Agreement-based Dynamic Active Learning with Least and Medium Certainty Query Strategies

Yue Zhang YUE.ZHANG1@IMPERIAL.AC.UK
Department of Computing, Imperial College London, London, United Kingdom

Eduardo Coutinho E.COUTINHO@IMPERIAL.AC.UK
Department of Computing, Imperial College London, London, United Kingdom

Zixing Zhang ZIXING.ZHANG@UNI-PASSAU.DE
Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany

Caijiao Quan CAIJIAO.QUAN@TUM.DE
Machine Intelligence and Signal Processing Group, Technische Universität München, Munich, Germany

Björn Schuller BJOERN.SCHULLER@IMPERIAL.AC.UK
Department of Computing, Imperial College London, London, United Kingdom

Abstract

In this contribution, we propose a novel method for active learning termed 'dynamic active learning' or DAL for short, with the aim of ultimately reducing the costly human labelling work for subjective tasks such as speech emotion recognition. Through an adaptive query strategy, the amount of manual labelling work is minimised by deciding for each instance not only whether or not it should be annotated, but also dynamically on how many human annotators' opinions are needed. Through extensive experiments on standardised test-beds, we show that DAL achieves the same classification accuracy of 'traditional' AL with a cost reduction of up to 79.17 %. Thus, the DAL method significantly improves the efficiency of existing algorithms, setting a new benchmark for the utmost exploitation of unlabelled data.

1. Introduction

Within the context of Computational Paralinguistics, speech patterns can be characterised using objective and subjective measures (Schuller & Batliner, 2014). In the case of *objective* measures (e. g., age, gender, weight), the labels attributed to speech are referred to as the 'ground truth'. On the other hand, there are speech phenomena (e. g., voice likeability, attractiveness, interest) that can only be reliably assessed (annotated/labelled) by perceptive judgements (Steidl et al., 2005). In consequence, the reliability of labels for the subjective speech phenomena highly depends on the annotators' stable and transient characteristics, including a myriad of subjective factors (Steidl et al., 2005; Schuller, 2015). Therefore, in contrast to the 'ground truth' that can be measured objectively, subjective annotations lead to what is known as the 'gold standard', and are necessarily assessed by inter-rater agreement procedures. Thus, a large number of annotators is necessary to establish a well grounded reference. This leads to the fact that subjective tasks are particularly affected by the major barrier of today's research: scarceness of human annotated data, which are time-consuming and expensive to obtain. On the other hand, there is a vast resource of unlabelled data which is nowadays pervasive in digital format and relatively easy and inexpensive to collect, e. g., from public resources such as social media.

Following the belief "there is no data like more data", many researchers in the area of Machine Learning (ML) developed approaches in machine learning for the exploitation of unlabelled data. The most common methods include semi-supervised learning (SSL) (Zhu, 2006), active learning (AL) (Settles, 2009), and diverse combinations thereof (Tur et al., 2005; Zhu et al., 2003). The essence of the conventional ML methods is to train a classifier on a small, labelled data set, and re-train the model iteratively by sequentially adding new (machine or human) labelled instances to the

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

training set. The active learner aims at achieving greater accuracy with fewer training labels by (actively) choosing the data from which it learns, and querying human annotators for labelling. The main drawback of conventional AL algorithms is that they define a fixed number of human annotators for all selected instances (hereinafter referred to as 'Static Active Learning' (SAL)). As result of this constraint, the SAL algorithms still require a considerable amount of human annotations, which can easily be avoided through a shift in perspective from standard majority voting among multiple raters to an agreement based annotation strategy (Zhang et al., 2014). This motivates us to introduce the 'Dynamic Active Learning' (DAL) approach that uses an adaptive query strategy to minimise the amount of human labelling work without sacrificing performance. The core underlying idea is simple: instead of requesting all available raters and then forming a majority of their votes, we adapt the number of annotations for each instance to a predefined agreement level, i. e., a certain number of votes for a common category (e. g., class label). Among many application possibilities, the proposed DAL approach is targeted at optimising existing crowd-sourcing systems (e. g., Amazon Mechanical Turk (Kittur et al., 2008)) in which tasks are distributed to paid click-workers to complete (Howe, 2006; Yuen et al., 2011).

In this paper, we describe the methods used in DAL in Section 2. Then, we introduce the database and the feature set in Section 3 and Section 4. Experimental setup and results are presented in Section 5. We conclude by discussing our findings and extensions for future research in Section 6.

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 is a rather inefficient method. For instance, if there are five annotators available and the first three annotate a specific instance with the same label, the annotations of the other two annotators seem to be abundant. In the following, we detail the methods of DAL by providing the mathematical definitions of prediction uncertainty and agreement level, and describing the algorithms.

2.1. SVM and Confidence Measure

Similar to 'traditional' AL, we apply Support Vector Machines (SVMs) that construct decision hyperplanes to separate instances of different classes by using the decision function $f(\mathbf{x})$, while maximizing 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 (Platt, 1999). For binary classification, the sigmoid function with the parameters A and B is defined as:

$$P_1(\mathbf{x}) = \frac{1}{1 + \exp(Af(\mathbf{x}) + B)} \tag{1}$$

$$P_0(\mathbf{x}) = 1 - P_1(\mathbf{x}) \tag{2}$$

The confidence value for the predicted class is obtained by forming the difference of the posterior probabilities $P_0(\mathbf{x}), P_1(\mathbf{x})$ for classes '0' and '1', respectively:

$$C(\mathbf{x}) = |P_1(\mathbf{x}) - P_0(\mathbf{x})| \tag{3}$$

In addition to the least certainty (lc) query strategy adopted from 'traditional' AL, we consider a medium certainty (mc) query strategy (Zhang & Schuller, 2012) that has the potential advantage of avoiding the selection of noisy data, which can be caused by distortions of acoustic patterns (You et al., 2006), unreliable or ambiguous annotations (Grimm & Kroschel, 2005) as it is usually the case for acoustic emotion recognition tasks due to their subjective nature. Formally, the query function for mc is defined as:

$$\mathbf{x}_{mc} = \arg\min_{\mathbf{x}} |C(\mathbf{x}) - C_m|, \tag{4}$$

where $C(\mathbf{x})$ denotes the confidence value assigned to the predicted label of a given instance $\mathbf{x}$. The confidence values are ranked and stored in a queue (in descending order). Accordingly, C_m represents the confidence value of the instance located in the centre of the ranking queue.

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. In practice, j raters would be required simultaneously in the first round of queries to minimise the related time-consumption as j is the minimum number of ratings to achieve the respective agreement level. The SAL performance that is achieved through majority voting among all n raters is set to the baseline in 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 based on the mc and lc query strategies. 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 $\mathcal{U}$. Depending on which query strategy is implemented, a subset $\mathcal{N}_a \subset \mathcal{U}$ is selected and submitted to human annotation. The sequential process is repeated until a predefined number of instances are annotated. The main improvement compared to the SAL method is presented in the fifth item. In the proposed adaptive query strategy, the stopping criterion for manual labelling of each instance is fulfilled when a predefined agreement level for a specific task is 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 or least 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 $v' = j$; **STOPP**
   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 and least certainty strategies for a predefined agreement level j.

3. Database

In our experiments, we use the FAU Aibo Emotion Corpus (AEC) (Steidl, 2009) of the INTERSPEECH 2009 Emotion Challenge (IS09 EC) (Schuller et al., 2009; Schuller, 2012). The database consists of recordings of children interacting with Sony's pet robot Aibo, which performs a fixed, predetermined sequence of actions. The recordings were taken from 51 children at two different schools, referred to as 'MONT' and 'OHM'. Five labellers (advanced students of linguistics) annotated each word independently from each other as neutral (default) or as one of ten specific emotional states. Each instance corresponds to a manually defined chunk that consists of multiple words according to the syntactic-prosodic criteria. For binary classification, the 11-class labels are mapped onto two-class labels by defining states with negative valence (e. g., *angry*, *reprimanding*) as **NEG**(agative), and all other states as **IDL**(e). A heuristic approach is applied to map the labels from the word-level to the chunk-level for each of the five labellers, where a chunk is defined as NEG if it contains at least one word with negative valence. To define the gold standard for the baseline results, we resort to majority voting to combine the labels from all five labellers to one single label for each chunk. The frequencies for the two-class problem are given in Table 1. Speaker independence is guaranteed by using the speech samples of the school 'OHM' for training and the data of the other school 'MONT' for validation. Specifically, the training data referred to as 'Pool' contains both the labelled training set $\mathcal{L}$ and the unlabelled data pool $\mathcal{U}$.

Table 1. Distribution of speakers and instances per partition of the FAU AEC. 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. 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 (Eyben et al., 2010; 2013) 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. Thus, the total feature vector per chunk contains $16 \times 2 \times 12 = 384$ attributes.

5. Experiments and Results

In this section, we investigate the performance of the DAL algorithm by evaluating the classification accuracy in relation to the number of manually annotated instances with regard to different agreement levels and query strategies. The optimised results are compared with the baseline performance achieved through the SAL method.

5.1. Experimental Setup

For transparency and reproducibility, we used open-source classifier implementations of SVMs from the WEKA data mining toolkit (Hall et al., 2009). As classifiers, we chose linear kernel SVMs trained with a complexity parameter C constant of 0.05 and with Sequential Minimal Optimisation (SMO). 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 different initialisations of the random generator. As the evaluation measure, we considered the unweighted average recall (UAR) in accordance with IS09 EC.

5.2. Discussion of Results

According to Figure 2, the sequential addition of human-labelled instances to the initial training set (200 per iteration) leads to continuous improvements in the performance of the classifier. The UAR first increases steeply with the total number of human annotations before reaching a plateau. Moreover, our results display that higher classification accuracy and greater stability are achieved with the mc strategy. This evidence reinforces our assumption that the mc strategy is more robust to noise disturbance (Zhang et al., 2014). Most importantly, the fact that the DAL curves are shorter than the SAL ones shows that the DAL method requires markedly less human annotations to achieve the same performance as SAL. In order to demonstrate the cost reduction, we compare the costs in terms of the numbers 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. In addition, it can be noted that mc allows higher CR for all tested agreement levels. Finally, the analysis of standard deviation shows that the stability of the model is enhanced during the learning process.

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 UAR$_{max}$

	mc		lc	
	UAR$_{max}$	CR (%)	UAR$_{max}$	CR (%)
SAL	68.79	–	68.48	–
j=3	68.79	25.58	68.48	23.53
j=2	68.84	54.83	68.52	47.98
j=1	68.38	79.17	68.16	78.26

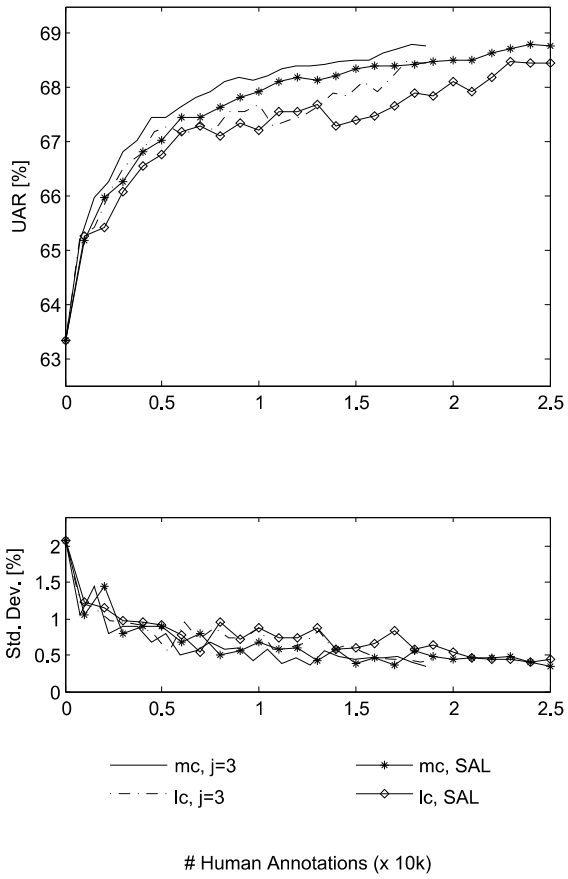

Figure 2. Dynamic Active Learning (DAL) with agreement level $j = 3$ vs. Static Active Learning (SAL) baseline: the performance measures show the UAR values averaged across 20 runs of the algorithm and the respective standard deviations vs. the number of human annotations for the FAU AEC with IS09 EC feature set by 200 initial training instances.

6. Conclusions and Future Work

In this paper, we introduced a novel approach for Dynamic Active Learning that allows utmost reduction of the 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 approach leads to the same performance of the trained model, but requires up to $79.17\,\%$ less human annotations for the medium certainty and $78.26\,\%$ for the least certainty query strategy. Finally, our results reinforce the assumption that the mc strategy achieves higher cost reduction than the lc strategy. For future research, we will extend the DAL algorithm by considering individual rater reliability and inter-rater correlation. In the long term, the full potential of self-optimising classifiers will be realised by combining SSL methods with enhanced DAL techniques.

Acknowledgments

This work was funded by the European Unions's Horizon 2020 and Seventh Framework Programmes under grant agreements No. 645378 (ARIA-VALUSPA) and No. 338164 (ERC Starting Grant iHEARu).

References

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

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

Grimm, M. and Kroschel, K. Evaluation of natural emotions using self assessment manikins. In *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 381–385, Cancun, Mexico, 2005.

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

Howe, J. The rise of crowdsourcing. *Wired magazine*, 14 (6):1–4, 2006.

Kittur, A., Chi, H., and Suh, B. Crowdsourcing user studies with mechanical turk. In *Proc. of the SIGCHI conference on human factors in computing systems*, pp. 453–456, Florence, Italy, 2008.

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

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

Schuller, B. Multimodal Affect Databases - Collection, Challenges & Chances. In Calvo, Rafael A. DMello, Sidney, Gratch, Jonathan, and Kappas, Arvid (eds.), *Handbook of Affective Computing*, Oxford Library of Psychology, chapter 23, pp. 323–333. Oxford University Press, 2015. invited contribution.

Schuller, B. and Batliner, A. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons, New York, NY, 2014.

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

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

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

Steidl, S., Levit, M., Batliner, A., Nöth, E., and Niemann, H. "of all things the measure is man": Automatic classification of emotions and inter-labeler consistency. In *Proc. of ICASSP*, pp. 317–320, Philadelphia, PA, 2005.

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

You, M., Chen, C., Bu, J., Liu, J., and Tao, J. Emotion recognition from noisy speech. In *Proc. of ICME*, pp. 1653–1656, Toronto, Canada, 2006. IEEE.

Yuen, M., King, I., and Leung, K. A survey of crowdsourcing systems. In *Proc. of Privacy, Security, Risk and Trust (PASSAT) and Proc. of Social Computing (SocialCom)*, pp. 766–773, Boston, MA, 2011. IEEE.

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

Zhang, Z., Eyben, F., Deng, J., and Schuller, B. An agreement and sparseness-based learning instance selection and its application to subjective speech phenomena. In *Proc. of 5th International Workshop on Emotion Social Signals, Sentiment & Linked Open Data, satellite of LREC 2014*, pp. 21–26, Reykjavik, Iceland, 2014.

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

Zhu, X., Lafferty, J., and Ghahramani, Z. 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*, pp. 58–65, Washington DC, 2003.