Dynamic Active Learning Based on Agreement and Applied to Emotion Recognition in Spoken Interactions

Yue Zhang
Department of Computing
Imperial College London
London, U.K.
yue.zhang1@imperial.ac.uk

Eduardo Coutinho
Department of Computing
Imperial College London
London, U.K.
e.coutinho@imperial.ac.uk

Zixing Zhang
Chair of Complex & Intelligent Systems
University of Passau
Passau, Germany
Zixing.Zhang@uni-passau.de

Caijiao Quan
Machine Intelligence & Signal Processing group, MMK
Technische Universität München
Munich, Germany

Björn Schuller
Department of Computing
Imperial College London
London, U.K.
bjoern.schuller@imperial.ac.uk

ABSTRACT

In this contribution, we propose a novel method for Active Learning (AL) - *Dynamic Active Learning (DAL)* - which targets the reduction of the costly human labelling work necessary for modelling subjective tasks such as emotion recognition in spoken interactions. The method implements an adaptive query strategy that minimises the amount of human labelling work by deciding for each instance whether it should automatically be labelled by machine or manually by human, as well as how many human annotators are required. Extensive experiments on standardised test-beds show that DAL significantly improves the efficiency of conventional AL. In particular, DAL achieves the same classification accuracy obtained with AL with up to 79.17 % less human annotation effort.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Affective Computing and Human-Machine-Interaction

General Terms

Algorithm, Experimentation

Keywords

Active Learning; Adaptive Query Strategy; Speech Emotion Recognition; Acoustics

ICMI 2015, November 9–13, 2015, Seattle, WA, USA.

DOI: http://dx.doi.org/10.1145/2818346.2820774.

1. INTRODUCTION

In recent years, research on recognition of human emotion from natural speech has received increasing attention due to its manifold applications in human-machine communication, human-robot communication, and multimedia retrieval. In this light, one of the major barriers is the scarcity of annotated realistic speech data, which are of paramount importance for subjective tasks [15, 14], but are time-consuming and expensive to obtain. To overcome this limitation, a plethora of approaches in machine learning have been proposed for the exploitation of unlabelled data, which is nowadays pervasive in digital format and relatively easy and inexpensive to acquire (e. g., from public resources such as social media). Among others, the most prevailing approaches are Semi-Supervised Learning (SSL) [18], Active Learning (AL) [9] and various combinations thereof [12, 19]. AL algorithms aim at improving a model's performance by 'actively' choosing the 'most informative' instances from a large pool of unlabelled data, and passing them to human oracles for labelling. There are various measures by which the informativeness of unlabelled samples can be accessed (commonly referred to as *query strategies* [9]). One of the easiest methods is to determine the uncertainty of the predictions on unlabelled data based on a previously trained model (uncertainty sampling), and then query an annotator for labelling the instances with the least certain classification [17]. Other query strategies evaluating the informativeness of unlabelled instances include the query-by-committee method, which compares multiple competing models for the same task [4]; the expected-error-reduction method, which aims to measure how much its generalization error is likely to be reduced [6]; the expected-model-change-based method, which selects those instances that have a greater impact on the current model [10]; and the diversity-density-related method, which aims to maximise the learning benefits of relevance feedback on retrieving documents [13]. One major drawback regarding these conventional AL algorithms is that a fixed number of human annotators is set for all instances to be labelled (here-

inafter referred to as 'Static' AL or SAL). As a consequence of this constraint, SAL still requires a considerable amount of human labelling work and thus can be impractical in many cases. This fact motivates us to propose a novel avenue of research for AL algorithms – Dynamic Active Learning (DAL) – that shifts the perspective from standard majority voting procedures to an agreement based annotation strategy. The core underlying principle is simple – instead of requesting all available annotators and then forming the majority of their votes, we adapt the number of annotators for each instance to a certain task-specific agreement level (i. e., a certain number of votes for a common category, e. g., class label). In this way, the DAL approach leads to the same gold standard as with majority voting while omitting dispensable annotations, thus dramatically reducing the annotation effort.

The rest of this paper is organised as follows: In Section 2, we describe in detail the DAL algorithm. Then, we introduce the database and the feature set used to evaluate the DAL algorithm in Sections 3 and 4, respectively. The experimental settings and results are presented in Sections 5.1 and 5.2. Finally, in Section 6 we discuss our findings and impulses for future work.

2. METHODOLOGY

A common and straightforward decision rule in SAL is majority voting among multiple raters, who are considered equally reliable. It is evident that querying a fixed number of annotators for each instance as being implemented in SAL is a rather ineffective method. The DAL approach aims to avoid this disadvantage by following the intuitive assumption that in the context of subjective tasks, the higher the inter-rater agreement is, the more established the resulting gold-standard label is.

2.1 SVM and Prediction Uncertainty

Similar to conventional AL, the dynamic active learner actively selects the data from which it learns by considering the prediction uncertainty of the trained classifier in terms of confidence values. For this purpose, we apply Support Vector Machines (SVMs) that construct decision hyperplanes to separate instances of different classes, while maximising the functional margin. For each instance, the output distances to the decision boundaries are then transformed into probability values through a parametric method of logistic regression [5]. For the selection of the instances to be annotated at each iteration, we use a medium certainty query strategy, which has the potential advantage of avoiding the selection of noisy instances, as shown in [16].

2.2 Agreement Levels

Given the number n of annotators who are available for labelling a specific database, we define the *agreement level* as the minimum number of raters agreeing on one common category. Accordingly, $j \in \{1, ..., \lfloor \frac{n+1}{2} \rfloor\}$, with $j, n \in \mathbb{N}$, agreement levels can be selected. For the upper limit of the interval, the floor is considered with regard to even numbers of annotators. Specifically, $n' \in \{j, ..., 2j-1\}, n' \in \mathbb{N}$ raters might be needed until a certain agreement level j is achieved. The SAL performance that is achieved through majority voting among all n raters is defined as the baseline of our experiments.

2.3 Algorithms and Data Structure

For the applied algorithm, we define the following notations: $\mathcal{L} = ([\mathbf{x}_1, y_1], \ldots, [\mathbf{x}_l, y_l]), i = 1, 2, ..., l$, denotes a small set of labelled training data, where $\mathbf{x}_i$ is a d-dimensional feature vector, and y_i is the assigned emotion-related label. Additionally, a large pool of unlabelled data $\mathcal{U} = (\mathbf{x}'_1, \ldots, \mathbf{x}'_u), k = 1, 2, ...u$, exists where $u \gg l$ and $\mathbf{x}'_k$ is a d-dimensional feature vector. The number of votes for a specific class label y' that is manually assigned to an example instance $\mathbf{x}' \in \mathcal{N}_a$ is named v'. Figure 1 shows the pseudo-code description of the DAL algorithm. The learning process starts by training a model on the labelled data $\mathcal{L}$ and subsequently using this model to classify all instances of the unlabelled data pool $\mathcal{U}$. According to the medium certainty query strategy, a subset $\mathcal{N}_a \subset \mathcal{U}$ is selected and submitted to human annotation. The sequential process is repeated until a certain number of instances are annotated. The main improvement compared to the SAL method is presented in the fifth item. The stopping criterion for manual labelling of each instance is fulfilled when the predefined agreement level for a specific task has been achieved.

Algorithm: *Dynamic Active Learning (DAL)*
Repeat:

1. (Optional) Upsample the training set $\mathcal{L}$ to obtain even class distribution $\mathcal{L}_D$
2. Use $\mathcal{L}/\mathcal{L}_D$ to train a classifier $\mathcal{H}$, and then classify the unlabelled data set $\mathcal{U}$
3. Rank the data based on the prediction confidence values C and store them in a queue
4. Select a subset $\mathcal{N}_a$ with medium certainty
5. **For** each instance $\mathbf{x}'$ in $\mathcal{N}_a$
   (a) Randomise the query order of raters
   (b) Submit $\mathbf{x}'$ to the first j raters
   (c) If $v' = j$; **STOP**
   else **repeat**: select one rater for annotation
   until agreement level j is achieved
   (d) Assign y' to $\mathbf{x}'$
6. Remove $\mathcal{N}_a$ from the unlabelled set $\mathcal{U}$, $\mathcal{U} = \mathcal{U} \setminus \mathcal{N}_a$
7. Add $\mathcal{N}_a$ to the labelled set $\mathcal{L}$, $\mathcal{L} = \mathcal{L} \cup \mathcal{N}_a$

Figure 1: Pseudocode description of the DAL algorithm based on the medium certainty strategy for a predefined agreement level j.

3. DATABASE

In our experiments, we use the FAU Aibo Emotion Corpus (AEC) [11] of the INTERSPEECH 2009 Emotion Challenge (IS09 EC) [8, 7]. The database contains spontaneous and emotionally coloured speech of children interacting with Sony's pet robot Aibo. The recordings were taken from 51 children (age 10-13, about 9.2 hours of speech without pauses) at two different schools, referred to as 'MONT' and 'OHM'. For binary classification, the emotional states are categorised into the classes **NEG**(agative) and **IDL**(e). The frequencies for the two-class problem are given in Table 1.

Table 1: Distribution of speakers and instances per partition of the FAU AEC dataset. M: male; F: female; NEG: negative emotions; IDL: neutral and positive emotions.

	# speakers		# instances per class		
FAU AEC	M	F	NEG	IDL	Σ
Pool	13	13	3 358	6 601	9 959
Validation	8	17	2 465	5 792	8 257
Σ	21	30	5 823	12 393	18 216

4. SELECTED ACOUSTIC FEATURES

The acoustic features used in our experiments are adopted from the baseline feature set of IS09 EC. This is created with the openSMILE framework [2, 1] by applying statistical functionals to frame-wise low-level-descriptors (LLDs). To each of the 16 LLDs, the delta coefficients are computed. Finally, the 12 functionals are applied on a per-chunk level. As result of the 'brute-forcing' method, the total feature vector per chunk contains $16 \times 2 \times 12 = 384$ attributes.

5. EXPERIMENTS AND RESULTS

In the following, we investigate the performance of the DAL algorithm by evaluating the classification accuracy in relation to the number of human annotations. The optimised results are compared with the SAL baseline performance.

5.1 Experimental Setup

For transparency and reproducibility, we used open-source classifier implementations of SVMs from the WEKA data mining toolkit [3]. As classifiers, we chose linear kernel SVMs trained with a complexity parameter C constant of 0.05 and with Sequential Minimal Optimization (SMO), as they are robust against over-fitting in high dimensional feature spaces. For initial training of the model, 200 instances were randomly selected from the training data, whereas the remaining instances were used as the unlabelled data pool. At each learning iteration, we selected a subset $\mathcal{N}_a$ comprising 200 instances to be submitted to manual annotation. The learning process stopped after 4 800 instances had been manually annotated, where the total number of human annotations differs in each experimental scenario. The training process was repeated 20 times with randomly generated initializations. As evaluation measure, we considered the unweighted average recall (UAR).

5.2 Discussion of Results

Table 2: Relative cost reduction (CR) measured by the number of human annotations and comparing DAL on agreement levels $j = 1, 2, 3$ with the SAL baseline at $\mathbf{UAR}_{max}$

	UAR_{max}	CR (%)
SAL	68.79	–
j=3	68.79	25.58
j=2	68.84	54.83
j=1	68.38	79.17

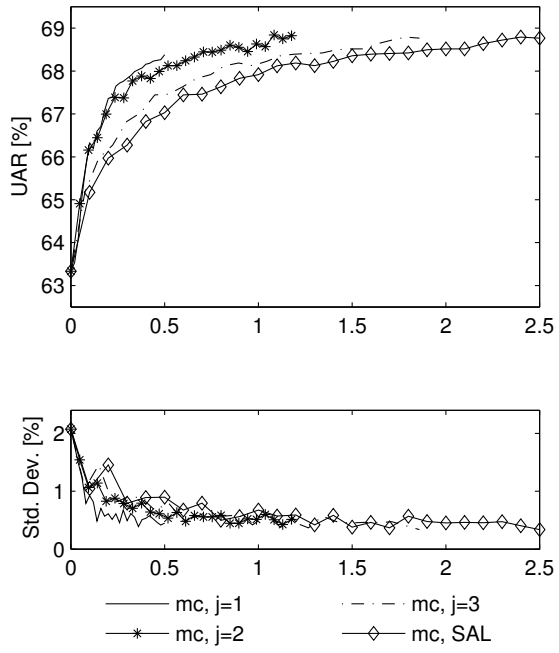

Figure 2: Dynamic Active Learning (DAL) vs Static Active Learning (SAL): the performance measures show the UAR values averaged across 20 runs of the algorithm and the respective standard deviations vs the number of human annotations.

Figure 2 shows the characteristic curve progression of AL: the sequential addition of human-labelled instances to the initial training set leads to continuous improvements in the performance of the classifier. More importantly, it can be clearly seen that all DAL curves corresponding to the different agreement levels are above the SAL baseline. This marked 'shrinking' effect demonstrates a dramatic cost reduction in the sense that less human annotations are required to achieve the same classification accuracy. In order to substantiate our findings, we compare the costs measured by the number of human annotations at the highest UAR (UAR_{max}) achieved by each method. According to Table 2, the relative cost reduction (CR) increases with lower agreement levels. Finally, the analysis of standard deviation shows that the stability of the model is enhanced during the learning process.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we introduced a novel approach for Dynamic Active Learning that allows significant reduction of the costly human labelling work by adapting the number of human annotators for each instance to a predefined agreement level. In particular, our results demonstrate that the DAL method leads to the same performance of the trained model, but requires up to 79.17 % less human annotations with the medium certainty (*mc*)query strategy. For future research, we will investigate the robustness of the DAL method by conducting experiments with multiple corpora, different feature sets, and varying amount of initial training instances.

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union's Framework Programme for Research and Innovation HORIZON 2020 under the Grant No. 645378 (ARIA-VALUSPA) and the European Union's Seventh Framework Programme under the ERC Starting Grant No. 338164 (iHEARu).

8. REFERENCES

[1] F. Eyben, F. Weninger, F. Groß, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proc. of ACM MM*, pages 835–838, Barcelona, Spain, 2013.

[2] F. Eyben, M. Wöllmer, and B. Schuller. openSMILE – the Munich versatile and fast open-source audio feature extractor. In *Proc. of ACM MM*, pages 1459–1462, Florence, Italy, 2010.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[4] R. Liere and P. Tadepalli. Active learning with committees for text categorization. In *Proc. of AAAI/IAAI*, pages 591–596, Providence, RI, 1997.

[5] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in large margin classifiers*, pages 61–74. MIT Press, Cambridge, MA, 1999.

[6] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proc. of ICML*, pages 441–448, Williamstown, MA, 2001.

[7] B. Schuller. The computational paralinguistics challenge. *IEEE Signal Processing Magazine*, 29(4):97–101, 2012.

[8] B. Schuller, S. Steidl, and A. Batliner. The INTERSPEECH 2009 emotion challenge. In *Proc. of INTERSPEECH*, pages 312–315, Brighton, UK, 2009.

[9] B. Settles. Active learning literature survey. Technical report, Department of Computer Sciences, University of Wisconsin–Madison, Wisconsin, WI, 2009.

[10] B. Settles and M. Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proc. of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1070–1079, Honolulu, HY, 2008.

[11] S. Steidl. *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Logos Verlag, Berlin, 2009.

[12] G. Tur, D. Hakkani-Tür, and R. E. Schapire. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45(2):171–186, 2005.

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

[14] Y. Zhang, E. Coutinho, Z. Zhang, M. Adam, and B. Schuller. On Rater Reliability and Correlation Based Dynamic Active Learning. In *Proc. 6th biannual Conference on Affective Computing and Intelligent Interaction (ACII 2015)*, Xi'an, P. R. China, 2015. AAAC, IEEE. 7 pages.

[15] Y. Zhang, E. Coutinho, Z. Zhang, C. Quan, and B. Schuller. Agreement-based Dynamic Active Learning with Least and Medium Certainty Query Strategy. In *Proc. Advances in Active Learning : Bridging Theory and Practice Workshop held in conjunction with the 32nd International Conference on Machine Learning, ICML 2015*, Lille, France, 2015. International Machine Learning Society, IMLS. 5 pages.

[16] Z. Zhang and B. Schuller. Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition. In *Proc. of INTERSPEECH*, Portland, OR, 2012. 4 pages.

[17] J. Zhu, H. Wang, B. K. Tsou, and M. Ma. Active learning with sampling by uncertainty and density for data annotations. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1323 –1331, 2010.

[18] X. Zhu. Semi-supervised learning literature survey. Technical Report TR 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2006.

[19] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. In *ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*, pages 58–65, Washington DC, 2003.