Multiple-Instance Multiple-Label Learning for the Classification of Frog Calls With Acoustic Event Detection

Jie Xie, Michael Towsey, Liang Zhang, Kiyomi Yasumiba, Lin Schwarzkopf, Jinglan Zhang, and Paul Roe

Electrical Engineering and Computer Science School, Queensland University of Technology, Brisbane, Australia
`xiej8734@gmail.com`
`https://www.ecosounds.org/`

Abstract. Frog call classification has received increasing attention due to its importance for ecosystem. Traditionally, the classification of frog calls is solved by means of the single-instance single-label classification classifier. However, since different frog species tend to call simultaneously, classifying frog calls becomes a multiple-instance multiple-label learning problem. In this paper, we propose a novel method for the classification of frog species using multiple-instance multiple-label (MIML) classifiers. To be specific, continuous recordings are first segmented into audio clips (10 seconds). For each audio clip, acoustic event detection is used to segment frog syllables. Then, three feature sets are extracted from each syllable: mask descriptor, profile statistics, and the combination of mask descriptor and profile statistics. Next, a bag generator is applied to those extracted features. Finally, three MIML classifiers, MIML-SVM, MIML-RBF, and MIML-kNN, are employed for tagging each audio clip with different frog species. Experimental results show that our proposed method can achieve high accuracy (81.8% true positive/negatives) for frog call classification.

Keywords: Frog call classification, acoustic event detection, multiple-instance multiple-label learning

1 Introduction

Recently, human activity and climate change put a negative effect on frog biodiversity, which makes frog monitoring become ever more important. Compared with the traditional monitoring method such as field observation, acoustic sensors have greatly extended acoustic monitoring into larger spatio-temporal scales [1]. Correspondingly, large volumes of acoustic data are generated, which makes it essential to develop automatic methods.

Several papers have already described automated methods for the classification of frog calls. Han et.al combined spectral centroid, Shannon entropy, Renyi entropy for frog call recognition with a k-nearest neighbour classifier [2]. Gingras

et al. proposed a method based on mean value for dominant frequency, coefficient of variation of root-mean square energy, and spectral flux for anuran classification [3]. Bedoya et al. used Mel-frequency cepstral coefficients (MFCCs) for the recognition of anuran species with a fuzzy classifier [4]. Xie et al proposed a method based on track duration, dominant frequency, oscillation rate, frequency modulation and energy modulation to do frog call [5]. All those previous methods achieve a high accuracy rate in recognition and classification, but recordings used in those papers are assumed that there is only a single frog species present in each recording.

Unfortunately, all the recordings used in this study are low signal to noise ratio and contain many overlapping animal vocal activities including frogs, birds, crickets and so on. To solve this problem, the multiple-instance multiple-label classifier for supervised classification is formulated [6]. In the previous study, Briggs et al has already introduced the MIML classifiers for acoustic classification of multiple simultaneous bird species [7]. In their method, a supervise learning classifier was employed for segmenting acoustic events, which required lots of annotations.

In this study, we introduced the MIML algorithm for frog call classification. Rather than using a supervised learning method for syllable segmentation, acoustic event detection is first employed to separate frog syllables. Then, three feature sets, mask descriptor, profile statistics, and the combination of mask descriptor and profile statistics, are calculated from each syllable. After applying a bag generator to those extracted feature sets, three classifiers, MIML-SVM [6], MIML-RBF [8], and MIML-kNN [9], are lastly used for the recognition of multiple simultaneous frog species. Experimental results show that our proposed method can achieve high classification accuracy.

2 Materials and methods

2.1 Materials

Digital recordings in this study were obtained with a battery-powered, weatherproof Song Meter (SM2) box. Recordings were two-channel, sampled at 22.05 kHz and saved in WAC4 format. Here, a representative sample of 342 10-s recordings was selected to train and evaluate our proposed algorithm for predicting which frog species are present in a recording. All those examples were collected between 02/2014 to 03/2014, because it is the frog breeding season with high calling activity. All the species that are present in each 10-s recording were manually labelled by an ecologist who studies frog calls. There are totally eight frog species in the recordings: Canetoad (CAD) (F_0=560 Hz), Cyclorana novaehollandiae (CNE) (F_0=610 Hz), Limnodynastes terraereginae (LTE) (F_0=610 Hz), Litoria fallax (LFX) (F_0=4000 Hz), Litoria nasuta (LNA) (F_0=2800 Hz), Litoria rothii (LRI) (F_0=1800 Hz), Litoria rubella (LRA) (F_0=2300 Hz), and Uperolela mimula (UMA) (F_0=2400 Hz). Here, F_0 is the mean dominant frequency for each frog species. Each recording contains between 1 and 5 species. Following

the prior work [7], we assume that recordings without any frog calls can be detected by acosutic event detection.

2.2 Signal processing

All the recordings were re-sampled at 16 kHz and mixed to mono. A spectrogram was then generated by applying short-time Fourier transform to each recording. Specifically, each recording was divided into frames of 512 samples with 50% frame overlap. A fast Fourier transform was then performed on each frame with a Hamming window, which yielded amplitude values for 256 frequency bins, each spanning 31.25 Hz. The final decibel values (S) were generated using $S_{tf} = 20*log_{10}(A_{tf})$, where A is the amplitude value, $t = 0, ..., T-1$ and $f = 0, ..., F-1$ represent frequency and time index, T and F are 256 frequency bins and 625 frames, respectively.

2.3 Acoustic event detection for syllable segmentation

Acoustic event detection (AED) aims to detect specified acoustic event in an audio stream. In this study, we use AED to segment frog syllables. Since all the recordings are collected from the field, there are much overlapping vocal activities. Traditional methods for audio segmentation are based on time domain information [10, 11], which cannot address those recordings. Here, we modified the AED method developed by Towsey et al [12] to segment recordings with overlapping activities. The detail of our AED method is described as follows:

Step 1: Wiener filter

To de-noise and smooth the spectrogram, a 2-D Wiener filter is applied to the spectrogram image over a 5×5 time-frequency grid, where the filter size is selected after considering the trade-off between removing the background graininess and blurring the acoustic events.

$$\hat{S_{tf}} = \mu + \frac{(\sigma^2 - \nu^2)}{\sigma^2}(S_{tf} - \nu) \tag{1}$$

where μ and σ^2 are local mean and variance, respectively. ν^2 is the noise variance estimated by averaging all local variances.

Step 2: Spectral subtraction

After Wiener filter, the graininess has been removed. However, some noises such as wind, insect, motor engine that cover the whole recording cannot be removed. Here, a modified spectral subtraction is used for dealing with those noise [13].

Step 3: Adaptive thresholding

After noise reduction, the next step is to convert the noise reduced spectrogram $\hat{S'_{tf}}$ into the binary spectrogram S^b_{tf} for events detection. Different from the hard threshold in Towseys work, an adaptive thresholding method named *Otsu thresholding* is used to convert the smoothed spectrogram into binary spectrogram. Otsus method assumes that the spectrogram is composed of two classes: acoustic events and background noise. An optimal threshold value is used for


Algorithm 1: Modified Spectral Subtraction

Data: $\hat{S_{tf}}$, spectrogram after Wiener filtering.

Result: $\hat{S'_{tf}} = \hat{S_{tf}}$, noise reduced spectrogram.

begin

    Construct an array of the modal noise values for all frequency bins;

    for $f \in F$ **do**

        1. calculate the histogram of the intensity value over each frequency bin

        2. smooth the histogram array with a moving average window of size 7

        3. regard the modal noise intensity at the position of maximal bin in the left-side of the histogram

    Smooth the array with a moving average filter with window of size 5;

    for $f \in F$ **do**

        1. subtract the modal noise intensity

        2. truncated negative decibel values to zero
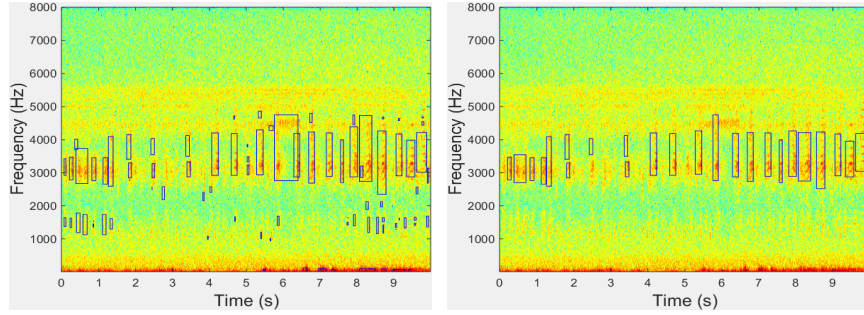
Fig. 1. Acoustic event detection results before (Left) and after (Right) event filtering based on dominant frequency. Here, blue rectangle means the time and frequency boundary of each detected event.

the decision. After thresholding, each group of contiguous positive pixels will be regarded as a candidate event.

Step 4: Events filtering using dominant frequency and event area

After aforementioned process, not all detected events are correspond to frog vocalizations. To further remove those events that are from the listed frog species in section 2.1, dominant frequency (F_0) and area within the event boundary (Ar) are used for filtering.

Step 5: Region growing

Region growing algorithm is utilized to obtain the contour of the particular acoustic event [14]. To get the accuracy boundary of each acoustic event and improve the discrimination of extracted features, a 2-D region growing algorithm is applied for obtaining the accuracy event shape within each segmented event. First, a maximal intensity value within each segmented event is selected as the seed point. Then, if the difference between the neighbourhood pixels and the seed(s) is smaller than the threshold, the neighbourhood pixels will be located

and assigned to the output image. Next, the new added pixels are used as seeds for further processing until all the pixels that satisfy the criteria are added to the output image. The final results after region growing are shown in Fig. 2. Here, the threshold value is empirically set as 5 dB.


Algorithm 2: Event filtering based on dominant frequency and event area

Data: S_{tf}^b, spectrogram; $t_s(n)$, $t_e(n)$, $f_l(n)$, $f_h(n)$, location of each acoustic event n; $F_0(i)$, dominant frequency of frog species i.

Result: $\tilde{S}_{tf}$, spectrogram after events filtering.

begin

    Calculate the area of each acoustic event n.
    $Area(n) = (t_e(n) - t_s(n)) * (f_h(n) - f_l(n))$
    for $n \in N_{e1}$ **do**
        if $Ar(n) \geq Ar_l$ **then**
            split event n into small events
    where Ar_l is set as 3000 pixels.
    Filter events using dominant frequency $f_d(n) = \sum_{t=t_s(n)}^{t_e(n)} F(t)/t_e(n) - t_s(n)$
    where $F(t)$ is the peak frequency of each frame within the event area
    for $n \in N_{e2}$ **do**
        for $i \in I$ **do**
            if $f_d(n) \geq F_0(i) + \theta$; $f_d(n) \leq F_0(i) - \theta$ **then**
                $f_d(n) = 0$;
    where θ is frequency range and set as 300 Hz.
    Remove small acoustic events except frequency band between θ_l and θ_h
    for $n \in N_{e2}$ **do**
        if $Ar(n) \leq Ar_s$ **then**
            remove event n
    where Ar_s is set at 300 pixels, θ_l and θ_h are set as 300 Hz and 800 Hz, respectively. Because the area of LTE is smaller than Ar_s.
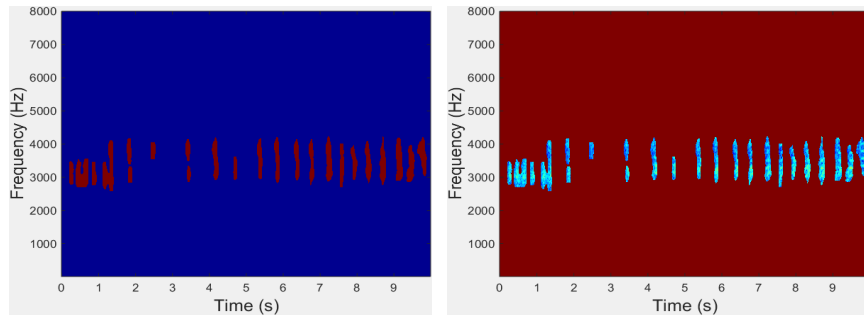
Fig. 2. Acoustic event detection results after region growing. Left: binary segmentation results; Right: segmented frog syllables.

2.4   Feature extraction

Based on acoustic event detection results, two feature sets are first calculated to describe each event (syllable): mask descriptor and profile statistic [7]. Here, we exclude histogram of orientation from our feature sets, because the previous study has already demonstrated its poor classification performance [7]. For mask descriptor, it is used to describe the syllable shape including minimum frequency, maximum frequency, bandwidth, duration, area, perimeter, noncompactness, rectangularity. For profile statistics, there are time-Gini, frequency-Gini, frequency-mean, frequency-variance, frequency-skewness, frequency-kurtosis, frequency-max, time-max, mask-mean, and mask standard deviation. The third feature set consists of all features.

Table 1. Accuracy measure for MIML classifiers with different feature sets. Here, ↓ indicates the smaller the better, while ↑ indicates the bigger the better.

Feature	Algorithm	Hamming loss ↓	Rank loss ↓	One-error ↓	Coverage ↓	Micro-AUC ↑
MD	MIML-SVM	0.253	0.186	0.308	3.147	0.745
MD	MIML-kNN	0.205	0.153	0.298	2.647	0.771
MD	MIML-RBF	**0.182**	**0.132**	**0.223**	**2.352**	**0.828**
PS	MIML-SVM	0.239	0.208	0.323	3.544	0.728
PS	MIML-kNN	0.211	0.153	0.298	2.647	0.777
PS	MIML-RBF	0.186	0.161	0.338	3.161	0.746
AF (MD+PS)	MIML-SVM	0.261	0.199	0.279	3.588	0.761
AF (MD+PS)	MIML-kNN	0.205	0.160	0.264	2.735	0.787
AF (MD+PS)	MIML-RBF	0.191	0.142	0.220	2.632	0.821

3   Multiple-instance multiple-label classifiers

After feature extraction, three MIML algorithms are evaluated for the classification of multiple simultaneous frog calls: MIML-SVM, MIML-RBF, and MIML-kNN. With some form of event-level distance measure, the MIML problem has been reduced to a single-instance multiple-label problem by associating each event with a event-level feature [7]. Here, the maximal and average Hausdorff distances between two syllables are used by MIML-SVM and MIML-RBF, separately. For MIML-kNN, the nearest neighbour is used to assign syllable-level features.

4   Experiment results

4.1   Parameter tuning

There are three modules whose parameters need to be discussed: signal processing, acoustic event detection, and classification. For signal processing, the window size and overlap are 512 samples and 50%, respectively. During the process of acoustic event detection, four thresholds for event filtering need to be

determined, which are small and large area threshold, and frequency boundary for events filtering. All those thresholds were determined empirically by applying various combinations of thresholds to a small number of randomly selected 10s clips. For MIML-SVM classifiers, the parameters used are (C, γ, r) and set as $(0.1, 0.6, 0.2)$ experimentally. For MIML-RBF, the parameters are (r, μ) and set as $(0.1, 0.6)$. For MIML-kNN, the number of references (k) and citers $(k^{'})$ are 10 and 20, respectively.

4.2   Classification

In this study, all the algorithms were programmed in Matlab 2014b. Each MIML algorithm is evaluated with five-fold cross-validation on the collection of 342 species-labelled recordings. Five measures including Hamming loss, rank loss, one-error, coverage, and micro-AUC are used to characterize the accuracy of each algorithm [15] [16]. The definition of each measure can be found in [7], the positive/negatives is defined as $1-$Hamming loss and it is 0.818 for MIML-RBF with MD. Mask descriptor (MD) and profile statistical (PS), and all features (AF) are put into the three classifiers, respectively. The performance of each MIML classifier is shown in Table 1. Here, the best classification accuracy is achieved by MIML-RBF using MD. For each classifier, the classification accuracy of MD is higher than PS and AF, which shows that the event shape have higher discrimination power than the event content. To give a concrete view of predictions, the results of 5 randomly selected recordings using MIML-RBF are shown in Table 2. Recordings of No.1 and No.3 are accurately predicted.

Table 2. Example predictions with MIML-RBF.

No.	Ground truth	Predicted labels
1	UMA	UMA
2	LNA, LRI, UMA	LNA, LRA, UMA
3	LNA, UMA	LNA, UMA
4	LNA, LFX, LRA	LNA, LFX, LRI, LRA
5	LNA, LFX, LRA	LNA, LRA

5   Conclusion

In this study, we propose a novel method for the classification of multiple simultaneous frog species in environmental recordings. To the best of our knowledge, this is the first study that applies the MIML algorithm to frog calls. Since frogs tend to call simultaneously, the MIML algorithm is more suitable for dealing with those recordings than single-instance single-label classification. After applying acoustic event detection algorithm to each 10s recording, each frog syllable is segmented. Then, three feature sets are calculated based on those segmented

syllables. Finally, three MIML classifiers are used for the classification of frog calls with the best accuracy (81.8% true positive/negatives). Future work will focus on the study of novel features and MIML classifiers for further improving the classification performance.

References

1. J. Wimmer, M. Towsey, B. Planitz, I. Williamson, and P. Roe, "Analysing environmental acoustic data through collaboration and automation," *Future Generation Computer Systems*, vol. 29, no. 2, pp. 560–568, February 2013.
2. N. C. Han, S. V. Muniandy, and J. Dayou, "Acoustic classification of australian anurans based on hybrid spectral-entropy approach," *Applied Acoustics*, vol. 72, no. 9, pp. 639–645, 2011.
3. B. Gingras and W. T. Fitch, "A three-parameter model for classifying anurans into four genera based on advertisement calls," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 547–559, 2013.
4. C. Bedoya, C. Isaza, J. M. Daza, and J. D. López, "Automatic recognition of anuran species based on syllable identification," *Ecological Informatics*, vol. 24, pp. 200–209, 2014.
5. J. Xie, M. Towsey, A. Truskinger, P. Eichinski, J. Zhang, and P. Roe, "Acoustic classification of australian anurans using syllable features," in *2015 IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (IEEE ISSNIP 2015)*, Singapore, Singapore, Apr. 2015.
6. Z.-H. Z. M.-L. Zhou, "Multi-instance multi-label learning with application to scene classification," *Advances in Neural Information Processing Systems*, p. 1609:1616, 2007.
7. F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. Hadley, A. S. Hadley, and M. G. Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4640–4650, 2012.
8. M.-L. Zhang and Z.-J. Wang, "Mimlrbf: Rbf neural networks for multi-instance multi-label learning," *Neurocomputing*, vol. 72, no. 16, pp. 3951–3956, 2009.
9. M.-L. Zhang, "A k-nearest neighbor based multi-instance multi-label learning algorithm," in *Tools with Artificial Intelligence (ICTAI), 2010 22nd IEEE International Conference on*, vol. 2.   IEEE, 2010, pp. 207–212.
10. P. Somervuo *et al.*, "Classification of the harmonic structure in bird vocalization," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 5.   IEEE, 2004, pp. V–701.
11. C.-J. Huang, Y.-J. Yang, D.-X. Yang, and Y.-J. Chen, "Frog classification using machine learning techniques," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3737–3743, 2009.
12. M. Towsey, B. Planitz, A. Nantes, J. Wimmer, and P. Roe, "A toolbox for animal call recognition," *Bioacoustics*, vol. 21, no. 2, pp. 107–125, 2012.
13. J. Xie, M. Towsey, J. Zhang, and P. Roe, "Image processing and classification procedure for the analysis of australian frog vocalisations," in *Proceedings of the 2Nd International Workshop on Environmental Multimedia Retrieval*, ser. EMR '15.   New York, NY, USA: ACM, 2015, pp. 15–20.
14. A. Mallawaarachchi, S. Ong, M. Chitre, and E. Taylor, "Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphin whistles,"

The Journal of the Acoustical Society of America, vol. 124, no. 2, pp. 1159–1170, 2008.

15. Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Miml: a framework for learning with ambiguous objects," *CORR abs/0808.3231*, 2008.
16. A. Dimou, G. Tsoumakas, V. Mezaris, I. Kompatsiaris, and I. Vlahavas, "An empirical study of multi-label learning methods for video annotation," in *Content-Based Multimedia Indexing, 2009. CBMI'09. Seventh International Workshop on.* IEEE, 2009, pp. 19–24.