Ecological Informatics 33 (2016) 75–84

Contents lists available at ScienceDirect

Ecological Informatics

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

Enhancing the dissimilarity-based classification of birdsong recordings

José Francisco Ruiz-Muñoz [a],*, German Castellanos-Dominguez [a], Mauricio Orozco-Alzate [a,b]

[a] *Universidad Nacional de Colombia, Sede Manizales, Signal Processing and Recognition Group, km 7 vía al Magdalena, Manizales 170003, Colombia*
[b] *Universidad Nacional de Colombia, Sede Manizales, Departamento de Informática y Computación, km 7 vía al Magdalena, Manizales 170003, Colombia*

ARTICLE INFO

Article history:
Received 18 December 2015
Received in revised form 5 April 2016
Accepted 6 April 2016
Available online 13 April 2016

Keywords:
Multiple instance classification
Dissimilarity-based classification
Bioacoustics

ABSTRACT

Classification of birdsong recordings can be naturally formulated as a multiple instance problem, where bags of instances are represented by either features or dissimilarities. In bioacoustics, bags typically correspond to regions of interest in spectrograms, which are detected after a segmentation stage of the audio recordings. In this paper, we use different dissimilarity measures between bags and explore whether the subsequent application of metric learning/adaptation methods and the construction of dissimilarity spaces allow increasing the classification performance of birdsong recordings. A publicly available bioacoustic data set is used for the experiments. Our results suggest, in the first place, that appropriate dissimilarity measures are those which capture most of the overall differences between bags, such as the modified Hausdorff distance and the mean minimum distance; in the second place, they confirm the benefit from adapting the applied dissimilarity measure as well as the potential further enhancement of the classification performance by building dissimilarity spaces and increasing training set sizes.

1. Introduction

Wildlife monitoring is very often related to the collection, analysis, and identification of bioacoustic signals coming from several species, which are heard more often than seen or even trapped (Brandes, 2008). With the aim of alleviating the repetitive and labor-intensive tasks derived from wildlife monitoring, biologists and ecologists have recently turned their attention to automatic pattern recognition. Among the specific advantages of the acoustic-based monitoring approach, the following are worth mentioning: i) relative easiness and cheapness for collecting acoustic information by using digital audio recording devices (Potamitis et al., 2014); ii) feasibility of acquiring a vocal activity during extended periods of time, allowing large scale coverages along both time and space domains (Kasten et al., 2012; Frommolt and Tauchert, 2014); and iii) ability to tackle challenges of labeling the enormous amount of available bioacoustic data, whose analysis might be too costly or even infeasible to be carried out by human experts (Ross and Allen, 2014). Consequently, automated bioacoustic monitoring becomes cheaper in the long term than the observations made by experts, providing even sometimes more accurate results (Hao et al., 2013).

In the field of bioacoustics, many approaches of signal processing and pattern recognition have been applied to the problem of automatic bird detection and classification. These approaches include time-frequency feature extraction, analysis of specific vocalization properties, dissimilarities between acoustic signals or their representations, and statistical classifiers. Earlier studies focused on the classification of syllables and songs: in Härmä (2003), the problem of classifying passerine bird syllables is studied by using a sinusoidal modeling and classification by matching, i.e., the one-nearest neighbor (1-NN) classification rule. A similar approach is proposed by Chen and Maher (2006) to tackle the bird strike avoidance problem in aviation. Somervuo et al. (2006) compare three feature sets: sinusoidal modeling, mel-frequency cepstrum coefficients (MFCC) and descriptive features, and use three classification techniques: 1-NN based on the dynamic time warping (DTW) distance, Gaussian mixture models (GMM) and hidden Markov models (HMM). In Fagerlund (2007), vocalizations are represented by MFCC and descriptive features and classification is carried out by using a decision tree with a support vector machine (SVM) classifier. Trifa et al. (2008) apply HMM for classifying songs of antbirds of a Mexican rainforest represented by MFCC and linear predictive coding (LPC). Likewise, Acevedo et al. (2009) propose a methodology for automatically classifying isolated calls of three common mountain bird species by using standard call variables and spectral features, and three classifiers: linear discriminant analysis, decision tree, and SVM.

Recently, the problems of classification of recordings and detection in continuous audio signals have been studied to face realistic problems. In Briggs et al. (2009), recordings of 6 species from the Cornell Macaulay Library are classified by using frame-level feature histogram representation and the proposed 1-NN on statistical manifolds. Briggs et al. (2012) propose a multi-instance multi-label classification framework for classifying birdsong recordings of the H. J. Andrews data set, which consists in the representation of each audio signal as a bag-of-instances and its

* Corresponding author.
E-mail addresses: jfruimu@unal.edu.co (J.F. Ruiz-Muñoz), cgcastellanosd@unal.edu.co (G. Castellanos-Dominguez), morozcoa@unal.edu.co (M. Orozco-Alzate).
URL: https://sites.google.com/a/unal.edu.co/mauricio/ (M. Orozco-Alzate).

http://dx.doi.org/10.1016/j.ecoinf.2016.04.001
1574-9541/© 2016 Elsevier B.V. All rights reserved.

classification using a SVM. Potamitis (2014) classifies the recordings of the Multi-label bird species classification challenge-NIPS 2013 by detecting bags of relevant segments from spectrograms and uses image-based features with a random forest classifier. Stowell and Plumbley (2014) introduce the concept of unsupervised feature learning for classifying recordings of four data sets of bird recordings from France, UK and Brazil. Among the detection studies, we highlight the following ones: Bardeli (2009) proposes a methodology of similarity search in audio recordings by using time–frequency trajectories and evaluates this approach with recordings of the Animal Sound Archive of Berlin. In a later study, Bardeli et al. (2010) apply a similar approach for detecting vocalizations of the Eurasian bittern and Savi's warbler. Potamitis et al. (2014) use time–frequency features and HMM for detecting bird species of North America, Eurasia, and North Africa. Ganchev et al. (2015) detect vocalizations of the *Vanellus chilensis lampronotus* extracting spectral features and using GMM and HMM. In some studies, species-specific parametrization is carried out, e.g.,, the methodology for detecting vocalization of a Hawaiian forest bird proposed by Sebastián-González et al. (2015).

Audio recordings are often treated as images by using spectrograms. In such a way, acoustic events appear as blobs in these two-dimensional representations. Therefore, any framework of image analysis can be followed. In addition to some of the studies mentioned above, in the following ones the audio recognition task is transformed into an image processing and classification problem: Jančovič and Kküer (2011) rely on the spectral shape to detect tonal bird sounds in noisy environments. In Aide et al. (2013), as a first stage in the recognition system, regions of spectrograms are automatically selected. Similarly, the detection system proposed by Ventura et al. (2015) extracts image-based features to classify bird species from Brazil.

In general, automatic recognition systems require an adequate *representation* of the objects or events to be recognized as well as accurate *classification* rules (Pekalska and Duin, 2002). In the particular case of bioacoustic applications, we group the options to represent the segmented and preprocessed recordings into two categories, namely i) feature-based representations and ii) dissimilarity-based representations. The most common alternatives for feature-based representations are feature vectors and bags of feature vectors (so-called bags of instances). The former is the classic representation that consists in the extraction of a set of characteristic and hopefully discriminative descriptors from each recording. Notice that feature vectors and instances refer to the same concept; however, in order to maintain consistency with the literature, we prefer the word instances hereafter. The other option, *bags of instances* (Li et al., 2013), represents each object as a set of feature vectors. In more detail, the representation by bags of instances allows representing each audio recording (one object) as a bag of regions from its spectrogram which are typically detected by an automatic procedure (see the procedure described in Section 2.2). It is worth clarifying that the segmentation algorithm may fail — in isolated cases, as indicated by Briggs et al. (2013) — when calls overlap and detect only one segment, instead of two, that represents two species. However, in this non-classical representation by bags of instances, it is not required that all regions exclusively belong to the target class, since a bag is positive if at least one of its instances is positive. In other words, a positive bag might contain some instances not associated to the target class. As explained in Cheplygina et al. (2015), the relative advantage of the bags of instances is that they are a flexible representation that allows preserving more information than a single feature vector representation. However, this representation increases the complexity of the classification stage. On the other hand, in dissimilarity-based representations, each object is described by a number of dissimilarity values, regarding its relative differences against a set of pre-selected ones. This representation is used by Keen et al. (2014) to compare information provided by several dissimilarity measures between bird calls — as whole units.

Bags of instances and dissimilarities have been very actively researched during the last years. Among their enhancement proposals, the following two are especially promising for simplifying the bag-of-instances classification process and improving the dissimilarity-based classification, respectively: i) to compute bag dissimilarities so that a single vector holds all pairwise dissimilarity values between each bag and a set of other bags selected beforehand. Therefore, the bag-of-instances problem is cast into a dissimilarity-based task while preserving its original representational power (Tax et al., 2011). ii) To optimize or adapt[1] a given dissimilarity measure by using the information from a training set (Duin et al., 2014). The first proposal might be further enhanced by applying the latter to it, that is, by optimizing or adapting dissimilarity measures between bags. Therefore, we propose such an adaptation for classifying birdsong recordings represented as multiple instance objects, resulting a classification strategy that takes the advantages from both approaches.

The basic outline of this paper is as follows: representation and classification methods are described in Section 2. The experiments and obtained results are described in Section 3 and discussed further in Section 4. Lastly, we present our concluding remarks in Section 5. Table 1 summarizes the notation used in this paper.

2. Methods

Our methodology is based on the multiple instance classification (MIC) approach and consists in the following four stages that are explained below: i) a preprocessing stage to extract bags of instances from the spectrograms computed for birdsong recordings; ii) selection of a dissimilarity measure between the estimated bags; iii) enhancement of the dissimilarity representation using metric learning and dissimilarity space approaches and *iv*) classification using either the 1-NN algorithm or a trained classifier in the dissimilarity space. According to the different configurations for this methodology, we formulate four classification strategies that are described at the end of this section.

2.1. Multiple instance classification

Given an input data set $\tilde{S} = \{\mathbf{s}_n \in \mathbb{R}^L : n = 1, \ldots, N\}$ that holds N objects (instances), described by L extracted features, their respective labels $\tilde{Y} = \{\tilde{y}_n \in \{-1, 1\}\}$ are estimated in standard two-class classification tasks. So, the classifier $\tilde{S} \rightarrow \tilde{Y}$ assigns the corresponding class label to each new incoming instance. In contrast, multiple instance classification (MIC) methods represent every object by a *bag-of-instances* $\mathbf{S}_n = \{\mathbf{s}_{nm} \in \mathbb{R}^L : m = 1, \ldots, N_n\}$ that includes N_n instances $\mathbf{s}_{nm}$. In that case, given the input data set $S = \{\mathbf{S}_n \in \mathbb{R}^{N_n \times L}\}$ and the two-class label set $Y = \{y_n \in \{-1, 1\}\}$, the MIC classifiers are designed for assigning a single label to each query bag of instances: $S \rightarrow Y$.

MIC methods, depending on the level where they hold discriminant information, can be grouped into two broad categories (Amores, 2013): instance level methods and bag level methods. The former category, for which objects are instances in the representation space, mostly focus on modeling the class probability of each instance; afterwards, the bag-level classification is carried out by an additional set of rules, which combine the results of instance classification. Methods in the latter category take into account information about global properties of bags represented in the bag space, avoiding an additional step for bag level classification. In turn, the bag level methods are grouped into two types as follows:

- *Dissimilarities between bags*: a dissimilarity function is defined to compare any two bags to be classified by a dissimilarity-based approach, e.g., by the k-nearest neighbor (k-NN) rule.
- *Embedded-space*: a mapping function extracts information from each

[1] Here, the term *adaptation* refers to procedures carried out in the training stage where the original dissimilarity values are modified to improve their discriminant ability.