Applied Acoustics 110 (2016) 91–98

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Using multi-label classification for acoustic pattern detection and assisting bird species surveys

Liang Zhang *, Michael Towsey, Jie Xie, Jinglan Zhang, Paul Roe

Eco-acoustics Group, Science and Engineering Faculty, Queensland University of Technology, Australia

ARTICLE INFO

Article history:
Received 3 December 2015
Received in revised form 22 March 2016
Accepted 22 March 2016
Available online 28 March 2016

Keywords:
Soundscape ecology
Computational ecology
Acoustic indices
Multi-label classification

ABSTRACT

Acoustics is a rich source of environmental information that can reflect the ecological dynamics. To deal with the escalating acoustic data, a variety of automated classification techniques have been used for acoustic patterns or scene recognition, including urban soundscapes such as streets and restaurants; and natural soundscapes such as raining and thundering. It is common to classify acoustic patterns under the assumption that a single type of soundscapes present in an audio clip. This assumption is reasonable for some carefully selected audios. However, only few experiments have been focused on classifying simultaneous acoustic patterns in long-duration recordings. This paper proposes a binary relevance based multi-label classification approach to recognise simultaneous acoustic patterns in one-minute audio clips. By utilising acoustic indices as global features and multilayer perceptron as a base classifier, we achieve good classification performance on in-the-field data. Compared with single-label classification, multi-label classification approach provides more detailed information about the distributions of various acoustic patterns in long-duration recordings. These results will merit further biodiversity investigations, such as bird species surveys.

1. Introduction

Natural acoustics offers a wealth of information for understanding dynamics of ecosystem and biodiversity, capturing an ecologically meaningful environment on top of what the visual cue can provide. Early research on the natural environment relies heavily on in-the-field human's visual and auditory surveys [1], but the quantity and quality of collected data are subject to perception and constrained at small temporal and spatial scales. Not until the use of acoustic sensor systems have massive amounts of acoustic data become accessible [2]. Once being deployed, acoustic sensors can record automatically in spite of poor external conditions such as darkness and visual obstruction. Furthermore, acoustic signals are cheap to store and compute.

Recently a new research area – soundscape ecology [3–5], has emerged in line with the escalating acoustic data. Soundscape ecology focuses on the study of ecological processes by utilising a community level of acoustic information emanated from a landscape. It considers spatiotemporal characteristics of a local ecosystem as a study subject rather than analysis of species-specific vocalisations. One example of related topics is concerned with species richness, which attempts to find the number of different bird species represented in an ecological region at a specified time [6]. Others may concentrate on the evenness [7] and heterogeneity [8] of an acoustic community.

Much effort has been placed on developing effective acoustic indices to characterise community-level ecological processes. Here, acoustic indices refer to statistical values that describe the spatiotemporal distribution of acoustic energy within recordings. According to a recent overview [9], acoustic indices can be categorised into two classes: within-group indices and between-group indices. Within-group indices aim to reflect the ecological information within a single acoustic community; whereas between-group indices are designed to measure the differences between two or more acoustic communities. In this paper, we use within-group indices because our interest lies in the analysis of concomitant acoustic patterns in individual audio recordings, rather than the differences between them. A typical within-group index is acoustic entropy. Experimental results have shown that it increases monotonically with the number of species using recordings collected from Tanzanian coastal forests [10]. Another example is the acoustic complexity index, which has been used to estimate the number of bird vocalisations [11]. A further case-study argues that acoustic complexity index can be used to monitor long-term acoustic fluctuations of an ecological community [12].

* Corresponding author at: Room 1002, Level 10, S Block, Gardens Point Campus, 2 George Street, Brisbane, QLD 4000, Australia.
E-mail address: l68.zhang@hdr.qut.edu.au (L. Zhang).

http://dx.doi.org/10.1016/j.apacoust.2016.03.027
0003-682X/

Although acoustic indices have been demonstrated to be a powerful tool to identify acoustic dynamics over time and through space, a single index is highly unlike to cover all facets of ecological information. It is essential to use a combination of these acoustic indices to achieve a complementary understanding of a soundscape [9]. A recent paper has shown that combinations of acoustic indices can provide more efficient estimation of species richness and detect acoustic patterns such as rain and insect chorus [13]. Acoustic indices have also been integrated for visualisation, which enables to facilitate navigation through long-duration audio recordings [14].

Automated classification of natural soundscapes provides a unique opportunity to study dynamics of the environment and biodiversity. Research in this field has become active in the last few years, but the work is remarkably less compared to that for urban soundscape analysis [15,16]. Automated recognition of unstructured natural acoustic patterns is still in its infancy. Some methods classify acoustic scenes with acoustic event-specific features derived from carefully selected movie and television tracks [17,18]. Others attempt to classify relatively short-duration (3- or 4-s) acoustic environments without pre-extracting event-specific features [19,20]. Although automated approaches are far from perfect due to the innate complexity of environmental data, they are evolving rapidly and have achieved promising results.

Efficient and effective methods are required to transmute the escalating acoustic data into useful knowledge. This study addresses the automated classification and detection of concomitant acoustic patterns in audio clips. Automated techniques facilitate the ecological surveys by providing informative distribution of various acoustic patterns, which can merit further 'details-on-demand' analysis. For example, the studies of bird species richness can be conducted with audio clips that have been automatically detected as containing birds.

The main contribution of this paper is to introduce a multi-label classification approach to detect co-occurring natural acoustic patterns using indices derived from raw audio data. Here the extraction of acoustic indices as features reduces the amount of data to be processed, providing a more efficient characterisation of audio clips; the use of multi-label classification can reveal naturally co-occurring acoustic patterns and the relationships between acoustic indices and general ecological processes.

The remainder of this paper is structured as follows. Section 2 provides the related work on multi-label classification. Section 3 presents the details on the proposed methods, including how to calculate acoustic indices and what classifier is used. Section 4 demonstrates the classification results and offers the related discussion. Finally, Section 5 concludes and describes our future work.

2. Related work

2.1. Multi-label classification

Traditional single-label classification is a supervised machine learning technique that associates a single label with each instance. A significant body of research in single-label classification can be found on vocal species detection, including marine mammals [21], birds [22–24], and insects [25]. High classification accuracy has been achieved in these studies. Nevertheless, these tasks are performed on carefully selected short-duration audio segments where only one species exists. However, for an in-field audio recording, co-occurrence of bird vocalisations and the background noise pose a new challenge to automated detection.

To resolve this problem, multi-label classification has been proposed. Unlike prior work on single-label classification, multi-label classification deals with the issue of associating a single instance with multiple labels. It has been applied to a wide range of areas, including text classification [26], audio and video classification [27,28], and bioinformatics [29]. For instance, an audio clip may contain multiple vocal bird species or multiple acoustic patterns such as rain, wind, and bird vocalisations; genres of a film can be labelled as 'action', 'adventure', and 'fantasy'.

The study of multiple simultaneous bird species in Briggs' paper [30] presented a multi-instance multi-label classification (MIML) approach to predict a set of species in a given recording. To clarify, the multi-instance in their paper denotes bird vocalisations in each audio clip; the objects to be classified are audio clips, and the labels are the species present. In other words, their task is to associate an audio clip with multiple species labels and multi-instance emphasise on the method of acoustic features extraction. In their work, high accuracy (96.1%) has been achieved on classifying 548 10-s recordings, each of which may contain one to five bird species labels.

A common procedure for dealing with a multi-label classification problem is to transform it into single-label problems. After the transformation, any single-label classifier can be applied. In the end, individual predictions are integrated into multi-label predictions. The most common and straightforward transformation method is binary relevance. It decomposes a multi-label problem into multiple binary problems. For each label, a binary classifier is trained and used to predict the present or absent of that label. Previous work includes using k-nearest neighbour [31] and perceptrons [32]. One argument against binary relevance method is it assumes label independence. In other words, this method ignores the correlations between labels and may cause information loss. Nevertheless, a subsequent paper [33] compares binary relevance method with ensemble of classifier chains (a method which considers the correlations between labels) using several datasets. The results show that binary relevance method is not only computationally efficient, but also effective in practical applications.

2.2. Acoustic indices

What facets of biodiversity can be represented by acoustic indices are still open questions [34]. Some indices have been demonstrated to correlate with vocal bird species [35] as well as local vegetation structure [36] and can uncover the spatiotemporal heterogeneity across landscapes [37]. Endeavour has also been put into using automated techniques to group similar one-minute acoustic patterns of a 24-h audio recording [38]. The authors discuss the use of clustering techniques on acoustic indices for characterising natural soundscapes. However, each one-minute acoustic pattern can only be assigned into a particular group and the interpretation of each group requires further analysis.

We aim with this paper to propose a multi-label classification approach for detecting co-occurring patterns using acoustic indices. By co-occurring patterns, we refer to ecological dynamics that have distinct attributes and are reflected in a one-minute recording. We define five commonly encountered acoustic patterns that contain different time–frequency characteristics: Birds, Insects, Low Activity, Rain, and Wind (Fig. 1). In this study, acoustic patterns differ from acoustic events such as particular types of bird vocalisations in temporal scales. Typically more than one pattern may appear in an audio recording. Consider, for example, vocal birds during rainy and/or windy days. The use of acoustic indices and multi-instance classification enables to learn the inherent characteristics of soundscapes and predict simultaneous acoustic patterns.

3. Materials and methods

3.1. Study site and data collection

Audio recordings are collected from the Samford Ecological Research Facility (SERF), Brisbane, Australia (27.39°S, 152.88°E) on 15 October 2010. The main vegetation is comprised of inland

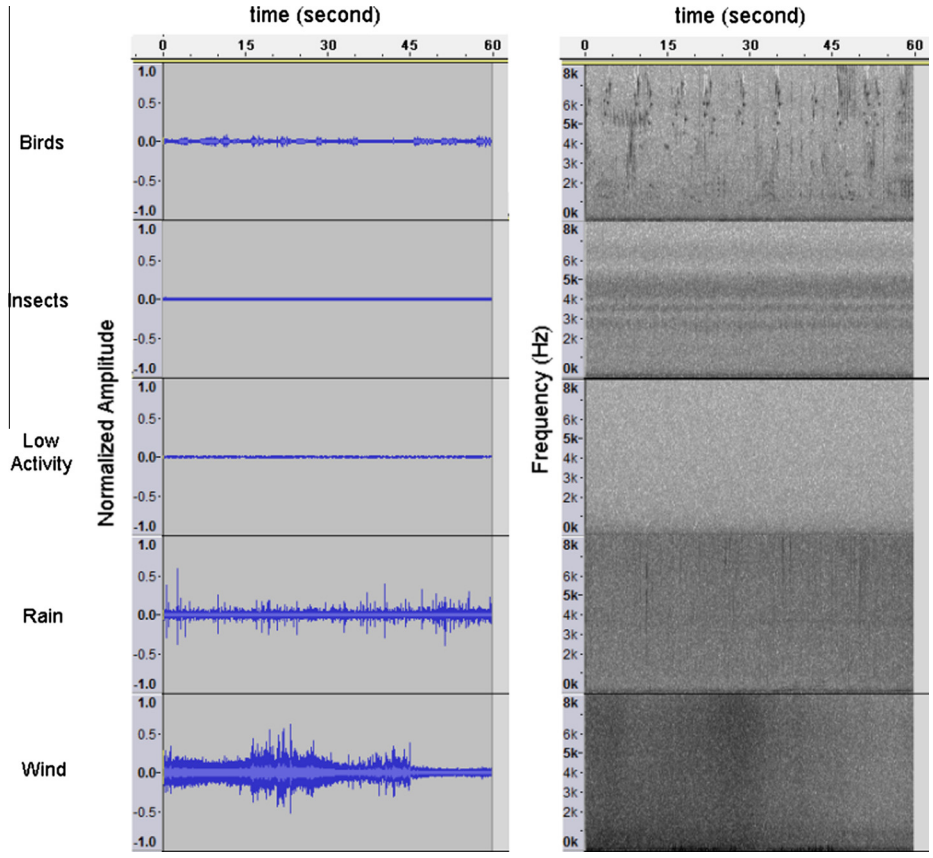

Fig. 1. Examples of five exclusive one-minute acoustic patterns have different time-frequency characteristics (left: waveforms; right: spectrograms).

open-forest and woodland including *Eucalyptus tereticornis*, *Eucalyptus crebra* and *Melaleuca quinquenervia* in moist drainage. Small areas of gallery rainforest and open pasture can be found along the southern boundary.

All recordings are recorded in stereo with a sampling rate of 22,050 Hz and 16 bits per sample. They are later down-sampled to 17,640 Hz and cut into one-minute audio clips to reduce the computational burden. The stereo signals are downmixed to a mono-signal for indices calculation. For each one-day recording, there are 1435 one-minute audio clips, excluding 5 min of data storage and acoustic sensor initialisation for the next continuous recording. One day is a reasonable ecological cycle for bird species surveys because the species compositions in a specific geological region are stable.

Each one-minute audio clip is associated with one to five labels manually annotated by the author. These labels are Birds, Insects, Low Activity, Rain, and Wind. Table 1 lists the label details associated with audio clips. Over 56% of audio clips contain multiple labels. The cardinality (average number of labels per instance) is 1.66. Particularly, the number of unique bird species in each one-minute audio clip is annotated by two experienced bird observers. There is a total of 62 bird species on this day.

Table 1
The number of audio clips associated with different combinations of labels.

Index	Number of labels	Labels	Number of audio clips
1	1	Birds	556
2	1	Insects	1
3	1	Low activity	1
4	1	Rain	63
5	1	Wind	3
6	2	Birds & Insects	19
7	2	Birds & Low activity	50
8	2	Birds & Rain	133
9	2	Birds & Wind	49
10	2	Insects & Low activity	251
11	2	Insects & Rain	127
12	2	Insects & Wind	25
13	2	Low activity & Wind	1
14	2	Rain & Wind	21
15	3	Birds & Insects & Low activity	61
16	3	Birds & Insects & Wind	6
17	3	Birds & Insects & Rain	17
18	3	Birds & Low activity & Wind	1
19	3	Birds & Rain & Wind	15
20	3	Insects & Low activity & Wind	6
21	3	Insects & Rain & Wind	25
22	4	Birds & Insects & Rain & Wind	4
23	Total		1435

3.2. Acoustic features

As with most classification tasks, extraction and selection of proper features are crucial for high classification accuracy. Although various acoustic indices have been proposed, previous papers have shown that adding more features does not always produce best results [39]. As the feature space expands, there are potentially correlated feature vectors that may introduce redundancy, hence impact the classification result. On the other hand, additional feature vectors also increase computational complexity. A forward stepwise feature selection method has been used to obtain a set of seven acoustic indices, resulting in promising

classification accuracy [40]. These acoustic indices can be categorised into two groups: temporal indices and spectral indices.

Temporal acoustic indices are derived from the waveform envelopes. Here a waveform envelope refers to the maximum consecutive 512-point non-overlapping rectangular window over the original signal. Given a time series $x(n)$ of a length N, the proposed two temporal acoustic indices are given as follows:

1. AveSignalAmplitude: It is the average amplitude of the waveform envelope. The values are converted to decibels.

$$\text{AveSignalAmplitude} = 10 \times \log_{10}\left(\frac{1}{N}\sum^{N} x(n)\right)^2 \tag{1}$$

2. Signal-to-residual ratio (SRR): It is derived from matching pursuit algorithm. Matching pursuit is a sparse approximation method that decomposes a complex waveform signal into a set of Gaussian-modulated sinusoids. The Matching Pursuit Toolkit (MPTK) has been used to implement the algorithm. To obtain a comparable time–frequency resolution with spectral indices, the parameters in MPTK are set to a 512-point Gaussian window with no overlaps. The procedure is an audio signal subtracts a sinusoid that has maximum energy reduction from a set of sinusoids generated from a Gabor function, resulting in a residual. Then the residual is used as the signal to run the same procedure iteratively until a certain number has been reached. The signal-to-residual ratio is calculated as:

$$\text{SRR} = 10 \times \log_{10}\ (\text{initial signal energy/residual energy}) \tag{2}$$

The rest of acoustic indices are derived from spectral information. Typically, a short-time Fourier transform is applied to a waveform signal. A 512-point rectangular window is used to segment the waveform signal without any overlap. Only amplitudes of the Fourier coefficients are retained. The transformed result is a matrix $\boldsymbol{S}$ of N time frames multiplied by M frequency bins. Note that the number of time frames is equal to the length of the waveform envelope and vectors and matrices are in bold. The matrix $\boldsymbol{S}$ can be described by following equations:

$$\boldsymbol{S} = (\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_N) \tag{3}$$

$$T_i = (a_1, a_2, \ldots, a_M)^{transpose} \quad i = 1, 2, \ldots, N \tag{4}$$

Here $\boldsymbol{T}$ denotes a time frame consisting of a vector of Fourier coefficients, a denotes each amplitude value. Note that multiplication between vectors or matrices are dot product. Five spectral acoustic indices include:

3. AcousticComplexity: It is the average absolute amplitude differences between adjacent time frames. The summed amplitude differences ($\boldsymbol{D}$) of all frequency bins are:

$$\boldsymbol{D} = \sum_{i=1}^{N-1} |\boldsymbol{T}_{i+1} - \boldsymbol{T}_i| \tag{5}$$

The AcousticComplexity of a spectrogram is calculated as:

$$\text{AcousticComplexity} = \frac{1}{M}\sum^{M} \boldsymbol{D} \tag{6}$$

4. AveEntropySpectrum: It is an entropy index of average amplitude calculated in each frequency bin from 482 Hz to 8820 Hz. This frequency range is selected to remove anthropophonic noise in low frequency bins. The average spectrum ($\boldsymbol{P}$) for all frequency bins is:

$$\boldsymbol{P} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{T}_i \tag{7}$$

So the entropy of average spectrum is calculated as:

$$\text{AveEntropySpectrum} = -\frac{1}{\log_2 M}\sum^{M} (\boldsymbol{P} \times \log_2 \boldsymbol{P}) \tag{8}$$

5. EntropyPeaks: It is an entropy index of amplitude that has maximum counts in each frequency bin from 482 Hz to 8820 Hz. The reason of selecting such a frequency range is the same with AveEntropySpectrum. Here, $\boldsymbol{C}$ refers to a vector of amplitudes that have maximum counts in each frequency bin.

$$\text{EntropyPeaks} = -\frac{1}{\log_2 M}\sum^{M} (\boldsymbol{C} \times \log_2 \boldsymbol{C}) \tag{9}$$

6-7. Ridge indices (verRidge and horRidge): If a spectrogram is considered as an image comprised of pixels, ridges are local maxima in at least one dimension in the spectrogram. Dong et al. [41] introduced ridge features for bird vocalization retrieval in massive acoustic data. Ridge indices used in this research are the average count of vertical and horizontal ridges of each spectrogram.

3.3. Classifiers

A binary relevance based multi-label classification is used in this study because of its scalability and flexibility [42]. As mentioned in Section 2.1, the basic idea of binary relevance approach is to consider a multi-label classification problem as multiple binary classification problems respectively. After this problem transformation, a classifier is required to solve each of the binary classification. In this study, we examine three classic single-label algorithms: k-nearest neighbour, decision tree, and multi-layer perceptron. To optimise three classifiers' overall accuracy, we implemented a grid search on the parameters. The number of nearest neighbour was evaluated ranging from 1 to 10. The classifier had the highest accuracy when there were 5 nearest neighbours. For decision tree, we examined the minimum number of instances per leaf from 2 to 6, and found the default setting of 2 provided the highest accuracy. For multi-layer perceptron, the number of nodes from 1 to 10 was examined with one hidden layer, and we found that the default setting of 6 had the highest accuracy.

A ten-fold cross validation is performed to evaluate the multi-label classification model on one-minute audio clips. All audio clips are divided into 10 groups at random. For each fold, nine groups are used for training and the remaining group is used for testing. The performance of classifier is estimated by the average values over ten folds.

3.4. Evaluations

In single-label classification problem, the prediction of each instance is either correct or incorrect. Evaluating the performance of a classifier is based on the number of correctly or incorrectly classified instances. Some standard evaluation metrics include accuracy, precision, and recall. However, in multi-label classification problem, predictions are a set of labels. Therefore, the prediction of each instance can be partially correct. None of the above mentioned evaluation metrics reflects this notion in their original forms.

To measure partially correct, one strategy is to average the differences between the predicted labels and the actual labels for all labels and instances. In this research, we select three of these measures to evaluate the performance of multi-label classification. They are ranked by the strength from weak to strong: hamming loss, accuracy, and exact match. Here x_i and y_i denote the prediction and the ground truth respectively; $|I|$ is the number of instances and $|L|$ is the number of labels.

Hamming loss accounts for the prediction error and the missing error, normalized over total number of labels and instances. It measures the average times that an instance is associated with an incorrect label. Here '*xor*' denotes exclusive or.

$$\text{HammingLoss}(x_i, y_i) = \frac{1}{|I|}\sum_{i=1}^{|I|}\frac{xor(x_i, y_i)}{|L|} \tag{10}$$

Accuracy is defined as the proportion of the correctly predicted labels to the total number of labels; this proportion is later averaged across all instances. Note that it is possible to calculate individual accuracy for each label.

$$\text{Accuracy}(x_i, y_i) = \frac{1}{|I|}\sum_{i=1}^{|I|}\frac{|x_i \cap y_i|}{|x_i \cup y_i|} \tag{11}$$

Exact match is a measure of precise match between predictions and actual labels. It extends the accuracy metric in single-label classification for multi-label problem, where partially correct predictions are considered as incorrect ones.

$$\text{ExactMatch}(x_i, y_i) = \frac{1}{|I|}\sum_{i=1}^{|I|}(x_i \equiv y_i) \tag{12}$$

The values of hamming loss, accuracy, and exact match range from 0 to 1 inclusive. For hamming loss, 0 corresponds to perfect prediction and 1 corresponds to wrong predictions for all labels of each instance; whereas for accuracy and exact match, the values have the completely opposite meanings.

4. Results

4.1. Comparison between multi-label classifiers

Table 2 illustrates the evaluation measures for two multi-label classification algorithms. We also calculate a baseline for hamming loss in order to interpret the performance of a classifier. The baseline is considered as a non-informative classifier that predicts empty label sets for all instances. Therefore, the baseline of hamming loss is calculated as the number of labels across all instances divided by the number of labels $|L|$ and the number of instances $|I|$.

Three multi-label classifiers perform more than three times better than a non-informative classifier (baseline). Note that a smaller hamming loss is better, while larger accuracy and exact match are better. From Table 2, we can see that multi-layer perceptron classifier outperforms the other two classifiers from all the evaluation metrics.

The hamming loss of a multilayer perceptron classifier is 4.18 times better than that of the baseline. This performance is as good as Briggs' classification, but our study objects (5 acoustic patterns at a one-minute resolution) differ from theirs (13 bird species at a 10-s resolution). In their work the value of exact match is not shown, but 20 randomly selected example recordings with actual labels and predictions are provided. Among them, 9 out of 20 predictions are exactly matched with the ground truth. We estimate that their exact match is approximately 0.45. By contrast, our exact match is 0.696, which means our classifier outweighs theirs. However, hamming loss shows that the performance of our study (0.079) is worse than theirs (0.039). The main reason causing such a controversial consequence, we believe, is the number of possible labels to be predicted. A large number of possible labels will lower both hamming loss and exact match according to the mathematical equations. For example, the calculation of hamming loss is required to average across the number of possible labels; whereas an increased number of possible labels make the exact prediction more difficult. Therefore, this result suggests that it is essential to use various evaluation metrics to measure the performance of multi-label classification.

Table 2
The performance of three different multi-label classifiers.

	Hamming loss	Accuracy	Exact match
ML-kNN	0.099 ± 0.008	0.827 ± 0.014	0.622 ± 0.030
ML-DecisionTree	0.090 ± 0.010	0.833 ± 0.020	0.661 ± 0.028
ML-MultilayerPerceptron	**0.079 ± 0.007**	**0.853 ± 0.014**	**0.696 ± 0.039**
Baseline	0.332	/	/

4.2. Comparison to single-label classification

The aforementioned evaluation metrics provides a convenient way to understand the performance of different classifiers. These measures are inappropriate when more subtle details are required. To determine where misclassification actually occurs, we calculate *precision* and *recall* on each of the five acoustic patterns. We also investigate single-label classification using the same dataset. The experimental settings of single-label classification such as classifiers, parameters, and cross validation are identical to those in multi-label classification. The label of each instance is determined by selecting a dominant acoustic pattern from the five possible labels. Note that, for each instance, the label in single-label classification is one of those in multi-label classification.

Fig. 2 shows the precision and recall for both single- and multi-label classifiers. Generally, all classifiers provide good performance on detecting Birds and Low activity, which are the major acoustic patterns in the dataset. Among the five acoustic patterns, precision and recall of Birds, Rain, Insects, and Low activity are higher than 0.7 except for Wind, which has the poorest classification performance. For Wind, the standard deviations of both metrics are about 0.1; for the rest four acoustic patterns, their standard deviations are less than 0.05. These standard deviations are not shown in the figures for clarity purposes. The reasons for poor classification accuracy of Wind might be insufficient training instances and inappropriate features for this particular acoustic pattern. However, current dataset is the only available one and the feature set used in this study is optimised to provide the best overall classification accuracy.

In multi-label classification, multilayer perceptron and k-nearest neighbour provide better performance than decision tree, but there are no apparent performance differences between single-label classifiers. Therefore, multi-layer perceptron classifier in multi-label classification is preferable for classifying concomitant classes in long-duration recordings. Although multi-label classifiers seem to have higher precisions and recalls than the corresponding single-label classifiers in most cases, a direct comparison between these two approaches is inappropriate because they deal with different classification problems and different number of labels for each acoustic pattern.

4.3. Investigation of bird species

Birds are important indicators of environmental health. The detection and analysis of bird species have attracted continuous attention over the years. Classifying acoustic patterns provides a potential way to remove irrelevant audio clips and improve the efficiency in such ecological studies, especially when the amount of audio recordings is huge.

Fig. 3 shows the number of bird species per minute in the minutes classified as Birds by using multi-layer perceptron classifiers. In the original 24-h recording, there are about 600 min that do not contain any species. Obviously, single-label classification enables to recognise the majority of non-bird minutes. However, a large portion of minutes containing birds are misclassified as one of

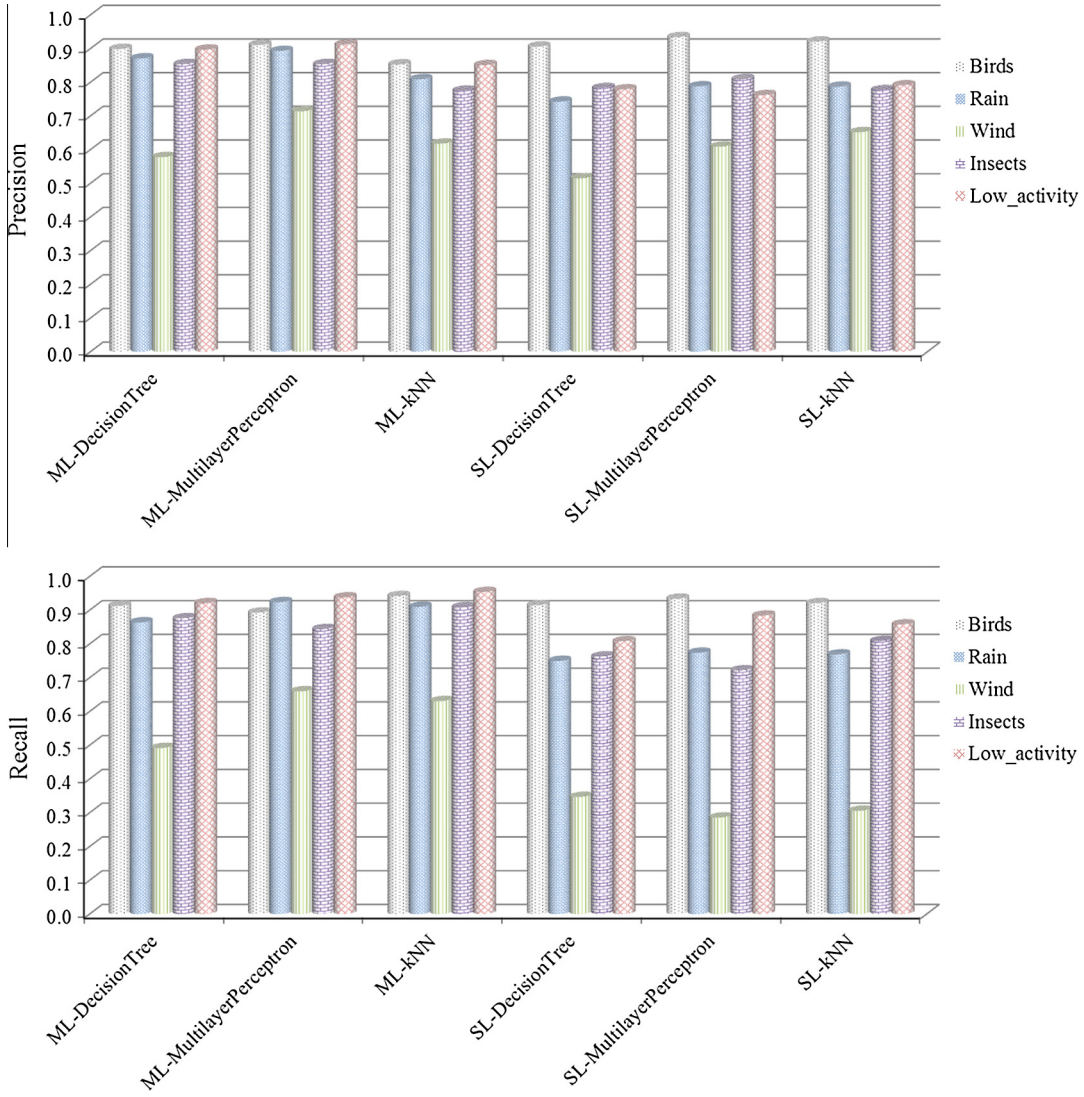

Fig. 2. Comparisons of precisions and recalls between single- and multi-label classifiers for five acoustic patterns (ML: multi-label; SL: single-label).

the other four acoustic patterns. Multi-label classification captures the minutes which contain one to seven bird species but are misclassified by the single-label method, increasing the number of true positives. Also note that the number of misclassified bird minutes (false positives) soars to 200 in multi-label classification.

We further analyse bird species loss and the efficiency in bird species surveys based on the annotations mentioned in Section 3.1. Compared to multi-label classification, where there is no species loss, single-label classification retain 59 out of 62 bird species within the classified bird minutes. Fig. 4 shows the species accumulative curves using five different methods. The theoretical best is derived from the bird annotations. The dawn sampling is proposed by Wimmer [43] where he suggested conducting bird species surveys three hours after dawn at random. Two classification methods are proposed in the current study and the baseline is derived from randomly selecting one-minute samples from a one-day recording. Each point in the curves represents the average percent of bird species found given a specific number of one-minute samples.

We can see that classification approaches (red[1] and blue) achieve higher efficiency in bird species surveys than randomly selecting minutes without any process (green) (t-test, $p < 0.001$). Take the line parallel to x axis at value 50 on y axis in Fig. 4 for example. Using classification methods, one needs to inspect 25 one-minute audio clips to find 50 percent of species on that day. However, without classification methods, 37 one-minute audio clips are required to achieve the same performance. Therefore, one can earn the time of inspecting 12 one-minute audio clips using classification methods. Most importantly, the earned time increases exponentially if more bird species are required to be found. Although multi-label classification increases the false positives (Fig. 3), no apparent difference can be found between the efficiency of finding bird species using single- and multi-label classification by calculating a pairwise t-test on both corresponding species accumulative curves ($p > 0.1$). Since single-label classification causes bird species loss, multi-label classification is a preferable approach for removing redundancy in bird species surveys.

We compared our classification methods with Wimmer's dawn sampling. He suggested that sampling audio clips at random from 3 h after dawn is an efficient strategy for bird species surveys. By implementing his dawn sampling method, we obtained another species accumulative curve (orange). This method has higher efficiency of finding bird species than our classification methods for the first 30 one-minute audio clips, but its performance decreases

[1] For interpretation of color in Fig. 4, the reader is referred to the web version of this article.

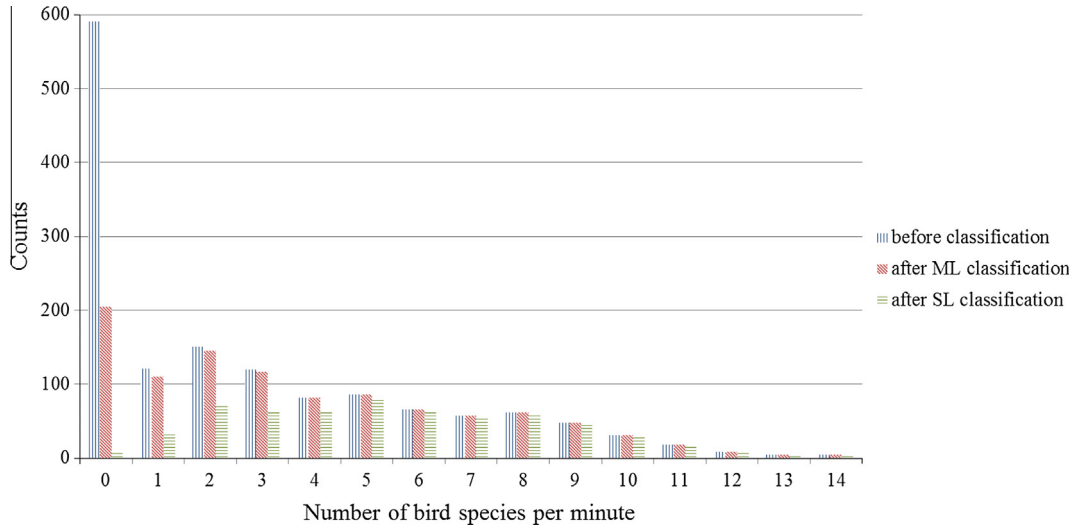

Fig. 3. Distributions of the number of bird species per minute after single- and multi-label classifications (ML: multi-label; SL: single-label). The classifier is the multilayer perceptron.

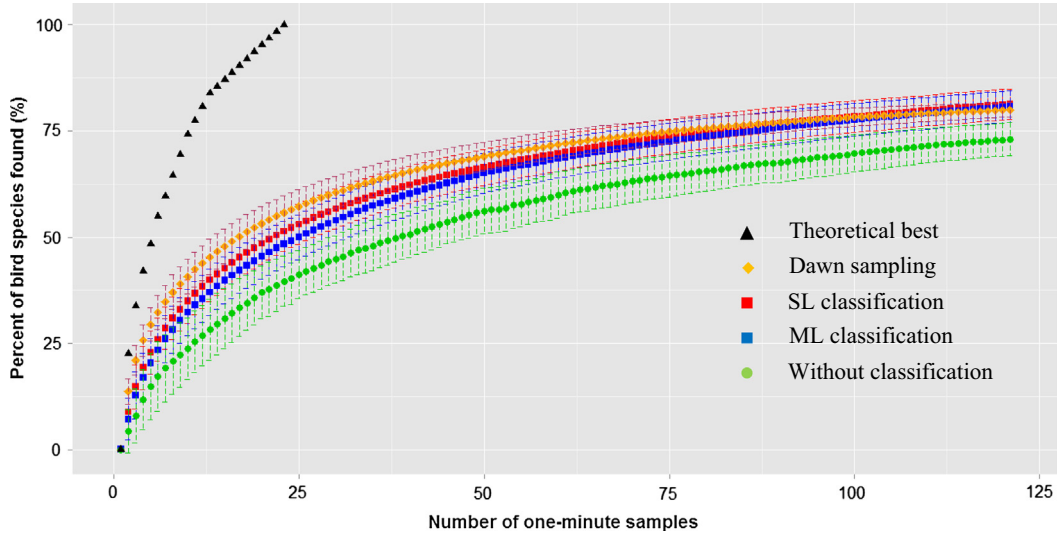

Fig. 4. Five species accumulative curves derived from different sampling methods: dawn sampling, single-label (SL) and multi-label (ML) classification, and two benchmarks.

when more audio clips are inspected. Dawn sampling is based on the prior knowledge that most bird species vocalise during morning chorus. It is susceptible to two factors: weather conditions and the time that bird species appear. For example, rain can interrupt the morning chorus and no further instructions are given for species surveys during the rest of the day. Dawn sampling also excludes species that are absent from the morning chorus. Therefore, our classification methods provide comparable efficiency in bird species surveys but are more resilient than dawn sampling in these two aspects.

5. Conclusion and future work

This paper proposed a multi-label classification method to detect simultaneous acoustic patterns in long-duration recordings. Using an in-the-field audio data set, we find that multi-label methods produce promising performance in classifying co-occurring acoustic patterns, and outperform single-label methods. Extensive information is also provided to demonstrate the advantages of using multi-label classification in a bird species study.

We introduce a novel set of acoustic indices to characterise one-minute audio clips. By yielding high accuracy in classifying five concrete acoustic patterns, we can infer that acoustic indices are proper indicators of general ecological processes. Since they are summary acoustic information over a period of time, it is inadvisable to use them for acoustic event detection. For example, they barely contain important frequency information of an acoustic event such as a bird vocalization. Nevertheless, acoustic indices can reflect general acoustic patterns in long-duration recordings and can serve as a pre-process step to remove redundant information, improving the efficiency in finding bird species. Moreover, the proposed classification method is more resilient than an empirical dawn sampling method.

One limitation of this study is that the proposed classification methods require a training data set. The performance of classifiers is dependent on the quality of these training data but the preparation of such a dataset is time-consuming. The data used in this study is a one-day recording at a specific location. Typically, a large training dataset is required to avoid overfitting of the classifier. Our field-collected data alleviate this effect to a certain extent.

This is a further step towards the understanding of complex ecological processes by using natural acoustics. In contrast with the manual analysis, automated techniques offer an efficient and scalable approach for investigating acoustic data in the long term. In our future work, the proposed method will be scaled up for analysing years or decades of audio recordings. This will effectively create presence/absence information that can facilitate the navigation through massive audio recordings.

Acknowledgements

This research was conducted with the support of the QUT Samford Ecological Research Facilities. The authors would like to acknowledge the contribution of the experts in annotating the bird species. We thank Anthony Truskinger for managing the datasets and Xueyan Dong for calculating a part of the acoustic features. The first author was supported by a joint scholarship provided by QUT and Chinese Scholarship Council.

References

[1] Hutto RL, Pletschet SM, Hendricks P. A fixed-radius point count method for nonbreeding and breeding season use. The Auk 1986;103(3):593–602.
[2] Acevedo MA, Villanueva-Rivera LJ. Using automated digital recording systems as effective tools for the monitoring of birds and amphibians. Wildl Soc Bull 2006;34(1):211–4.
[3] Pijanowski BC, Farina A, Gage SH, Dumyahn SL, Krause BL. What is soundscape ecology? An introduction and overview of an emerging new science. Landscape Ecol 2011;26(9):1213–32.
[4] Pijanowski BC, Villanueva-Rivera LJ, Dumyahn SL, Farina A, Krause BL, Napoletano BM, et al. Soundscape ecology: the science of sound in the landscape. Bioscience 2011;61(3):203–16.
[5] Farina A. Soundscape ecology: principles, patterns, methods and applications. Netherlands: Springer; 2014.
[6] Spellerberg IF, Fedor PJ. A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener' Index. Glob Ecol Biogeogr 2003;12(3):177–9.
[7] Mulder CPH, Bazeley-White E, Dimitrakopoulos PG, Hector A, Scherer-Lorenzen M, Schmid B. Species evenness and productivity in experimental plant communities. Oikos 2004;107(1):50–63.
[8] Silva RA, Martins IA, Rossa-Feres DdC. Environmental heterogeneity: Anuran diversity in homogeneous environments. Zoologia (Curitiba) 2011;28(5):610–8.
[9] Sueur J, Farina A, Gasc A, Pieretti N, Pavoine S. Acoustic indices for biodiversity assessment and landscape investigation. Acta Acustica United Acustica 2014;100(4):772–81.
[10] Sueur J, Pavoine S, Hamerlynck O, Duvail S. Rapid Acoustic Survey for Biodiversity Appraisal. PLoS ONE 2008;3(12):e4065.
[11] Pieretti N, Farina A, Morri D. A new methodology to infer the singing activity of an avian community: the acoustic complexity index (ACI). Ecol Ind 2011;11(3):868–73.
[12] Farina A, Pieretti N, Piccioli L. The soundscape methodology for long-term bird monitoring: A Mediterranean Europe case-study. Ecol Inform 2011;6(6):354–63.
[13] Towsey M, Wimmer J, Williamson I, Roe P. The use of acoustic indices to determine avian species richness in audio-recordings of the environment. Ecol Inform 2014;21:110–9.
[14] Towsey M, Zhang L, Cottman-Fields M, Wimmer J, Zhang J, Roe P. Visualization of long-duration acoustic recordings of the environment. Procedia Comput Sci 2014;29:703–12.
[15] Rey Gozalo G, Trujillo Carmona J, Barrigón Morillas JM, Vílchez-Gómez R, Gómez Escobar V. Relationship between objective acoustic indices and subjective assessments for the quality of soundscapes. Appl Acoust 2015;97:1–10.
[16] Jennings P, Cain R. A framework for improving urban soundscapes. Appl Acoust 2013;74(2):293–9.
[17] Lian-Hong C, Lie L, Hanjalic A, Hong-Jiang Z, Lian-Hong C. A flexible framework for key audio effects detection and auditory context inference. IEEE Trans Audio Speech Lang Process 2006;14(3):1026–39.
[18] Radhakrishnan R, Divakaran A, Smaragdis P. Audio analysis for surveillance applications. In: IEEE workshop on applications of signal processing to audio and acoustics. p. 158–61.
[19] Chu S, Narayanan S, Kuo CCJ. Environmental sound recognition with time-frequency audio features. IEEE Trans Audio Speech Lang Process 2009;17(6):1142–58.
[20] Ma L, Milner B, Smith D. Acoustic environment classification. ACM Trans Speech Lang Process 2006;3(2):1–22.
[21] Shamir L, Yerby C, Simpson R, von Benda-Beckmann AM, Tyack P, Samarra F, et al. Classification of large acoustic datasets using machine learning and crowdsourcing: application to whale calls. J Acoust Soc Am 2014;135(2):953–62.
[22] Briggs F, Raich R, Fern XZ. Audio classification of bird species: a statistical manifold approach. Nineth IEEE international conference on data mining, Miami, 2009. p. 51–60.
[23] Bardeli R, Wolff D, Kurth F, Koch M, Tauchert KH, Frommolt KH. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. Pattern Recogn Lett 2010;31(12):1524–34.
[24] De Oliveira AG, Ventura TM, Ganchev TD, De Figueiredo JM, Jahn O, Marques MI, et al. Bird acoustic activity detection based on morphological filtering of the spectrogram. Appl Acoust 2015;98:34–42.
[25] Chen Y, Why A, Batista G, Mafra-Neto A, Keogh E. Flying insect classification with inexpensive sensors. J Insect Behav 2014;27(5):657–77.
[26] Nam J, Kim J, Loza Mencía E, Gurevych I, Fürnkranz J. Large-scale multi-label text classification — revisiting neural networks. Mach Learn Knowl Disc Databases 2014;8725:437–52.
[27] Markatopoulou F, Mezaris V, Kompatsiaris I. A comparative study on the use of multi-label classification techniques for concept-based video indexing and annotation. MultiMedia Model 2014;8325:1–12.
[28] Cakir E, Heittola T, Huttunen H, Virtanen T. Polyphonic sound event detection using multi label deep neural networks. International joint conference on neural networks (IJCNN), Killarney, 2015. p. 1–7.
[29] Fabris F, Freitas AA. Dependency network methods for hierarchical multi-label classification of gene functions. IEEE symposium on computational intelligence and data mining (CIDM), Orlando, 2014. p. 241–8.
[30] Briggs F, Lakshminarayanan B, Neal L, Fern XZ, Raich R, Hadley SJK, et al. Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. J Acoust Soc Am 2012;131(6):4640–50.
[31] Spyromitros E, Tsoumakas G, Vlahavas I. An empirical study of lazy multilabel classification algorithms. Artif Intel: Theories, Models Appl 2008;5138:401–6.
[32] Fürnkranz J, Hüllermeier E, Loza Mencía E, Brinker K. Multilabel classification via calibrated label ranking. Mach Learn 2008;73(2):133–53.
[33] Luaces O, Díez J, Barranquero J, del Coz J, Bahamonde A. Binary relevance efficacy for multilabel classification. Prog Artif Intel 2012;1(4):303–13.
[34] Servick K. Eavesdropping on ecosystems. Science 2014;343(6173):834–7.
[35] Depraetere M, Pavoine S, Jiguet F, Gasc A, Duvail S, Sueur J. Monitoring animal diversity using acoustic indices: implementation in a temperate woodland. Ecol Ind 2012;13(1):46–54.
[36] Farina A, Pieretti N. Sonic environment and vegetation structure: a methodological approach for a soundscape analysis of a Mediterranean maqui. Ecol Inform 2014;21:120–32.
[37] Rodriguez A, Gasc A, Pavoine S, Grandcolas P, Gaucher P, Sueur J. Temporal and spatial variability of animal sound within a neotropical forest. Ecol Inform 2014;21:133–43.
[38] Sankupellay M, Towsey M, Truskinger A, Roe P. Visual fingerprints of the acoustic environment: the use of acoustic indices to characterise natural habitats. IEEE international symposium on big data visual analytics, Tasmania, Australia, 2015.
[39] Chu S, Narayanan S, Kuo CCJ, Mataric MJ. Where am I? Scene recognition for mobile robots using audio features. IEEE international conference on multimedia and expo, Toronto, Canada, 2006. p. 885–8.
[40] Zhang L, Towsey M, Zhang J, Roe P. Computer-assisted sampling of acoustic data for more efficient determination of bird species richness. IEEE international workshop on environmental acoustic data mining, Atlantic City, New Jersey, 2015.
[41] Dong X, Towsey M, Zhang J, Banks J, Roe P. A novel representation of bioacoustic events for content-based search in field audio data. Digit Image Comput: Techn Appl (DICTA), 2013. p. 1–6.
[42] Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. Mach Learn 2011;85(3):333–59.
[43] Wimmer J, Towsey M, Roe P, Williamson I. Sampling environmental acoustic recordings to determine bird species richness. Ecol Appl 2013;23(6):1419–28.