ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 103 (2017) 361 – 364

Fast Track Articles
DOI 10.3813/AAA.919064

Active Learning for Bird Sounds Classification

Kun Qian[1,2], Zixing Zhang[2], Alice Baird[2], Björn Schuller[2,3]

[1] Chair of Human-Machine Communication, Technische Universität München, Theresienstr. 90, 80333 München, Germany. andykun.qian@tum.de

[2] Chair of Complex & Intelligent Systems, Universität Passau, Innstr. 43, 94032 Passau, Germany

[3] Department of Computing, Imperial College London, 180 Queens' Gate, Huxley Bldg., London SW7 2AZ, UK

Summary

It has been shown that automatic bird sound recognition can be an extremely useful tool for ornithologist and ecologists, allowing for a deeper understanding of; mating, evolution, local biodiversity and even climate change. For a robust and efficient recognition model, a large amount of labelled data is needed, requiring a time consuming and costly effort by expert-human annotators. To reduce this, we introduce for the first time, active learning, for automatic selection of the most informative data for training the recognition model. Experimental results show that our proposed; sparse-instance-based and least-confidence-score-based active learning methods reduce respectively 16.0% and 35.2% human annotated samples than compared to passive learning methods, achieving an acceptable performance (unweighted average recall > 85%), when recognising the sound of 60 different species of birds.

© 2017 The Author(s). Published by S. Hirzel Verlag · EAA. This is an open access article under the terms of the Creative Commons Attribution (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent decades, more and more scholars of ecology, bioacoustics, signal processing, and machine learning, are working towards better machine listening for bird sounds, which is now seen as an essential indicator for climate change tracing [1], and species evolution [2]. Bird species recognition by sound will be a very important supplement, or even substitute for traditional telescope based approaches due to its ability for long-term non-human monitoring, unrestricted by adverse weather conditions. From the early work done by McIlraith et al. [3], using two-layer perceptrons to recognise six species of birds, to recent work on handling larger amounts of bird sound data [4], or the use of a more robust, and efficient classification model with limited data [5].

However, there are few studies which deal with the problem of unlabelled bird sound data. Unlabelled data is much more easily acquired in much larger quantities, compared to labelled ones. Asking human-experts to annotate the data would be massively time-consuming, and very expensive. Inspired by our work on active learning (AL) for speech emotion recognition [6], we introduce this methodology for bird sound. In our study, we propose and investigate two AL methods, e. g., the sparse-instance-based AL (SI-AL), and the least-confidence-score-based AL (LCS-AL). We extend the strategy of selecting the sparse data from a multi-class database rather than a two-class problem [6]. As indicated by the database provider[1], the recordings in their database are high quality, and the determination of species and the locations of the recordings are highly reliable, considering the experimental experience of our previous work, we chose a least-confidence-score rather than the medium-confidence-score in [6].

This paper is organized as follows: related prior work will be shown in Section 2. Then, Section 3 will describe our current methodology, and the the databases utilised. Experimental results, and discussion will be presented in Section 4 before conclusions are made in Section 5.

2. Related Prior Work

Inspired by our successful work on active learning for speech emotion recognition [6], we introduce this methodology for bird sounds. There are two main differences between this work, and the work in [6]: 1) we extend the sparse-instances-based active learning into a multi-class problem; 2) we use a least-certainty method to select ranked samples rather than the medium certainty method used in [6]. This work is also an extended work of [7],

Received 6 January 2017,
accepted 2 March 2017,
published online 5 April 2017.

[1] http://www.animalsoundarchive.org/RefSys/ProjectDescription.php?CurLa=en

© 2017 The Author(s). Published by S. Hirzel Verlag · EAA.
This is an open access article under the terms of the CC BY 4.0 license.

which proposed a framework for bird sound detection, classification, and classifier modification.

Algorithm 1: *Passive Learning*

Repeat:
1) Randomly select N samples ω_n from the unlabelled set Ω.
2) Let human experts annotate the selected subset ω_n.
3) Remove ω_n from the unlabelled set Ω, i. e., $\Omega = \Omega \setminus \omega_n$.
4) Add ω_n to the labelled set Σ, i. e., $\Sigma = \Sigma \cup \omega_n$.

End: When iteration reaches a defined number, or the trained classifier achieves a certain performance on the associated validation set.

3. Methodology

A common approach for coping with the issue of limited data, is passive learning (PL), which randomly, and independently selects samples from unlabelled data and asks for additional human-expert annotation. This method can be extremely time-consuming, and costly [6]. It has been reported that, in a typical data mining project, approximately 80 % of the work is completed during data collection, cleaning, and annotation [8]. The detailed PL algorithm is shown in Algorithm 1.

Another method, known as 'Active Learning (AL)', uses only the 'most informative' manually labelled samples, using are several methods to select them [9]. The most popular is based on uncertainly sampling. Such a strategy uses a confidence measure criterion to select data. This assumes that the unlabelled data predicted with the least-confidence-score are the most informative, and should be selected for human annotation. Compared with other AL approaches, this methods is simple and effective, and widely used in many other applications [6]. Thus, it is one of the methods investigated in this paper, named as least-confidence-score-based AL (LCS-AL). In addition, another state-of-the-art active learning approach, sparse-instance-based AL (SI-AL), has shown promising performance for emotion recognition in [6]. Compared with other AL approaches, it assumes that the instances predicted as the sparse classes, are the most helpful to enhance discrimination of the classification model. More detailed information about these approaches is to follow.

3.1. Sparse-Instance-based Active Learning

Bird sound data is naturally unbalanced due to the varied distribution of bird species, and the collected recordings. Therefore, we take the sparse-instance-based active learning (SI-AL) as our method. In SI-AL, we consider the unbalanced characteristics of the bird sound data set. We randomly select N samples from this data, and throughout each iteration, classify these as the 'sparse class', and 'most informative' samples. The pseudo code of this method is shown in Algorithm 2. We need to note that, if it happens that no data is classified as a 'sparse class', the Algorithm 2 will make random selections as in Algorithm 1.

Algorithm 2: *Sparse-Instance-based Active Learning*

Repeat:
1) Train a classifier Ψ based on the labelled set Σ.
2) Randomly Select N samples ω_n from the unlabelled set Ω as predicted by Ψ to be the 'sparse class', whose value is less than a specified threshold, i. e., $N_s < N_{max} \times sparse\_fraction$, where N_s is the sample value of a certain class, N_{max} is the maximum sample value among all data, and $sparse\_fraction$ is a predefined ratio.
3) Let human expert annotate ω_n.
4) Remove ω_n from the unlabelled set Ω, i. e., $\Omega = \Omega \setminus \omega_n$.
5) Add ω_n to the labelled set Σ, i. e., $\Sigma = \Sigma \cup \omega_n$.

End: When the iteration reaches a defined number, or the Ψ achieves a certain performance on the validation set.

Algorithm 3: *Least-Confidence-Score-based Active Learning*

Repeat:
1) Train a classifier Ψ based on the labelled set Σ.
2) Predict the unlabelled data Ω by Ψ, and rank the data by its prediction *confidence scores*.
3) Randomly select N samples ω_n from the last c % of ranked data in Ω.
4) Let human expert annotate ω_n.
5) Remove ω_n from the unlabelled set Ω, i. e., $\Omega = \Omega \setminus \omega_n$.
6) Add ω_n to the labelled set Σ, i. e., $\Sigma = \Sigma \cup \omega_n$.

End: When iteration reaches a defined number, or the Ψ achieves a certain performance on the validation set.

3.2. Least-Confidence-Score-based Active Learning

Compared with the SI-AL, least-confidence-score-based active learning (LCS-AL) considers the capacity of classifier trained at initial steps. In LCS-AL, we treat 'the least-confidence-scores' samples ranked by classifier's estimated posterior probabilities as the 'most informative' samples. As the *confidence scores* have a consistent corresponding relationship to the *posterior probabilities* estimated by the classifier, we use the latter to rank the unlabelled data in our method. The *cross entropy* of the estimated probability vector is used to represent the *confidence score* (as inverted ranking) of the unlabelled data. The pseudo code of this method is shown in Algorithm 3.

4. Experimental Results

4.1. Bird Sound Database

The bird sound data utilised for these experiments has been provided by the Museum für Naturkunde Berlin (MNB)[2], Germany. Due to copyright restrictions defined

by the author, we removed all audio files which were labelled as protected. In addition, we eliminated the subclasses of bird species which had less than 20 audio samples. In total, our bird sound dataset has 3 483 audio files, including 60 species (sub-classes) of birds. As Table I shows, we separated the database into three parts: a smaller labelled dataset, a larger pool dataset without labelled information, and a test dataset.

4.2. Acoustic Feature Set & Classifier

As previously, our toolkit, openSMILE [10], has been used within the bird sound recognition task [7]. In this study, we use the 'ComParE' set, which contains a total of 6 373 features. The detailed information about the low-level-descriptors (LLDs), and the functions applied to those LLDs, can be found in [11]. Before feeding into the classifier model, all the original features are standardized to eliminate any effect of outliers. As a classifier, we chose Support Vector Machines (SVM) [12]. We implemented the SVM training, and testing process with LIBSVM [13]. Based on previous work, we select SVM with a *linear kernel*, and a complexity value of 0.01. The method to calculate the estimated *posterior probabilities* of the SVM classifier were described in [14].

4.3. Experimental Setup

All experiments were done in the software environment of Matlab 2016b by Math Works. The *sparse_fraction* mentioned in Algorithm 2 is set to 0.5, and the c value mentioned in Algorithm 3 is set to 20, respectively. The iteration number is set to use all of the unlabelled data due to each learning strategy, and N is 100. To make a fair comparison, we randomly repeat 20 rounds of independent experiments of PL. In this work, we use unweighted average recall (UAR), i. e., the averaged accuracy of each class of data, as our evaluation metric due to the imbalanced distribution of bird sound data.

4.4. Results and Discussion

Figure 1 shows the unweighted averaged recalls vs. the used human annotated samples by PL, and AL during iterations. We can see that, both the SI-AL, and the LCS-AL use less human annotated samples than PL, once the best trained model is a achieved. We observe that, LCS-AL, is better than SI-AL in selecting the 'most informative' samples. LCS-AL is the fastest method to build an acceptable model (UAR above 85.0 %) among all strategies. When using around 2k (2 020 samples in 14-th iteration) human annotated samples, LCS-AL can reach approximately 86.9 % UAR, higher than performance by PL (81.6 %), with a significant level at $p < 0.005$ within one-tailed z-test. It should be noted that, in LCS-AL, the UAR can be improved quickly at earlier iterations, but later will gradually become stable or even decrease slightly. This is caused due to an increased use of the most informative samples,

[2] http://www.animalsoundarchive.org/RefSys/Statistics.php.

Table I. Bird sounds data set.

$\sum$	Labelled	Pool	Test
3 483	720	2 090	673

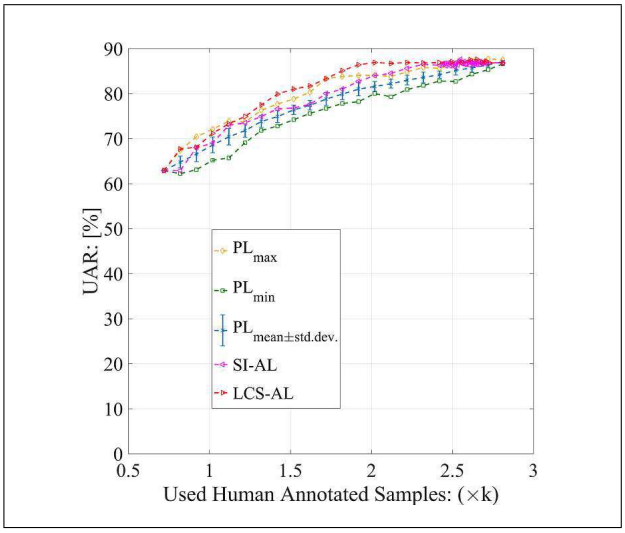

Figure 1. Unweighted Average Recall (UAR) vs. used human annotated samples. 20 independent runs were carried out with PL.

the remaining samples bring unimportant information to the classifier, which in return, produces a negative effect on the model. Figure 2 shows the percentage of the least used human annotated samples from the whole unlabelled data when achieving an acceptable model (UAR above 85.0 %) for the proposed three methods. LCS-AL required the least number of human annotated samples (1 100, 52.6 %) to train the best performing model, 35.2 % absolute less than PL (1 835, 87.8 %, averaged by 20 independent experiments), and 19.2 % less than SI-AL (1 500, 71.8 %). In this study, SI-AL is inferior to LCS-AL, but it still requires 16.0 % less human annotated samples when compared to PL. With improved classifier performance, these experimental results prove that, the two proposed AL methods can considerably reduce the need for human experts' when comparing this to PL. The two proposed methods, i. e., SI-AL, LCS-AL can also be easily extended to *multi-label* cases if we modify the definition of *sparse class* (for SI-AL), and *cross entropy* (for LCS-AL).

5. Conclusion

In this work, we proposed and compared two active learning methods, namely sparse-instance-based active learning, and least-confidence-score-based active learning, for the task of bird sound classification. Experimental results have demonstrated that, both active learning methods can considerably reduce the need for human annotation, when compared with passive learning, once an acceptable model is achieved. The least-confidence-score-based active learning method can outperform passive learning

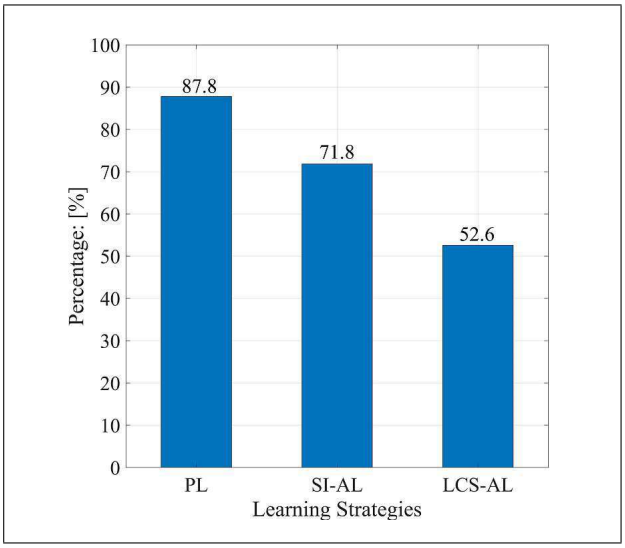

Figure 2. Minimal amount, in percent of human annotations needed from an unlabelled dataset, to build an acceptable model (UAR> 85.0 %) using different learning strategies.

when fed a small number of human annotated samples during earlier iterations. In this study, the performance of sparse-instance-based active learning falls behind least-confidence-score-based active learning. Future work will focus on extending our study to *multi-label* cases, and comparing with other state-of-the-art active learning methods.

Acknowledgement

This work was partially supported by the China Scholarship Council (CSC), and the European Union's Seventh Framework under grant agreements No. 338164 (ERC Starting Grant iHEARu).

References

[1] C. Parmesan, G. Yohe: A globally coherent fingerprint of climate change impacts across natural systems. Nature **421** (2003) 37–42.

[2] C. K. Catchpole, P. J. Slater: Bird song: Biological themes and variations. Cambridge University Press, Cambridge, UK, 2003.

[3] A. L. McIlraith, H. C. Card: Birdsong recognition using backpropagation and multivariate statistics. IEEE Transactions on Signal Processing **45** (1997) 2740–2748.

[4] M. Lasseck: Large-scale identification of birds in audio recordings. Working Notes of CLEF 2014 Conference, Sheffield, UK, 2014, 643–653.

[5] L. N. Tan, A. Alwan, G. Kossan, M. L. Cody, C. E. Taylor: Dynamic time warping and sparse representation classification for birdsong phrase classification using limited training data. The Journal of the Acoustical Society of America **137** (2015) 1069–1080.

[6] Z. Zhang, B. Schuller: Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition. Proc. of INTERSPEECH, Portland, Oregon, USA, 2012, 362–365.

[7] K. Qian, Z. Zhang, F. Ringeval, B. Schuller: Bird sounds classification by large scale acoustic features and extreme learning machine. Proc. of GlobalSIP, Orlando, Florida, USA, 2015, 1317–1321.

[8] D. Braha: Data mining for design and manufacturing: Methods and applications. Kluwer Academic, 2001.

[9] B. Settles: Active learning literature survey. Computer Sciences Technical Report University of Wisconsin-Madison, 2010.

[10] F. Eyben, M. Wöllmer, B. Schuller: OpenSMILE: the munich versatile and fast open-source audio feature extractor. Proc. of ACM MM, Firenze, Italy, 2010, 1459–1462.

[11] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al.: The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. Proc. of INTERSPEECH, Lyon, France, 2013, no pagination.

[12] C. Cortes, V. Vapnik: Support-vector networks. Machine Learning **20** (1995) 273–297.

[13] C.-C. Chang, C.-J. Lin: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[14] H.-T. Lin, C.-J. Lin, R. C. Weng: A note on platt's probabilistic outputs for support vector machines. Machine Learning **68** (2007) 267–276.