issue of data scarcity. A semi-supervised learning method is evaluated in [1], which demonstrates a significant accuracy enhancement when adding automatically transcribed samples – with high classifier confidence – from an unlabelled set to the original training set.

As for supervised learning, a common approach is passive learning (PL), where samples are randomly and independently selected from the underlying distribution, and then annotated by human experts manually. A crucial problem for this approach is that this process is extremely time-consuming and costly. It is well-known that data collection, cleaning, and annotation consume about 80 percent of the effort in a typical data mining project [2].

Active learning (AL) – another supervised learning approach – aims to minimize the amount of human supervision required and maximize the performance given transcribed and untranscribed data. The goal thereby is to identify the "most informative" samples in the unlabelled data – i. e., those that we would gain most by, if they were manually labelled – and then present only these sample to human labellers. Several approaches have been investigated for selecting the most informative samples [3]. A well known method is uncertainty-based AL in which the active learner determines the certanties of the predictions on the unlabelled data based on posterior probabilities. The samples with least certainty are generally presented to the labellers for annotation. This method is well established in automatic speech recognition [4] and information extraction [5], for example. Another common AL strategy is the committee-based method which utilizes multiple classifiers and is investigated in [6, 7] for text categorization. Predictions for unlabelled data are made by multiple classifiers. The samples considered as most informative are those with the lowest agreement. Other AL methods include the expected-error-reduction method, which aims to measure how much its generalization error is likely to be reduced [8], the expected-model-change-based method which chooses the instances that impact the current model most greatly [9], and the diversity-density-related method which aims to maximize the learning benefits of relevance feedback on retrieving documents [10]. A major drawback of the above methods is that they ignore the problem of class unbalance or the issue of sparsity of certain classes. In this paper, we propose two novel AL strategies for binary acoustic emotion recognition based on sparse instances and on medium confidence scores. Sparse-instances-based AL (SI-AL) selects the samples "likely to be" the fewest ones to be annotated manually from the candidate data in the pool. It is very useful to boost the unweighted accuracy in an unbalanced scenario as usually encountered when dealing with realistic emotion data. Medium-confidence-scores-based AL (MCS-AL), instead of the widely used least confidence AL, selects the samples with a confidence close to 0.5, if the confidences are normalised to the range 0 to 1. An upsampling strategy is implemented for dealing with the class unbalance problem. This is beneficial for unlabelled class data.

In the following, Section 2 and 3 introduce the selected database, the acoustic feature set and the classifier used. Section 4 describes the algorithms of two proposed AL approaches based on sparse instances and based on medium confidence scores. Results are shown in Section 5. Finally, the paper is summarized and an outlook on future work is given in Section 6.

2. Database

Recently, the more diverse emotions covered in spontaneous databases with larger amounts of instances (up to 10 k and more) of more subjects (> 50) and annotated by more labellers are attracting more and more interest in the field of emotion recognition. Trying to meet this requirement, the FAU AIBO Emotion Corpus is chosen for this paper as in [11].

This corpus deals with recording of children interacting with Sony's pet robot Aibo. In such a recording environment, the children were led to believe that the Aibo would be responding to their commands, which sparked Aibo producing series of fixed and predetermined behaviours. In the realistic communication scenario, however, the Aibo sometimes disobeyed the commands, which stimulated the children various of emotional reactions. So, German speech spoken from the children was spontaneous and coloured by emotion. These experiments were executed at two different schools of MONT and OHM, involving 51 children with

21 males and 30 females, age covering from 10 to 13. Finally, 9.2 hours of speech recording without pauses was obtained by using a DAT-recorder (16 bit, 48 kHz down-sampled to 16 kHz) and high quality wireless head set.

After that, the recordings were segmented into turns that are annotated by 5 advanced students of linguistic independently from each other. Then the turns were divided into chunks on the word levels.

For our experiments, the whole corpus consisting of 18 216 chunks and the 2-class labelling from the INTERSPEECH 2009 Emotion Challenge [11] is used: **NEG**ative (subsuming $angry$, $touchy$, $reprimanding$, and $emphatic$) and **IDL**e (consisting of all nonnegative states). Further, speaker independence is guaranteed by using the data recorded at one school (OHM, 13 male, 13 female) as initial training set and pool – serving as a large set of candidate data requiring for labelling, and the data recorded at another school (MONT, 8 male, 17 female) as test set.

Table 1: FAU AIBO 2-class task. Pool: served as large set of candidate data requiring for labelling

#	**NEG**	**IDL**	Σ
Train	156	344	500
Pool	3 202	6 257	9 459
Test	2 465	5 792	8 257
Σ	5 823	12 393	**18 216**

3. Acoustic Features and Classifier

The acoustic feature set from the INTERSPEECH 2009 Emotion Challenge feature [11] contains 384 features resulting as a systematic combination of 16 low-level-descriptors (LLDs) and corresponding first order delta coefficients with 12 functionals. The features are extracted with our feature extraction toolkit openSMILE [12]. For the details of LLDs and functionals, please refer to [11].

As classifier we implement Support Vector Machines (SVM) which have solid mathematical and statistical foundation [5] and excellent performance in emotion recognition [1]. Thus, for representative results in our experiments, we chose SVM with linear kernel and complexity of 0.05, and pairwise multi-class discrimination Sequential Minimal Optimization (SMO). The WEKA toolkit is used for keeping in-line with the INTERSPEECH 2009 Emotion Challenge baseline.

4. Active Learning

Compared with PL, AL repeatedly queries untranscribed data and selects the most informative samples to annotate manually, and updates its learned rules. In the following, two proposed AL strategies are described.

4.1. Sparse-instances-based active learning

The algorithm of SI-AL is described as follows: at the initial step, a small set of labelled data D_l (500 utterances in this paper) is employed, and a pool of data D_u (9 459 utterances in this paper) serves as unlabelled data. By testing the pool data, K samples which are predicted as the sparse class 'NEG' will be selected randomly as D_k. We then add these selected data D_k with their original transcribed labels (to simulate additional manual labelling) into the training set D_l as a new training set $D_l + D_k$. We iterate this process until there is no data in the pool predicted automatically as 'NEG', since there will be less and less data in the pool predicted as 'NEG' when iteration continues. This strategy bases on the idea that by adding the 'likely to be NEG' samples, the acoustic model can improve on this class more and more since the initial training set includes generally less information for 'NEG' for its scarcity.

The following summarises the SI-AL algorithm:

Algorithm 1: SI-AL
Given:

- a small set of labelled data D_l
- a pool of unlabelled data D_u

Repeat:

- Train an acoustic model *M* on D_l
- Use *M* to test on D_u
- *Randomly select K samples from D_u that are predicted as belonging to the sparse class ('NEG' in our case) D_k above*
- Ask human experts to label these
- Add the labelled examples into the training data $D_l + D_k$, and delete them from the pool D_u / D_k

End: A defined number of data are annotated manually or *M* achieves a certain performance or there is no data in the pool predicted as belonging to the sparse class ('NEG' in our case)

4.2. Medium-confidence-scores-based active learning

In contrast to SI-AL which focuses on solving the instance sparsity problem, MCS-AL focuses on identifying and then annotating uncertain samples, as these may provide most information for the classification model. In contrast to the widely used least certainty AL, MCS-AL is most suitable for binary class recognition, because a low confidence for one class means a high confidence for another class for two-class recognition tasks. Hence, only the samples which have been assigned a medium confidence score are selected for additional labelling.

In this proposed approach, the query function is defined as:

$$Q(x) = \begin{cases} 1, & \text{if } c1 \leqslant \textit{confidence\_score (x)} \leqslant c2 \\ 0, & otherwise \end{cases} \quad (1)$$

where $confidence\_score(x)$ evaluates the posterior probability of sample x. If the posterior probability are normalised to the range 0 to 1, the classifier has the highest confidence to predict if the value equals 1, and has the lowest confidence to predict if the value equals 0. $c1$ and $c2$ are chosen close to 0.5, and are the lower boundary and the higher boundary of confidence scores for the selected samples, respectively. The exact values of $c1$ and $c2$ are determined by the given training set and given pool.

For our experiment, we iterate this process until all data from the pool has been selected into the training set. In this approach, upsampling is optional for dealing with the class imbalance problem.

The details of this algorithm are as follows:

Algorithm 2: MCS-AL
Given:

- a small set of labelled data D_l
- a pool of unlabelled data D_u

Repeat:

- (optional) Upsample training data to even class distribution D_l^s
- Train an acoustic model *M* on D_l^s
- Use *M* to test D_u
- *Based on the SVM prediction confidence scores*, rank the data
- Select *K* samples 'out of the middle' of the ranking D_k
- Ask human experts to label these
- Add the labelled examples into the training data $D_l + D_k$, and delete them out of the pool D_u/D_k

End: A defined number of data are annotated manually or *M* achieves a certain performance

5. Experimental Results

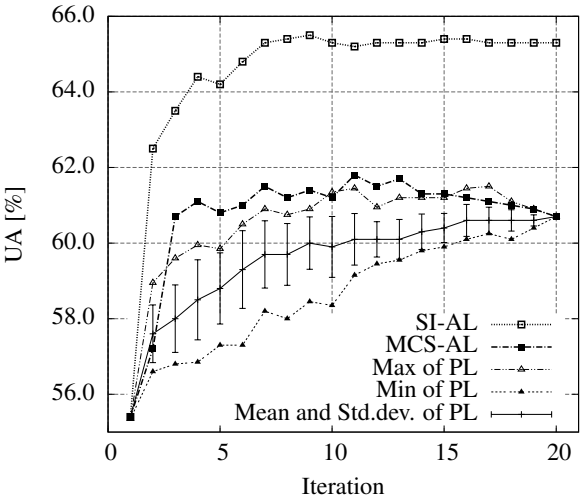
In the experiments, we evaluated two proposed active learning strategies: based on sparse instances and based on medium confidence scores. In order to evaluate the performance, both of AL and PL use the same 500 initial utterances set to train the acoustic model. For each iteration, the number of selected samples K is set equal to 500 instances unless there are not enough data to be selected from the pool. In this case, an accordingly lower K is chosen. As baseline result of PL, 10 rounds of the whole iterations process were executed to reduce the impact of statistical effects. Furthermore, the average unweighted accuracy (UA) along with the standard deviation, maximum, and minimum UA are represented (cf. Figure 1).

As to the performance evaluation measure, we chose UA, the sum of the recalls of the 'NEG' and 'IDL' classes divided by two, which has been the official competition measure of the first of its kind INTERSPEECH 2009 Emotion Challenge [11].

Table 2: Active learning based on sparse instances: Recall of binary IDL/NEG and the number of samples selected from the pool for each of the first 10 iterations.

		Recall [%]	
Iteration	Nr.	IDL	NEG
1	500	96.2	14.5
2	500	86.3	38.7
3	500	81.6	45.3
4	500	81.8	46.9
5	500	81.7	46.6
6	462	80.9	48.6
7	66	80.9	49.7
8	37	81.0	49.8
9	28	81.1	49.9
10	8	81.0	49.6

Figure 1: Unweighted accuracy vs. number of iterations. Unbalanced training set. Comparison between passive learning (PL) in 10 independent runs and active learning (AL) based on medium confidence scores (MCS-AL) and sparse instances (SI-AL) for binary NEG/IDL classification on the FAU Aibo Emotion Corpus.



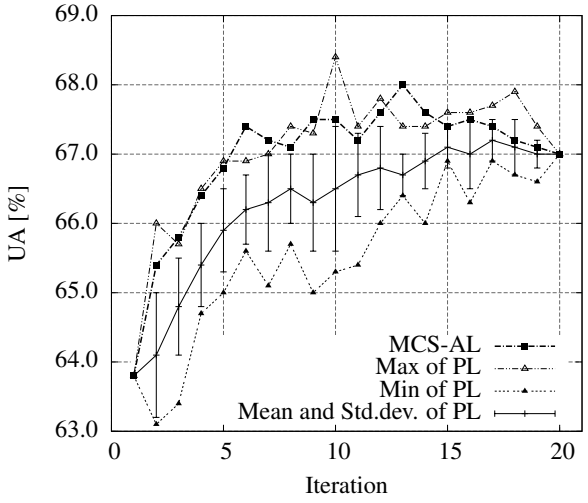
5.1. AL based on Sparse Instances

For SI-AL, we select K samples randomly from the subset including all of the instances which are predicted as 'NEG', whose instances number is far less than 'IDL' (cf. Table 1). These selected samples actually include the correctly assigned 'NEG' instances and falsely assigned 'IDL' instances. Then these samples with own originally correct labels 'NEG' or 'IDL' are added to the former training set simulating human labelling to form a new training set for the next iteration.

Table 2 displays the performance of SI-AL for the first ten iterations. From this table, we can see that, even though the recall slightly decreases for 'IDL', it boosts greatly for 'NEG' from 14.5 % to 49.7 % by the training set size growing up during the first few iterations, especially during the first two ones. After that, it remains quite stable, even if one adds more instances. This can be expected: SI-AL accelerates the learning rate for selecting the "right" instances within first few iterations, which helps to improve the acoustic model rapidly for the sparse class – 'NEG'.

Apart from that, one should also notice that the performance of 'NEG' and 'IDL' tends to be stable after the seventh iteration. That is because in the pool, there are fewer and fewer instances predicted as 'NEG' by selecting such instances again and again. Figure 1 depicts the UA comparison without upsampling strategy between PL and AL. For MCS-AL, a detailed discussion will be given in the next section. In Figure 2, it can be seen as well that the UA boots significantly within the first few iterations for its correctly selecting the most informative samples firstly. Further, it clearly shows that the best performance of AL based on sparse instances achieves 65.5 % UA by using only a small amount of training data (3 565 utterances), which significantly overtakes the best PL UA of 60.7 % using all training data (9 959 utterances). In other words, the sparse-instances-based selective sampling method boosts UA with roughly 5 % absolutely and reduces transcribed training data by 64.2 % compared to random instance selection without the application of upsampling.

Figure 2: Unweighted accuracy vs. number of iterations. Balanced training set. Comparison between passive learning (PL) in 10 independent runs and active learning (AL) based on medium confidence scores (MCS-AL) for binary NEG/IDL classification on the FAU Aibo Emotion Corpus.



5.2. AL based on Medium Confidence Scores

For MCS-AL, all the data in the pool are sorted by the prediction confidence scores in each iteration. Only a certain number of instances with medium confidence scores are included in the training set for the next iteration. In Figure 1, the dot-dash line with black solid squares represents the UA of MCS-AL. This curve is much higher than the curve of the mean UA and also higher than the one of maximum UA as achieved by random sampling. The best UA in this curve is 61.8 % which is 1.1 % better than the mean UA reached by usage of all data, and 1.7 % point better than using the same amount of data with random selection of instances. Furthermore, we obtain the best UA by using 85.0 % less transcribed data with MCS-AL as compared to random selection of instances. Further note that, the curve for MCS-AL is found above the one obtained by SI-AL.

In order to enhance baseline performance and cope with cases of unbalanced classes, we apply an upsampling strategy within which unbalanced binary classes are randomly upsampled to an approximately uniform distribution before classification in each iteration step. Figure 2 shows the results of this experiment. Based on the according plot, one observes that MCS-based selective sampling again greatly outperforms random sampling of instances: The best UA of 68.0 % is achieved by MCS-AL outperforming random sampling by 1.0 % UA absolute when the random sampling strategy uses all training data, and by 1.3 % UA absolute when the random sampling strategy uses the same amount of data as the MCS-based selective sampling uses. Note that the latter result is significant at the common 5 % level in a one-sided z-test.

6. Conclusions

In this paper, we proposed and evaluated two novel active learning strategies for binary acoustic emotion recognition tasks based on tracking of sparse instances of the minority class, and iterative re-training and labelling of data based on medium level confidence scores. The results show that both of the suggested strategies can efficiently minimize the amount of human-labelled training data required. This saves time and cost when collecting new data sets. By SI-AL, the amount of labelled training data required for a fixed unweighted accuracy on a test set is reduced by 64.2 % in comparison to the amount of data required for passive learning. Moreover, the unweighted accuracy improves by approximately 5 % absolute in the best case. By MCS-AL, the amount of labelled data required is reduced even further by 85.0 % and by 75.0 %, respectively, without and with balancing of the training instances, compared to passive learning with the same resulting UA on the test set. This result is beneficial for real-world application of automatic emotion recognition such as anger detection, where large unlabelled sets can be easily collected in automated ways, but labelling is expensive and time consuming. While both presented methods not only save costs and time for labelling, they also improve the overall performance compared to training with all labelled data.

Our future work will focus on analysing both of the two active learning strategies with various initial training set sizes and unlabelled data pool sizes to investigate its robustness wrt. to these parameters. Furthermore, combining both confidence based and sparse-instances based strategies for active learning seems a promising future avenue.

7. Acknowledgement

Zixing Zhang acknowledges funding from the Chinese Research Council.

8. References

[1] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised Learning in Cross-Corpus Acoustic Emotion Recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Big Island, HY, USA, 2011, pp. 523–528.

[2] D. Braha, Ed., *Data Mining for Design and Manufacturing: Methods and Applications*. Kluwer Academic, 2001.

[3] B. Settles, "Active learning literature survey," Computer Sciences Technical Report, University of Wisconsin-Madison, Tech. Rep., 2009.

[4] G. Riccardi and D. Hakkani-Tur, "Active learning: theory and applications to automatic speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 504–511, 2005.

[5] M. Li and I. Sethi, "Confidence-based active learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 8, pp. 1251–1261, Aug. 2006.

[6] R. Liere, "Active learning with committees: An approach to efficient learning in text categorization using linear threshold algorithms," Ph.D. dissertation, Oregon State Univ., Portland, 2000.

[7] A. McCallum and K. Nigam, "Employing em in pool-based active learning for text classification," in *Proc. Int'l. Conf. on Machine Learning (ICML)*, Madison, WI, July 1998, pp. 359–367.

[8] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," in *Prof. Int'l. Conf. on Machine Learning (ICML)*, Williamstown, MA, June 2001, pp. 441–448.

[9] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *Proc. Empirical Methods in Natural Language Processing(EMNLP)*, Honolulu, Hawaii, USA, October 2008, pp. 1070–1079.

[10] Z. Xu, R. Akella, and Y. Zhang, "Incorporating diversity and density in active learning for relevance feedback," in *Proc. European Conference on Information Retrieval (ECIR)*, Rome, Italy, April 2007, pp. 246–257.

[11] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. INTERSPEECH*, Brighton, UK, 2009, pp. 312–315.

[12] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia (MM)*, Florence, 2010, pp. 1459–1462.