Dissimilarity-based multiple instance classification and dictionary learning for bioacoustic signal recognition

Clasificación multi-instancia basada en disimilitudes y aprendizaje de diccionarios para el reconocimiento de señales bioacústicas

Dissimilarity-based multiple instance classification and dictionary learning for bioacoustic signal recognition

José Francisco Ruiz Muñoz

Universidad Nacional de Colombia
Facultad de Ingeniería y Arquitectura,
Departamento de Ingeniería Eléctrica, Electrónica y Computación
Manizales, Colombia
2017

Dissimilarity-based multiple instance classification and dictionary learning for bioacoustic signal recognition

José Francisco Ruiz Muñoz

Tesis presentada como requisito parcial para optar al título de:
Doctor en Ingeniería - Automática

Director:
Dr.Ing. Mauricio Orozco-Alzate

Línea de Investigación:
Aprendizaje de Máquina
Grupo de Investigación:
Grupo de Control y Procesamiento Digital de Señales

Universidad Nacional de Colombia
Facultad de Ingeniería y Arquitectura,
Departamento de Ingeniería Eléctrica, Electrónica y Computación
Manizales, Colombia
2017

A mi familia

Acknowledgements

I would like to thank my advisor Mauricio Orozco-Alzate for his guidance and valuable suggestions.

I am grateful to Dr. Raviv Raich, who supervised my internship during six months at Oregon State University.

I thank Prof. Dr. Germán Castellanos-Domínguez and the members of the Signal Processing and Recognition Group for their collaboration on handling academic as well as administrative issues. Also, I thank the juries of my thesis, Dr. Claudia Victoria Isaza Narváez, Dr. Juan Bernardo Gómez Méndoza, and Dr. T. Mitchell Aide, for their comments and suggestions in the evaluation process.

This research was financially supported by "Programa COLCIENCIAS Doctorado Nacional 567," "Sistema Nacional de Becas de la Universidad Nacional de Colombia," "Programa de Internacionalización del Conocimiento - Modalidad 3 de la Universidad Nacional de Colombia - Sede Manizales," and "Departamento de Ingeniería Eléctrica, Electrónica y Computación de la Universidad Nacional de Colombia - Sede Manizales."

This thesis is dedicated to my family, my granny Beatriz, my aunt Luz, my mom Cecilia, my dad Bayardo, my brothers Diego and Kike, and my girlfriend Laura Sofía. Their love and encouragement were the force that I needed to carry out this work.

Agradecimientos

Agradezco sinceramente a mi director, Mauricio Orozco Alzate, por su asesoría y valiosos consejos durante el desarrollo de mi proyecto de tesis.

Expreso mi gratitud con el Dr. Raviv Raich, quien supervisó la pasantía académica que realicé en Oregon State University durante seis meses.

Gracias al Prof. Dr. Germán Castellanos Domínguez y a los miembros del Grupo de Procesamiento y Reconocimiento de Señales por su colaboración en asuntos académicos y administrativos. También agradezco a la profesora Dra. Claudia Victoria Isaza Narváez, y a los profesores Dr. Juan Bernardo Gómez Méndoza y Dr. T. Mitchell Aide, quienes conformaron el comité de jurados que evaluó mi tesis.

Esta investigación fue financiada por el Programa COLCIENCIAS Doctorado Nacional 567, el Sistema Nacional de Becas de la Universidad Nacional de Colombia, el Programa de Internacionalización del Conocimiento - Modalidad 3 de la Universidad Nacional de Colombia - Sede Manizales y el Departamento de Ingeniería Eléctrica, Electrónica y Computación de la Universidad Nacional de Colombia - Sede Manizales.

Dedico esta tesis muy especialmente a mi familia, a mi abuela Beatriz, mi tía Luz, mi mamá Cecilia, mi papá Bayardo, mis hermanos Diego y Kike, y mi novia Laura Sofía. Gracias a su amor, apoyo y motivación, he tenido fuerza para realizar este trabajo y culminar esta etapa de mi vida.

Abstract

In this thesis, two promising and actively researched fields from pattern recognition (PR) and digital signal processing (DSP) are studied, adapted and applied for the automated recognition of bioacoustic signals: (i) learning from weakly-labeled data, and (ii) dictionary-based decomposition. The document begins with an overview of the current methods and techniques applied for the automated recognition of bioacoustic signals, and an analysis of the impact of this technology at global and local scales. This is followed by a detailed description of my research on studying two approaches from the above-mentioned fields, multiple instance learning (MIL) and dictionary learning (DL), as solutions to particular challenges in bioacoustic data analysis. The most relevant contributions and findings of this thesis are the following ones: 1) the proposal of an unsupervised recording segmentation method of audio birdsong recordings that improves species classification with the benefit of an easier implementation since no manual handling of recordings is required; 2) the confirmation that, in the analyzed audio datasets, 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; 3) the adoption of dissimilarity adaptation techniques for the enhancement of dissimilarity-based multiple instance classification, along with the potential further enhancement of the classification performance by building dissimilarity spaces and increasing training set sizes; 4) the proposal of a framework for solving MIL problems by using the one nearest neighbor (1-NN) classifier; 5) a novel convolutive DL method for learning a representative dictionary from a collection of multiple-bird audio recordings; 6) such a DL method is successfully applied to spectrogram denoising and species classification; and, 7) an efficient online version of the DL method that outperforms other state-of-the-art batch and online methods, in both, computational cost and quality of the discovered patterns.

Keywords: pattern recognition, digital signal processing, multiple instance learning, dictionary learning, bioacoustics.

Resumen

En esta tesis se estudian, adaptan y aplican dos prometedoras y activas áreas del reconocimiento de patrones (PR) y procesamiento digital de señales (DSP): (i) aprendizaje débilmente supervisado y (ii) descomposiciones basadas en diccionarios. Inicialmente se hace una revisión de los métodos y técnicas que actualmente se aplican en tareas de reconocimiento automatizado de señales bioacústicas y se describe el impacto de esta tecnología a escalas nacional y global. Posteriormente, la investigación se enfoca en el estudio de dos técnicas de las áreas antes mencionadas, aprendizaje multi-instancia (MIL) y aprendizaje de diccionarios

(DL), como soluciones a retos particulares del análisis de datos bioacústicos. Las contribuciones y hallazgos más relevantes de esta tesis son los siguientes: 1) se propone un método de segmentación de grabaciones de audio que mejora la clasificación automatizada de especies, el cual es fácil de implementar ya que no necesita información supervisada de entrenamiento; 2) se confirma que, en los conjuntos de datos analizados, las medidas de disimilitudes que capturan las diferencias globales entre bolsas funcionan apropiadamente, tales como la distancia modificada de Hausdorff y la distancia media de los mínimos; 3) la adopción de técnicas de adaptación de disimilitudes para mejorar la clasificación multi-instancia, junto con el incremento potencial del desempeño por medio de la construcción de espacios de disimilitudes y el aumento del tamaño de los conjuntos de entrenamiento; 4) se presenta un esquema para la solución de problemas MIL por medio del clasificador del vecino más cercano (1-NN); 5) se propone un método novedoso de DL, basado en convoluciones, para el aprendizaje automatizado de un diccionario representativo a partir de un conjunto de grabaciones de audio de múltiples vocalizaciones de aves; 6) dicho método DL se utiliza exitosamente como técnica de reducción de ruido en espectrogramas y clasificación de grabaciones bioacústicas; y 7) un metódo DL, de procesamiento en línea, que supera otros métodos del estado del arte en costo computacional y calidad de los patrones descubiertos.

Palabras clave: reconocimiento de patrones, procesamiento digital de señales, aprendizaje multi-instancia, aprendizaje de diccionarios, bioacústica.

List of Figures

List of Tables

Contents

[1] This chapter was published as: Paula Catalina Caycedo-Rosales, José Francisco Ruiz-Muñoz and Mauricio Orozco-Alzate. Automated recognition of bioacoustic signals: A review of methods and applications. In Ingeniería y Ciencia ISSN: 1794-9165 ed: Universidad Eafit vol. 9, pages 171-195, 2013.

[2] This chapter was published as: José Francisco Ruiz-Muñoz, Mauricio Orozco-Alzate, and Germán

Castellanos-Domínguez, Multiple instance learning-based birdsong classification using unsupervised recording segmentation, in Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015, 2015, pp. 2632–2638.

[3] This chapter was published as: José Francisco Ruiz-Muñoz, Mauricio Orozco-Alzate, and Germán Castellanos-Domínguez. Enhancing the dissimilarity-based classification of birdsong recordings in Ecological Informatics, Elsevier, Vol. 33, 2016, pp. 75–84.

[4] This chapter was published as: José Francisco Ruiz-Muñoz, Zeyu You, Raviv Raich, and Xiaoli Z. Fern. Dictionary Learning for Bioacoustics Monitoring with Applications to Species Classification in Journal of Signal Processing Systems, Springer US, 2016, pp. 1–15.

[5] This chapter was accepted for presentation on *The 42nd IEEE International Conference on Acoustics, Speech and Signal Processing* ICASSP 2017, March 5-9, 2017, New Orleans, USA.

1. Introduction

1.1. Context of the research

With the aim of alleviating the repetitive and labor-intensive tasks derived from wildlife monitoring, biologists and ecologists have recently turned their attention to new technologies. Especially, pattern recognition (PR) and digital signal processing (DSP) techniques facilitate the assessment of biodiversity conservation through automated recognition processes [3, 12]. Therefore, despite the fact that automated recognition is not a new concept —because, since decades ago, automatics has been applied on computation and robotics [26]— it is recently starting to be popular in bioacoustic applications. Moreover, the multidisciplinary collaboration among experts in bioacoustics, taxonomy, ecology, computer science and electronics has encouraged new advances in tools for automatically analyzing bioacoustic data.

In this thesis, the following state-of-the-art PR and DSP approaches are applied and adapted to handle challenges in bioacoustics: i) multiple instance learning (MIL), which is a weakly supervised technique that reduces the training effort since it does not require explicit *a priori* information of the concept [38, 47], and ii) dictionary learning (DL), which is used for searching time-varying patterns in audio signals [82, 102].

1.2. Objectives

This thesis aims to facilitate the automated recognition of bioacoustic signals by adapting and improving state-of-the-art PR and DSP techniques. To this end, the following aspects have been considered:

– **Identifying state-of-the-art PR and DSP techniques applicable to automated recognition of bioacoustic signals:** an overall review of the literature was carried out and published in [21]. In Chapter 2, an up-to-date version of it is presented.

- **Enhancing dissimilarity-based multiple instance classification:** we explored the MIL approach since it conveniently allows reducing the information needed for training classification systems. We combine both, dissimilarity-based multi-instance classification and metric-learning, in order to tackle the MIL problem as a standard classification problem and enhance the information provided by dissimilarity estimations. We published the results of this study in [95].

- **Detecting relevant time-frequency patterns:** we proposed a new convolutive dictionary-learning approach to detect fundamental time-frequency patterns. It was published in [96], showing its performance in both artificial and real-world data.

1.3. Contributions and outline

This thesis contributes to the fields of PR and DSP, and their applications to bioacoustics. Particularly, a framework for enhancing the dissimilarity-based multiple instance classification approach was proposed. This formulation allows both obtaining a good classification performance and facilitating MIL tasks in bioacoustics. Furthermore, a novel convolutive DL method was proposed to learn a representative dictionary from a collection of spectrograms.

The remainder of this document is organized as follows: Chapter 2 describes an analysis of challenges and opportunities for the application of bioacoustic technology in Colombia. Afterward, this thesis is divided into two parts: Part I contains the following three chapters related to MIL: Chapter 3 describes the procedure for applying MIL to bioacoustic recognition tasks and compares two segmentation methods; a baseline supervised method against a novel unsupervised method; in Chapter 4, metric learning techniques are applied to adapt dissimilarity measures between spectrograms, which are represented as bags of feature vectors also known as *bags of instances*; and, in Chapter 5, a method to enhance the one nearest classification on multiple instance datasets is proposed. Part II contains the following two chapters related to DL: in Chapter 6, we propose an efficient approach for learning and using a sparse convolutive model to represent a collection of spectrograms; subsequently, in Chapter 7, we present an online method to learn recurring time-frequency patterns from spectrograms. Finally, general conclusions of this thesis are drawn in Chapter 8 along with some recommendations and guidelines for future work.

2. Automated recognition of bioacoustic signals

Abstract

Among the most widely used methods to perform surveys, characterizations and monitoring of wildlife is the acquisition and analysis of acoustic signals that animals emit to communicate. DSP and PR methods offer promising opportunities for the automatic and remote bioacoustic monitoring of sound-emitting animals such as insects, fishes, frogs, birds, and mammals. During the past decade, numerous research studies and applications on automated bioacoustic monitoring have been published; however, such studies are scattered in the literature of engineering and life sciences. This chapter presents a review of fundamental concepts of automated acoustic monitoring. We aim to compare and categorize —in a taxonomy of DSP/PR techniques— the contributions of published research studies and applications in order to guide future research and highlight challenges and opportunities related to the deployment of this technology in Colombia.

2.1. Introduction[1]

Wildlife monitoring is carried out in order to allow that researchers and conservation planners acquire a good overview of the state and the outlook of the environment [77]. This activity is often related to the collection, analysis, and identification of bioacoustic signals coming from several species, which are frequently heard more than seen or even trapped [16]. Among the specific advantages of the acoustic-based monitoring approach, the following ones are worth mentioning: i) relative easiness and cheapness for collecting acoustic information by using digital audio recording devices [90]; ii) feasibility for acquiring acoustic signals

[1] This chapter was published as: Paula Catalina Caycedo-Rosales, José Francisco Ruiz-Muñoz and Mauricio Orozco-Alzate. Automated recognition of bioacoustic signals: A review of methods and applications. In Ingeniería y Ciencia ISSN: 1794-9165 ed: Universidad Eafit vol. 9, pages 171-195, 2013.

during extended periods of time, allowing large scale coverages along both time and space domains [63, 48]; and iii) ability to tackle the challenge of labeling the enormous amount of available bioacoustic data, whose analysis might be too costly or even non-feasible to be carried out by human experts [94]. Consequently, automated bioacoustic monitoring becomes cheaper in the long term than the observations made by experts, sometimes providing even more accurate results [53].

2.1.1. Importance of acoustic communication in wildlife

Two of the most commonly known functions of bioacoustic signals are the attraction of mating animals and defense of territory [29]. Both of these aspects strongly affect the genetic flow and the distribution patterns of the species [51]. Furthermore, acoustic signals emitted by taxa, such as insects, frogs, birds, fish, and mammals, strengthen the differences among species [109]. For instance, those signals have been used in several conservation studies because they are very efficient to confirm the presence of particular species, especially, in environments with reduced visibility (e.g., rainforest, aquatic ecosystems) and to study nocturnal species [10, 32, 45, 79, 81]. Additionally, acoustic signals have been used for identifying gender, age, and individuals [50, 67].

2.1.2. Bioacoustics and technology

In the field of bioacoustics, many approaches have been proposed for analyzing time-frequency patterns. For instance, the spectrogram representation of audio signals has been widely used [20]. However, analyzing this data by visual inspection might be non-feasible due to the amount of data that is usually required to be processed.

Technological advances allow for the application of new approaches and devices for collecting bioacoustic information [10]. For example, the autonomous recording units (ARUs) that are specialized hardware to record and store a large amount of data [5]. Such data can be transmitted, processed and analyzed with computational tools. Thus, the automated processing of acoustic data facilitates the organization and search of information [64]. Furthermore, it is possible to incorporate detection and location systems of sound-emitting animals based on DSP/PR techniques [69]. Also, the extracted information can be used for answering a broad range of questions about individuals, populations, communities and ecosystems [10, 36].

2.2. Collecting data

The acoustic monitoring is a non-invasive method, which implies that it does not require capturing specimens. Furthermore, technological tools can strengthen this type of observation [27], reducing costs and obtaining a wide spatial and temporal coverage. The primary devices used for acquiring bioacoustic signals are microphones, audio recorders, batteries, mechanisms for initiating and ending recordings, and weather-proof housing for the equipment [16]. There are specialized microphones used according to the physical properties of the signals and the environmental conditions. For instance, infrasound microphones are used for recording elephant and cetacean vocalizations whereas ultrasonic microphones are used for recording vocalizations of bats and some insects. Some technical specifications and recommendations about collecting devices are found in [12, 26, 27].

Currently, there is a wide variety of devices for recording bioacoustic signals. They vary in price, sound sensitivity, signal-to-noise ratio (SNR), quality and design of hardware, configuration, directionality of microphones, frequency sensitivity, and programmability of recording schedules [92]. The ARUs are an example of these devices, they are programmable audio recorders that can be handled automatically or by a wireless connection [78]. The most widely used ARUs are songmeters,[3] E3A recording systems [56, 111], and smartphones[4].

In monitoring programs, due to the technological differences, it is advisable to use a single type of equipment for comparative purposes and thus reducing bias in detectability. In [92], it is recommended to use the same equipment for no longer than five years because older versions usually leave the market. In such a way, researchers can take advantage of the improvements that manufacturers have made.

2.3. Automated recognition of bioacoustic signals

In general, automation techniques are applied in bioacoustics for i) detecting relevant patterns in recordings, which is known as segmentation, ii) extracting features, and iii) classifying patterns. Frequently, segmentation methods assume that relevant patterns are continuous and high energy regions [54]. Among the methods usually employed for extracting features are Linear Predictive Coding (LPC) [85], Mel-frequency cepstral coefficients (MFCC) [69] and descriptive parameters in the frequency and time domain [44]. Methods for acoustic classification of bird species can be grouped into two types [19]: those that classify individ-

[3] `http://www.wildlifeacoustics.com`

[4] `http://arbimon.com/arbimon/index.php/products-acoustics`

ual units such as calls or syllables, and those that classify longer recordings. Most common algorithms for classifying units are nearest-neighbor rule, linear discriminant analysis and support vector machines (SVM) [1]. The classification of recordings has been carried out by using Hidden Markov Models (HMM) [110], Gaussian Mixture Models (GMM) [104] and Multi-Instance Learning (MIL) [19]. Figure **2-1** shows a diagram of a typical automated recognition system. Even though we emphasize on audio signals, this scheme easily fixes to a recognition system based on images, or other quantitative data, e.g., location, date, or weather.

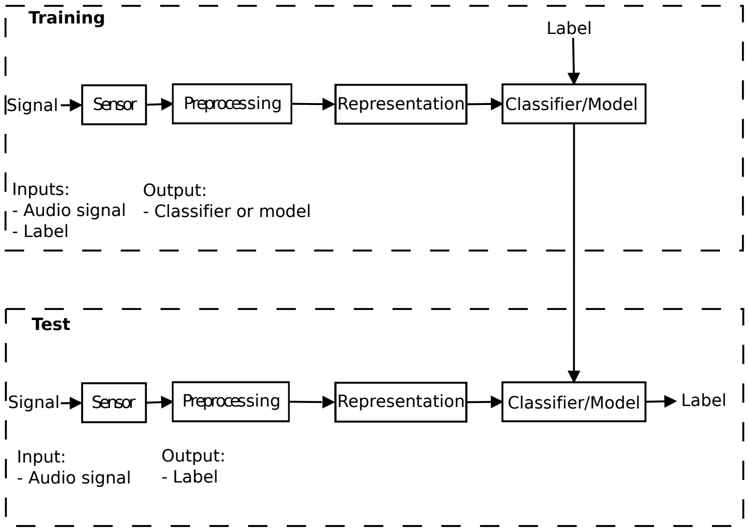


Figure 2-1.Stages of a typical automated recognition system. It is divided into (i) *training* at top, which consists in building a model according to *a priori* information, and (ii) *test* at bottom, which consists in labeling recordings of a test set.

Many approaches of DSP and PR have been applied to the problem of automatic bird detection and classification. These approaches include time-frequency feature extraction, analysis of specific vocalization properties, computation of dissimilarities between acoustic signals or their representations, and statistical classifiers. Earlier studies focused on the classification of syllables and songs. In [54], the problem of classifying syllables of passerine birds is studied by using sinusoidal modelling and classification by matching, i.e., applying the one-nearest neighbor (1-NN) classification rule. A similar approach is proposed in [23] to tackle the bird strike avoidance problem in aviation. In [104], three feature sets are compared: sinusoidal modelling, MFCC and descriptive features; the authors used three classification techniques: 1-NN based on the dynamic time warping (DTW) distance, GMM and HMM. In [44], 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. In [110], HMM is applied for classifying songs of antbirds from a Mexican rainforest represented by MFCC and LPC. Likewise, in [1], a methodology is proposed for automatically classifying isolated calls of three common mountain bird species by using standard call variables and spectral features, as well as 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 in order to face realistic problems. In [17], recordings of 6 species from the Cornell Macaulay Library are classified by using a frame-level feature histogram representation and a the 1-NN rule on statistical manifolds. In [19], a multi-instance multi-label classification framework is proposed for classifying bird song recordings of the H. J. Andrews dataset, which consists in the representation of each audio signal as a bag-of-instances and its classification using a SVM. In [89], the recordings of the Multi-label bird species classification challenge-NIPS 2013 are classified by detecting bags of relevant segments from spectrograms and using image-based features with a random forest classifier. In [105], the concept of unsupervised feature learning is introduced and the recordings of birds, from France, UK and Brazil are classified in four datasets. Among the detection studies, we highlight the following ones: in [9], a methodology of similarity search in audio recordings is proposed by using time-frequency trajectories; this approach is evaluated with recordings from the Animal Sound Archive of Berlin. In [10], a similar approach for detecting vocalizations of the Eurasian bittern and Savi's warbler is applied. Time-frequency features and HMM for detecting bird species of North America, Eurasia, and North Africa are used in [90]. Vocalizations of the *Vanellus chilensis lampronotus* are detected in [49] by 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 described in [100].

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, for some of the studies mentioned above, the audio recognition task is cast into an image processing and classification problem; for example in the following ones: in [62], the proposed recognition method relies on the spectral shape to detect tonal bird sounds in noisy environments. In [3], as a first stage in the recognition system, regions of spectrograms are automatically selected. Similarly, the detection system proposed in [112] extracts image-based features to classify bird species from Brazil.

2.3.1. Estimating the performance of an automated recognition system

Estimating the performance of an automated recognition system is crucial for knowing its reliability. One of the easiest ways to carry out this procedure is by counting the number of objects correctly labeled, which is known as accuracy. However, it might not be the best option for unbalanced data, i.e., when the number of objects per class is significantly different, since a system that assigns only the label of the majority class could show an unreal positive performance. Alternatively, in [15], several measures are described for estimating the performance of two-class classifiers.[5] To apply those measures to multi-class problems, the classification problem can be restated as several two-class problems. This is carried out by taking each class as the target class and the others as non-target class (which is known as one-against-all approach) or evaluating the performance of each pair of classes (which is known as one-against-one approach) [5, 44, 68].

In bioacoustics, classification of recordings is usually a multi-label problem —where an object can belong to several classes. For instance, each recording can contain the sound of different sources, such as vocalizations of several species and environmental noise. There are measures specialized for estimating the performance in multi-label problems: Hamming loss, rank loss, one-error, coverage and micro(macro)-AUC [19].

2.4. Projects of environmental monitoring based on bioacoustic signals

A common purpose of the automated recognition systems of bioacoustic signals is to facilitate the permanent extraction of environmental information [64, 78, 119]. This knowledge is fundamental for communicating with the decision makers and the general public about the state of the environment. The following projects of environmental monitoring from bioacoustic signals are currently being developed. In general, they are equipped with technological tools, specialized hardware and software, communication devices, and expert staff.

- In the United States of America (USA), the Bioacoustic Research Program[6] from The Cornell Lab of Ornithology aims to promote innovative technologies for collecting and interpreting sounds in nature. In this program, hardware and software are being developed to record and analyze acoustical information around the globe.

- In Puerto Rico, the Automated Remote BIodiversity MONitoring Network (ARBIMON[7]) develops several tools for acoustic monitoring. Researchers and technical staff of this project are specialized in environmental monitoring through audio recordings and satellite imagery. They implement permanent recording stations for long-term monitoring in real time. Furthermore, they rent and sell portable recorders, and offer

[5] A two-class classifier assigns to each object one of two possible labels. It is common to denote one of them as positive (or target class) and the other as negative (or non-target class).

[6] `http://www.birds.cornell.edu/brp/`

[7] `http://arbimon.com/`

access to a cloud-computing platform.

- In Europe, the Automatic acoustic Monitoring and Inventorying of BIOdiversity (AMIBIO[8]) project constructs autonomous multi-sensor monitoring stations and software to analyze the acquired data. The goals of this project are: i) biodiversity assessment and inventorying of an area; ii) estimation of the density of animals in the monitored areas; iii) monitoring and alarming about the presence or absence of rare and threatened species at inaccessible areas as well as night-migrating birds; iv) estimation of the health of certain species from their vocalizations; v) Monitoring and alarming of specific atypical sound events such as those related to potentially hazardous human activities (e.g., gun shots and trees falling); and, vi) permanent monitoring for danger and crisis events (e.g., fires and storms).

- In the Remote Environmental Assessment Laboratory (REAL[9]) from Michigan State University, an architecture for automatically collecting acoustic signals in natural areas has been developed. This architecture consists of communication and data processing infrastructure used to transmit, store, and analyze environmental data.

2.5. Software for bioacoustics

The software programs listed below contain tools for processing audio signals. These programs are widely used to analyze and recognize bioacoustic signals by researchers in this area:

- Avisoft (`http://www.avisoft.com/`): there are two types of this software: Avisoft-SASLab Pro that provides sound analysis, editing and classification tools; and Avisoft-RECORDER that is used to trigger recording systems.

- PAMGUARD (`http://www.pamguard.org/`): this open source software is designed particularly to facilitate the passive monitoring in marine environments. It contains tools for acoustic detection, localization and classification.

- Raven (`http://www.birds.cornell.edu/brp/raven/RavenOverview.html`): this is a free software specialized in the analysis, visualization, and measurement of animal sounds.

[8] `http://www.amibio-project.eu/`

[9] `http://www.real.msu.edu/`

- Song Scope (http://www.wildlifeacoustics.com/products/song-scope-overview): this is a computational tool designed to review recordings made by conventional bioacoustic recording equipment.

- SoundID (http://www.soundid.net/): this is a sound recognition application. The primary intention of its developers was to offer a system that could be used to search rare parrots.

- XBAT (http://www.birds.cornell.edu/brp/software): this is a bioacoustics toolbox developed in the numerical computing environment MATLAB.

- SonoBat (http://www.sonobat.com/): this is a software to display and analyze spectrograms of bat echolocation calls recorded from time-expansion bat detectors.

2.6. Challenges and opportunities in Colombia

Colombia is a country rich in biodiversity, especially in acoustically active wildlife, with more than 1.860 bird species [14], 763 frog species, 479 mammal species, 2.000 marine fish species, and 1.435 freshwater fish species.[10] Particularly, birds have been used as an indicator of changes in the environment, since they are widely distributed, easy to detect through their vocalizations and, compared with other groups of animals, there is a good knowledge of their biology. In Colombia, there are almost 20 ornithological associations that are interested in bioacoustic monitoring. Globally, the contribution of ornithological associations and birdwatchers has increased awareness of population trends of birds, especially in Europe and North America. In Colombia, since 1987, an annual census has been done on birds. These manual yearly surveys —since the sampling method has been based on counting points— have contributed to increased knowledge and cohesion of ornithological associations. However, if ornithological associations implement automated acoustic monitoring under a systematic experimental design, they could operate more than just once a year. Thus, it would be possible to collect data on wider temporal and spatial scales.

Nowadays, the interest of different organizations —such as foundations, non-governmental organizations (NGOs), industrial companies, and academic institutions— in automated bioacoustic monitoring is increasingly growing. The IAVH (Instituto de Investigación de Recursos Biológicos Alexander von Humboldt) can support them through its centers and services, particularly the following two: i) the Collection of Sounds, and ii) the Laboratory of Applied Biogeography and Bioacoustics. Thus, these institutions together could set up monitoring

[10] https://www.siac.gov.co/contenido/contenido.aspx?conID=1252&catID=52

nodes under a standardized system of field data collection and species identification. Therefore, it might be possible to guarantee the long-term protection of bioacoustic specimens, develop tools for automating the recognition of vocally active wildlife species, implement an online information system about this type of monitoring, and disseminating this information.

Automated bioacoustic monitoring is an attractive tool in different scenarios. For example, for ornithological tourism in Colombia, it is possible to facilitate the recognition of the species in the field by smartphone applications. Furthermore, this type of monitoring allows private companies to have reliable and efficient systems to detect the presence of wildlife in their areas of operation. Therefore, this technology is ideal for impact studies and environmental management plans that the ANLA (Autoridad Nacional de Licencias Ambientales) requires.

With respect to the design and implementation of an automated recognition system, there are three bottlenecks: a) collecting a statistically sufficient number of examples of vocalizations of target and non-target species; b) segmentation of the recordings to remove and discard audio vocalizations that are not of interest; c) transformation of a prototype system to a final product. This refers to the development of user interfaces, manuals, licensing and legalization and in general everything that should be added to the system to operate it. Tasks a) and b) require more than just technical expertise in computing and electronics. They also need the contribution of experts in ecology. Thus, these tasks can be carried out through cooperation agreements between groups or institutions with trained staff in science and technology. Finally, the transformation of a prototype automated recognition system into one appropriate for real-world scenarios and its installation are activities that require investment in personnel, purchasing equipment and software.

2.7. Discussion

The automated bioacoustic monitoring has the potential to simultaneously supply real-time information from various taxonomic groups —in a systematic way— covering both large spatial and temporal scales. This type of monitoring can provide accurate information about the behavior of the dynamics in nature. For instance, mining, agriculture and infrastructure; as well as factors derived from human development, such as climate change. Using this technology enables early warning systems with high efficiency, which provides environmental decision makers with information based on up-to-date data.

In the world, some research on biodiversity and monitoring projects have been successfully developed worldwide automated bioacoustic tools. In Colombia, cooperation between institutions would facilitate the implementation and use of this technology. A standard bioacoustic scheme of automated monitoring could unify isolated research efforts related to this topic. Since Colombia is a very diverse country —with a large number of species, and endemic and endangered ecosystems— efforts to implement this type of monitoring are valuable and strategic.

Part I.

Multiple Instance Learning

3. Birdsong classification based on multiple instance learning

Abstract

This chapter focuses on the extraction of local information as units –called instances– from audio recordings. The methodology for instance extraction consists in the segmentation carried out on spectrograms using image processing techniques and the estimation of a needed threshold by the Otsu's method. The multiple instance classification (MIC) approach is used for the recognition of the sound units. A public dataset was used for the experiments. The proposed unsupervised segmentation method has a practical advantage over the compared supervised method, which requires the training from manually segmented spectrograms. Results show that there is no significant difference between the proposed method and its baseline. Therefore, it is shown that the proposed approach is feasible to design an automatic recognition system of recordings which only requires, as training information, labeled examples of audio recordings.

3.1. Introduction[1]

In the field of pattern recognition, classifiers are designed to discriminate concepts, e.g., images, audio signals or documents. Traditionally, classifiers are provided with labeled training objects that correspond to feature vector representations —also called instances— of the concepts [55]. However, single feature vectors might be not appropriate to describe complex objects [6]. Alternatively, more advanced representations have been proposed, for example, the multi-instance representation. In the classification in this context, known as

[1] This chapter was published as: José Francisco Ruiz-Muñoz, Mauricio Orozco-Alzate, and Germán Castellanos-Domínguez, Multiple instance learning-based birdsong classification using unsupervised recording segmentation, in Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015, 2015, pp. 2632–2638.

multiple instance classification (MIC), *a priori* information of the concept is not explicitly provided because classifiers are trained from labeled objects represented by sets of feature vectors —called bag-of-instances— where each single instance can or can not be associated with the concept that corresponds to the label of its bag.

Methods for acoustic classification of bird species can be categorized into two types [19]: 1) those that classify individual syllables, and 2) those that classify recordings having sounds of multiple sources. The former ones require a detailed annotation of each segment while in the latter ones, for training, it is only required to label the presence of the species of interest in each recording. In either case, segmentation is a high-priority step [80]. However, methods of the first type are more sensitive to the segmentation quality because omitted syllables become false negatives and those sounds incorrectly detected could become false positives. In contrast, for methods of the second type, it is possible to achieve correct classification even if the two above-mentioned miss-segmentation cases occur since every recording may contain multiple syllables.

In general, segmenting recordings into smaller recognition units is assumed as a part of the preprocessing stage and it is done either manually or automatically. Yet, automatic recognition should not require manual segmentation [110]. For this purpose, segmentation algorithms have been developed mostly using energy and entropy as criteria to identify onset and offset times of the regions of interest [44]. Under ideal conditions, when the vocalization call is the only sound in the recording, an increase in energy clearly reveals a region of interest, making segmentation procedures simple enough [80]. However, in real conditions, recorded signals are degraded due to the presence of many sound sources, e.g., wind streams, background noise from other animals and surrounding events. In spite of that, several research studies on automated species recognition clarify that their methods work well when the recognition units are correctly detected, often this issue is not discussed in depth (making only a brief description). Furthermore, as indicated in [53], it should be taken into account that achieving perfectly segmented data is at least as difficult as the classification step.

Considering that the recognition of recordings has the advantage of not making the impractical assumption of requiring perfectly segmented data, in this chapter, it is proposed a classification methodology for audio recordings which only uses —in the training phase— labels from training recordings, that is, isolated and labeled vocalization segments are not required beforehand since the methodology includes a novel unsupervised segmentation method for birdsong recordings. Interest sounds are detected from the Short-time Fourier transform (STFT). In the segmentation method described in Sec. 3.2.1, the output is a matrix of the same size of the corresponding spectrogram, where interest elements (or pixels) are marked with “one” and non-interest elements with “zero”. Classification is carried out using the

MIC approach, as follows: 1) neighboring interest pixels are grouped into regions, 2) each region is described by a feature representation and 3) a classifier based on multiple instance learning (MIL) is trained considering each spectrogram as a bag of instances. Its classification performance is estimated when using the unsupervised segmentation method proposed in Sec. 3.2.1 and compared against the performance obtained when using the supervised segmentation method proposed in [19]. Both methods consist in the detection of regions in the spectrogram likely associated with vocalizations; however, in the latter, it is required to provide a set of manually annotated spectrograms where pixels have been labeled according to whether or not they correspond to bird sounds.

3.2. Material and methods

3.2.1. Segmentation of birdsong recordings

Time-frequency analysis of audio recordings is usually carried out through spectrograms representing power intensity at each time-frequency point. Particularly, spectrograms are considered as recognizable images to identify bird species [35], whose vocalizations are represented by intensity variations. Thus, under the assumption that segment vocalizations give a form of continuous regions holding the highest power values, our unsupervised segmentation methods consists in the following stages (see Fig. **3-1**):

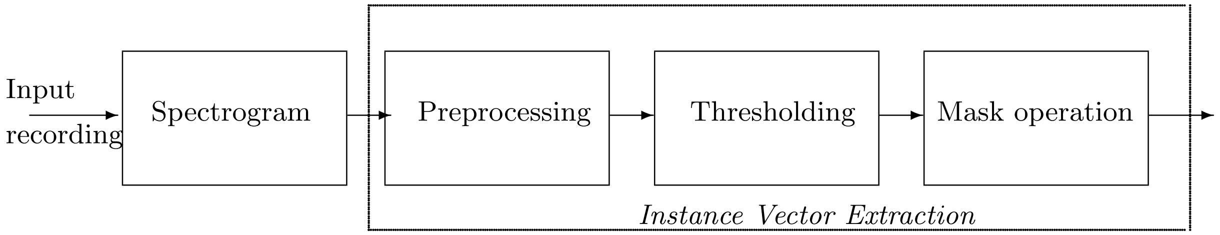


Figure 3-1.Flow diagram of the proposed segmentation method.

- **Spectrogram estimation:** Based on the STFT decomposition, we compute a spectrogram matrix $\boldsymbol{Y} = [y_{ij} : i = 1, \ldots, F; j = 1 \ldots, T] \in \mathbb{R}^{F\times T}$, where indexes i and j stand for the frequency and time domains (i.e., F points in the frequency domain that are estimated in each one of T time frames). We use the Hann window lasting 512 samples and overlapping 256 samples as in [19].

- **Preprocessing:** After applying the two-dimensional Wiener filter, a denoised and smoothed spectrogram, $\widetilde{\boldsymbol{Y}} = [\tilde{y}_{ij}] \in \mathbb{R}^{F\times T}$ is estimated with elements:

$$\tilde{y}_{ij} = \mu + y_{ij}(\sigma^2 + \sigma_\eta^2)/\sigma^2$$

where $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}$ are the local mean and variance, respectively, and $\sigma_\eta^2 \in \mathbb{R}$ is the noise variance. The first two values are estimated at a $k \times k$ neighborhood centered on each point (we set $k=5$ as in [91]), and σ_η^2 is the average over all local variances σ^2. Then, we suppress structures that are lighter than their surroundings and are connected to image borders by considering a value of 8 as connectivity parameter. An erosion [13] is performed before by estimating the lowest number of image elements that are not connected to the edges. The erosion procedure results in a smoothed image $\boldsymbol{I} \in \mathbb{R}^{F \times T}$, from which we perform morphological reconstruction of $\widetilde{\boldsymbol{Y}}$ to remove all intensity fluctuations (except the intensity peak). As a result, we get the matrix $\boldsymbol{H} \in \mathbb{R}^{F \times T}$ that only holds objects with neighboring borders. At the same time, the difference matrix, $\widetilde{\boldsymbol{H}} = \widetilde{\boldsymbol{Y}} - \boldsymbol{H}$ is also computed holding only those objects from the original image not having neighboring borders.

- **Thresholding:** We fix a threshold to binarize each image $\widetilde{\boldsymbol{H}}$ using the nonparametric and unsupervised Otsu's method of automatic threshold selection [84]. Extracted only from a computed gray-level histogram, the optimal threshold T_o is selected by maximizing an introduced discriminant measure of separability among all resultant gray level classes, as follows:

$$\hat{k} = \arg \max_{k \in [1,L]} \{(\phi_T \boldsymbol{\omega}(k) - \boldsymbol{\phi}(k))^2 / (\boldsymbol{\omega}(k)(1 - \boldsymbol{\omega}(k)))\}$$

where $\phi_T = \sum_{j=1}^{L} j p_j$, $\boldsymbol{\omega}(k) = \sum_{j=1}^{k} p_j$, $\boldsymbol{\phi}(k) = \sum_{j=1}^{k} j p_j$, $p_j = n_j / N$ are the values of the normalized gray level histogram, L is the number of gray levels, n_j is the number of pixels at level j, and N is the total number of pixels over the whole difference image $\widetilde{\boldsymbol{H}}$. Therefore, the optimal threshold is computed as $T_o = (\hat{k} - 1)/(L - 1)$.

- **Mask operation:** To select the most relevant pixels from the spectrogram at hand, a binary matrix $\boldsymbol{B} = [b_{ij}] \in \mathbb{R}^{F \times T}$ is obtained by thresholding as follows:

$$b_{ij} = \begin{cases} 1, & \text{if } \tilde{h}_{ij} > T_o; \\ 0, & \text{otherwise.} \end{cases}$$

- **Instance vector extraction:** From computed arrangements $\boldsymbol{B}$ and $\boldsymbol{Y}$, we compute a spectrogram region set $\mathcal{Z} = \{\boldsymbol{Z}_i : i = 1, \ldots, l\}$, holding the respective matrices of all-connecting points $\boldsymbol{Z}_i \in \mathbb{R}^{F_i \times T_i}$, with $\boldsymbol{Z}_i \subset \boldsymbol{Y}$, being $F_i < F$ and $T_i < T$. The number of regions l is automatically fixed as indicated in [122]. Lastly, each region $\boldsymbol{Z}_i$ supplies a single feature vector (or instance), denoted as $\boldsymbol{x}_i \in \mathbb{R}^d$. The number of features d is fixed in accordance with the training scenarios explained in Sec. 3.3.1.

3.2.2. Multiple instance classification

Within the standard supervised classification framework, the training set consists of n feature vector examples or instances $\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^d : i = 1, \ldots, n\}$ and their labels, in a two-class problem, $\mathcal{Y} = \{y_i \in \{0, 1\} : i = 1, \ldots, n\}$. Thus, any classifier function, $\mathcal{X} \rightarrow \mathcal{Y}$, is trained to predict labels for each novel instance. On the other hand, in MIC, an object is represented by a set, or bag, $\boldsymbol{X}_i = \{\boldsymbol{x}_{ij} \in \mathbb{R}^d : j = 1, ..., m_i\}$ of m_i instances $\boldsymbol{x}_{ij}$ and a label $\tilde{y}_i$, i.e., each label is associated with the entire bag but labels of the individual instances are unknown. Therefore, the training set consists of n bags $\widetilde{\mathcal{X}} = \{\boldsymbol{X}_i \in \mathbb{R}^{m_i \times d} : i = 1, ..., n\}$ and their corresponding labels $\widetilde{\mathcal{Y}} = \{\tilde{y}_i \in \{0, 1\} : i = 1, \ldots, n\}$. Then, the classifier function, $\widetilde{\mathcal{X}} \rightarrow \widetilde{\mathcal{Y}}$, is trained to predict labels for each novel bag of instances.

MIC methods, depending on the level where they hold discriminant information, can be grouped into two broad categories [4]: 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 bag to a single feature vector. As opposed to methods based on dissimilarities between bags, a set of relevant features is derived.

We address the problem of classifying recordings using the MIC approach because, in our case, we have labels for recordings, i.e. bags, but they are represented for several instances (feature vectors extracted from each region detected after the segmentation stage).

Particularly in this chapter, we use the MILES (MIL via Embedded Instance Selection) classification algorithm because it has been experimentally shown that it performs well with bioacustic signals [25]. MILES transforms the original MIC problem into a standard supervised learning framework injectively relating instances and labels [22]. It maps each bag into a feature space defined by instances in the training bags using an introduced instance (dis)similarity measure. Thus, bags are represented by the maximum (dis)similarity to all other instances. On this (dis)similarity representation, a sparse linear classifier is trained [108].

3.2.3. Segmentation performance measures

Since the manual recording segmentation is a very fatiguing task, rather than directly comparing between automated and manual outputs, we indirectly estimate the quality of the segmentation method proposed in Sec. 3.2.1 by the recording classification performance that must be strongly influenced by the used segmentation procedure, as discussed in [19]. The most common performance measures for a classifier are the following ones: accuracy $a = (T_P + T_N)/(P + N)$, specificity $s = T_N/N$, recall-rate $r = T_P/P$ and precision-rate $p = T_P/(T_P + F_P)$; where T_P is the number of recordings correctly classified as positives, T_N is the number of recordings correctly classified as negatives, F_P is the number of recordings incorrectly classified as positives, F_N is the number of recordings incorrectly classified as negatives, P and N are the total number of positives or negatives recordings, respectively. However, these performance measures are affected by the relative size of the classes Therefore, to overcome that drawback, the two-class F-score is used as the performance measure defined as follows:

$$F = \frac{2T_P}{2T_P + F_P + F_N}, \tag{3-1}$$

so that F ranges from 0 to 1 where the higher its value, the better the classification performance. In this study, the one-against-all reduction from multi-class task to binary classification technique is used where each species is selected as objective (positive) class since the F-score is a two-class measure. As regards the classifier performance, we carry out validation using the Bootstrapping technique where input audio data are randomly split into two sets: one-half for training and one-half for testing. This procedure is carried out ten times.

As suggested in [15], each one of the eight training strategies explained in Sec. 3.3.1 are compared over the 13 classes (see Sec. 3.3.2) in terms of the paired t-test, for which the null hypothesis states that the performance of two classification strategies can be statistically assumed as the same. The pseudo-code to determine whether to accept the null hypothesis is presented in Algorithm 1. Otherwise, when the null hypothesis is rejected, we select as the best strategy the one having the highest pair performance difference computed in average over all considered classification strategies.

Algorithm 1 Process to determine whether or not to accept the null hypothesis of the paired t-test. Based on [15].

```
Let $\boldsymbol{z} = [z_1,\ z_2,\ ...\ z_n]$ be the vector of the differences in classification performances between strategies 1 and 2 and let $n$ be the number of classes.
Assign $t_{level}$ considering the following: for 13 degrees of freedom, $t \geqslant 1.771$ would only be expected to occur by chance with probability 0.10 or less, and it is said that the hypothesis is rejected at 10% level. Therefore, if $t_{level} = 1.771$, $t_{level} = 2.160$ or $t_{level} = 3.012$, and $|t| > t_{level}$, the null hypothesis is rejected at the 10%, 5% or 1% level, respectively.
procedure T-TEST($\boldsymbol{z}$, $n$, $t_{level}$)
    Compute $a = (\sum_{i=1}^{i=n} z_i^2) - (\sum_{i=1}^{i=n} z_i)^2/n$.
    The sample variance $s^2$ is equal to $a/(n-1)$.
    The sample standard deviation is the square root of $s^2$.
    Divide $s$ by the square root of $n$ to get the standard error $e_{std}$.
    The $t$ statistics is computed dividing the average value of $z$ by $e_{std}$.
    if $|t| \geqslant t_{level}$ then
        Return The hypothesis is rejected.
    else
        Return The hypothesis is assumed as true.
    end if
end procedure
```

3.3. Experiments

3.3.1. Training strategy

Figure **3-2** shows the training scheme used through all experiments to test the proposed birdsong classification methodology based on unsupervised segmentation of audio recordings and multiple instance learning.

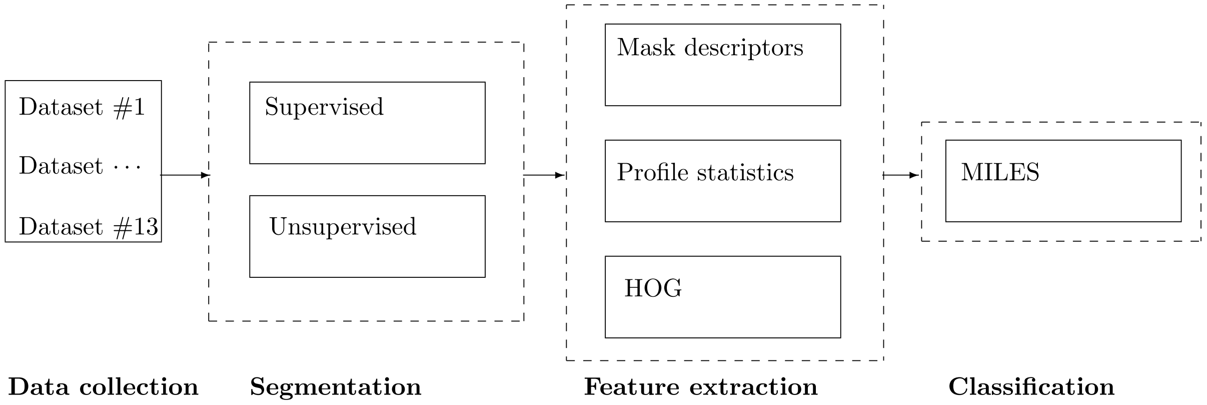


Figure 3-2.Training scheme used through all experiments to test the proposed birdsong classification methodology.

In the segmentation stage, only one approach, either the unsupervised (see Sec. 3.2.1) or the supervised (proposed in [19]), is used. Besides, the following four representation scenarios are separately considered: 1) **Mask descriptors**, denoted as "MD", that describe region shape; the following set of features are computed: minimum frequency, maximum frequency, bandwidth, duration, area, perimeter, non-compactness, and rectangularity. 2) **Profile statistics**, "PS", a set of fourteen features are computed, which are based on statistical segment properties in time and frequency: frequency-Gini, time-Gini, frequency-mean, frequency-variance, frequency-skewness, frequency-kurtosis, time-mean, time-variance, time-skewness, time-kurtosis, frequency-max, time-max, mask-mean, and mask-standard deviation. 3) **Histogram of Gradients**, "HOG", this set consists of 16 features characterizing shape and texture of each region where gradient directions over the pixels of the region are computed; each histogram holds 16 bins equally spaced over the angle range $[-\pi, \pi]$ and features are extracted from the normalized 1-D histograms [33]. 4) **All-features set**, "AF", that merges all above feature sets into a single one, as in [19].

Therefore, eight training strategies are tested for the feature extraction stage. In case of supervised segmentation training, the affix "-Br" (meaning Briggs) is added to the end of every feature notation, in accordance to the segmentation method proposed in [19]. Lastly, we use the MILES classification algorithm. We employ an exponential kernel $\exp\{-(||a -$

$b||)/p\}$ as the instance (dis)similarity function, where parameter $p \in [2 \dots 5]$ is heuristically fixed and notation $\| \cdot \|$ stands for Euclidean-norm.

3.3.2. Dataset of recordings

For the sake of comparison, we perform experiments with the publicly available[2] dataset used in [19]. This data collection holds 548 recordings sampled at 16 *kHz* that were manually labeled. The dataset contains 13 bird species, often vocalizing simultaneously and perturbed with environmental noise, though each recording lasting ten-seconds holds between one and five species. Table **3-1** shows the amount of recordings holding each considered species.

Table 3-1.Amount of recordings where each considered bird species (objective class) is labeled.

	Classes	*Labeled species name*	*Number of recordings*
1	BRCR	Brown Creeper	197
2	WIWR	Winter Wren	109
3	PSFL	Pacific-slope Flycatcher	165
4	RBNU	Red-breasted Nuthatch	82
5	DEJU	Dark-eyed Junco	20
6	OSFL	Olive-sided Flycatcher	90
7	HETH	Hermit Thrush	15
8	CBCH	Chestnut-backed Chickadee	117
9	VATH	Varied Thrush	89
10	HEWA	Hermit Warbler	63
11	SWTH	Swainson's Thrush	79
12	HAFL	Hammond's Flycatcher	103
13	WETA	Western Tanager	46

3.3.3. Results of compared classification strategies

Table **3-2** shows the estimated F-scores for the different considered feature sets; notice that the HAFL class has the highest F-score (> 0.99). In contrast, both classes DEJU and HETH achieve the lowest values (< 0.21) and get zero-value F-score for several features because those classes hold very few recordings (see Table **3-1**): 20 and 15, respectively.

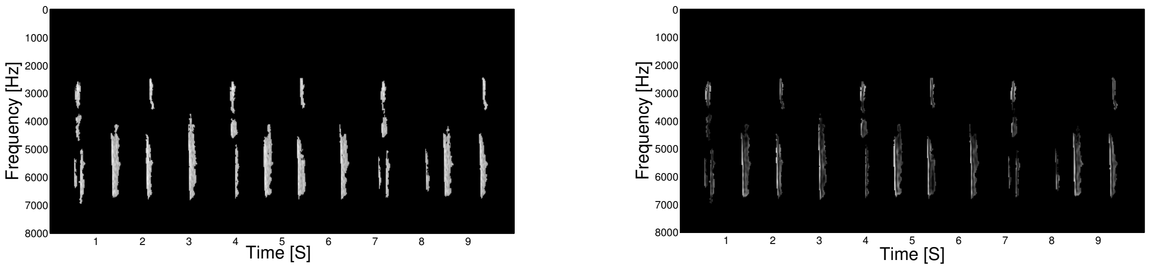
[2] Audio recordings of the dataset are available at http://www.miproblems.org/datasets/birds/

Table 3-2. Performed F-score values for all considered objective classes and each training scenario. The best reached F-score is marked in bold for each objective class. Besides, the notation "–" stands for null-value performance.

	Training scenario							
Class	*MD*	*MD-Br*	*PS*	*PS-Br*	*HOG*	*HOG-Br*	*AF*	*AF-Br*
BRCR	0.762	0.792	**0.848**	0.824	0.694	0.847	0.774	0.819
WIWR	**0.938**	0.870	0.917	0.870	0.896	0.900	0.933	0.905
PSFL	0.791	0.771	0.781	0.777	0.768	0.736	**0.817**	0.802
RBNU	0.853	0.742	0.793	0.725	0.759	0.784	**0.871**	0.831
DEJU	–	–	–	–	–	–	0.095	**0.214**
OSFL	0.701	0.704	0.718	0.681	0.667	0.694	**0.756**	0.738
HETH	–	–	–	–	–	–	–	–
CBCH	**0.742**	0.569	0.687	0.602	0.643	0.514	0.685	0.600
VATH	0.967	0.931	0.902	0.971	0.909	0.889	0.911	**0.983**
HEWA	0.729	0.718	0.643	**0.764**	0.641	0.652	0.715	0.741
SWTH	0.521	0.587	0.594	**0.813**	0.451	0.566	0.627	0.796
HAFL	**1**	0.996	0.999	0.996	**1**	0.994	**1**	0.996
WETA	**0.852**	0.762	0.795	0.757	0.340	0.372	0.840	0.823
Average	0.805	0.767	0.789	0.798	0.706	0.723	0.812	0.821

In terms of the considered feature sets, the use of AF-Br reaches the highest average performance value ($F = 0.821$) that is averaged over all considered objective classes, while HOG gets the lowest one, $F = 0.723$, for the baseline supervised approach. Once again, AF ($F = 0.812$) and HOG ($F = 0.706$) are the best and worst cases, respectively, for the unsupervised method. However, the MD feature set, in average, benefits the most from the use of the unsupervised segmentation, while the HOG set degrades the worst. It must be noted that the average performance is taken without DEJU and HETH classes because of their accomplished anomaly values. This situation may be explained since the MD feature set encodes region shape attributes that are far from being easy to be manually computed, as seen in Fig. **3-3 (a)** showing a typical birdsong spectrogram. In contrast, the texture-based HOG set implies calculation of gradient directions over region pixels, as illustrated in Fig. **3-3 (b)**; this procedure includes enhanced Gaussian filtering and histogram binning that are very sensitive to their parameter tuning.

In order to provide an intuitive illustration, all F-score values estimated in Table **3-2** are graphically presented in Fig. **3-4** where the vertical axis represents the performance obtained by the proposed unsupervised-based segmentation method, while the horizontal axis corresponds to the one obtained by the supervised reference method. Since the larger the F-score – the better the performance, any point above the diagonal line indicates that the



(a) Segmented spectrogram **(b)** Gradient directions

Figure 3-3.Example of a segmented spectrogram from a given audio birdsong recording.

proposed method outperforms the reference. In most of the cases, no remarkable difference in terms of performance is observed between both segmentation methods tested for the same representation scenario.

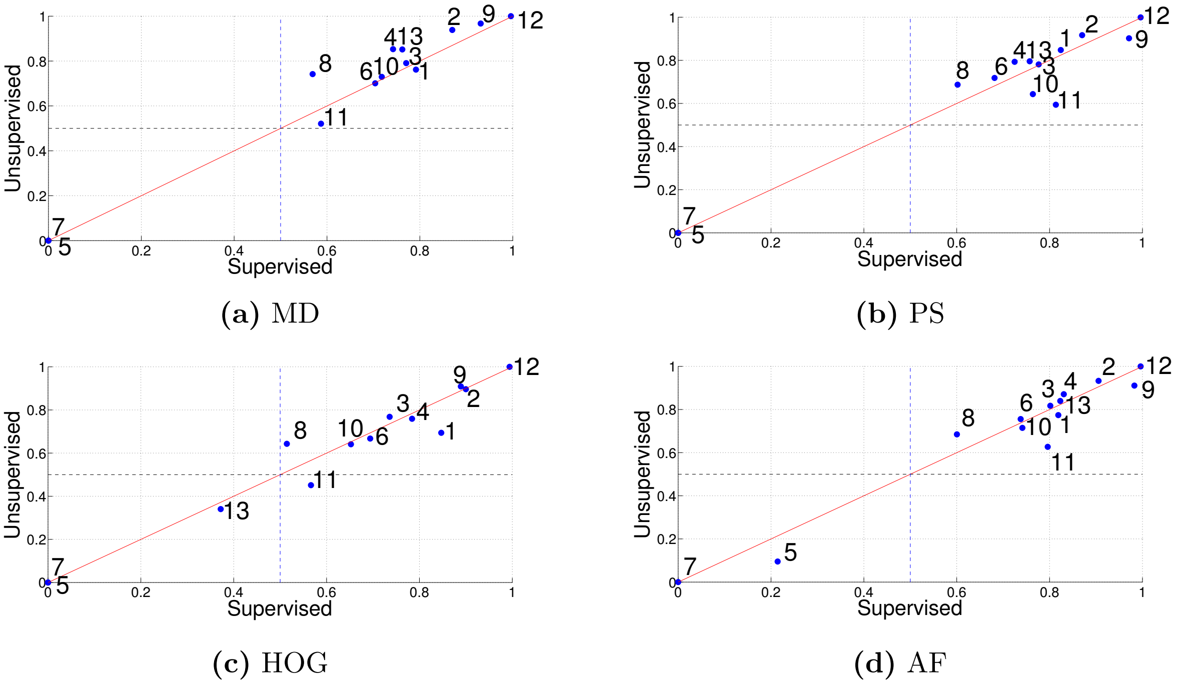


Figure 3-4.Relationship of performed F scores for each considered feature set between both methods: the proposed unsupervised-based segmentation (vertical axis) and supervised reference (horizontal axis).

In case of the worst performance when the classifier guesses at random with equal frequency (i.e., $T_p = P/2$, $F_n = P/2$, $F_N = N/2$, and $T_n = N/2$). Thus, one may calculate a threshold $F_t = P/(1.5P+0.5N)$ value, under which the classifier is just guessing. In our case, $F_t = 0.42$, corresponding to BRCR class, is the lowest one. As shown in Fig. **3-4**, most of the computed

F-scores overcome by far this F_t threshold.

Table **3-3** shows several performance measures (namely, recall, r, specificity, s, precision, p, accuracy, a, and the F-score) obtained for the whole feature sets, i.e., AF and AF-Br feature sets. Typically, both training strategies may be differently influenced by each considered class. Particularly, the former training strategy gets the better specificity for `DEJU`, `HETH`, `VATH`, `HAFL`, and `WETA` classes, while the latter strategy instead of `WETA` the `HETH` class also gets the highest value. At the same time, only the `HALF` class remains the best in terms of accuracy.

To provide a better illustration, Table **3-4** shows the obtained results of the paired F-score difference, ΔF, computed between the AF and AF-Br features. Since the latter set is the reference, the estimated ΔF value gets negative sign when the reference feature set is better, otherwise ΔF becomes a positive value. The best and worst achieved ΔF values are marked in bold. Particularly, the `SWTH` gets the lowest difference (-0.169), that is, that class is the most negatively influenced by the proposed strategy while the `CBCH` class achieves the best influence (0.084). However, though the average value is $\Delta F = -0.017$ in support of AF-Br set, the corresponding estimated t value gets as high as 0.905, meaning that neither of the considered feature sets are statistically different at levels of 10%, 5%, and 1%.

Lastly, the null-hypothesis values are shown in Table **3-5** in order to make clear the influence of each considered segmentation strategy, in terms of performed F-score classification measure. Values are computed at 10% level (a strong assumption) to either admit (value 0) or deny (value ± 1) the null hypothesis about their statistical similarity between each pair of contrasted feature sets (columns stand for the reference supervised feature sets and rows for the proposed unsupervised sets). In case the statistical similarity is rejected, the null hypothesis gets the value 1 if the column-wise feature set reaches a better performance than the row-wise set. Otherwise, the null hypothesis gets the value -1.

As seen from the obtained main diagonal matrix values, one can infer that each compared feature set, except MD, has no statistical difference regardless of the used segmentation approach. In case of the MD set, its unsupervised version turns out to be better. This situation should be expected due to the above-given explanation about the advantage of automated MD feature extraction. As a result, the proposed birdsong classification methodology based on unsupervised segmentation of audio recordings and multiple instance learning has no statistical difference with the baseline supervised version, at least, when using the three extracted feature sets: PS, HOG, and AF.

Table 3-3.Performance classification measures for the feature sets: AF (above) and AF-Br (below). The best column-wise measures achieved overall data-sets are highlighted in bold. Performance is estimated by: , recall (r), specificity (s), precision (p), accuracy (a), and F-score (F).

	Class	r	s	p	a	F
AF	BRCR	0.91	0.76	0.68	0.81	0.77
	WIWR	0.92	0.99	0.94	0.97	0.93
	PSFL	0.87	0.89	0.77	0.88	0.82
	RBNU	0.86	0.98	0.88	0.96	0.87
	DEJU	0.05	**1**	**1**	0.96	0.10
	OSFL	0.86	0.92	0.67	0.91	0.76
	HETH	0	**1**	0/0	0.97	0
	CBCH	0.74	0.89	0.64	0.86	0.68
	VATH	0.84	**1**	**1**	0.97	0.91
	HEWA	0.57	**1**	0.95	0.95	0.72
	SWTH	0.47	0.99	0.93	0.92	0.63
	HAFL	**1**	**1**	**1**	**1**	**1**
	WETA	0.77	0.99	0.92	0.98	0.84
AF-Br	BRCR	0.87	0.85	0.77	0.86	0.82
	WIWR	0.88	0.98	0.94	0.96	0.90
	PSFL	0.78	0.93	0.83	0.88	0.80
	RBNU	0.82	0.97	0.84	0.95	0.83
	DEJU	0.12	**1**	**1**	0.97	0.21
	OSFL	0.85	0.91	0.65	0.90	0.74
	HETH	0	**1**	0/0	0.97	0
	CBCH	0.46	0.98	0.85	0.87	0.60
	VATH	0.97	**1**	**1**	**1**	0.98
	HEWA	0.80	0.95	0.69	0.94	0.74
	SWTH	0.70	0.99	0.91	0.95	0.80
	HAFL	0.99	**1**	**1**	**1**	**1**
	WETA	0.72	**1**	0.97	0.97	0.82

Table 3-4. Computation example of the paired F-score difference, ΔF, that gets negative sign when the reference feature set is better, otherwise ΔF becomes positive.

Dataset	AF	AF-Br	ΔF
BRCR	0.774	0.819	-0.045
WIWR	0.932	0.905	0.027
PSFL	0.817	0.801	0.016
RBNU	0.871	0.831	0.040
DEJU	0.095	0.214	-0.119
OSFL	0.756	0.738	0.018
HETH	0	0	0
CBCH	0.685	0.600	**0.084**
VATH	0.911	0.983	-0.072
HEWA	0.715	0.741	-0.026
SWTH	0.627	0.796	**-0.169**
HAFL	1	0.996	0.004
WETA	0.840	0.823	0.017
Average	0.694	0.711	-0.017

Table 3-5. Values of null-hypothesis test computed at 10% level for both considered training segmentation strategies: supervised and unsupervised. Main diagonal elements marked in bold.

	MD	PS	HOG	AF
MD-Br	**1**	0	0	1
PS-Br	0	**0**	-1	0
HOG-Br	0	0	**0**	1
AF-Br	0	-1	-1	**0**

3.4. Discussion

In this chapter, the use of unsupervised segmentation of audio birdsong recordings is investigated along with multiple instance learning to classify among a given number of bird species. The proposed unsupervised segmentation of audio birdsong recordings is contrasted against its baseline reference supervised version requiring manual annotation of properly computed spectrograms, as described in [19]. Yet, since this manual recording segmentation poses as a very fatiguing task, we indirectly estimate the quality of the proposed segmentation method by the introduced two-class F-score as classification performance measure that is not affected by the relative class size. Afterwards, each one of the considered feature sets are compared in terms of the paired t-test, for which the null hypothesis states that the achieved F-score performance of two given classification sets can be statistically assumed as the same. The univariate paired t-test is preferred due to its simple interpretation though other multivariate tests may be used, for example, the multivariate paired Hotelling's T^2 test that provides similar results in our work. Both segmentation approaches are validated for the four feature sets: MD, PS, HOG, as well as the all-features set. In average, the MD feature set benefits the most from the use of the unsupervised segmentation, while the HOG set degrades the worst, as seen in Table **3-2**. The main reason for the latter results is the fact that parameter tuning of the automated HOG feature extraction should be improved. However, this procedure is out of the scope of the present work. Nonetheless, according to the accomplished values of the null-hypothesis test shown in Table **3-5**, the introduction of the unsupervised segmentation of audio recordings has no statistical difference with the baseline supervised version, at least, when using the following three extracted feature sets: PS, HOG, and AF.

This work provides a birdsong recognition framework using the MILES classifier, which in turn uses an exponential kernel as (dis)similarity measure. Even though performed F-score values are high, this classifier is sensitive to a low number of training recordings. As a conclusion, the proposed unsupervised recording segmentation of audio birdsong recordings improves species classification with the benefit of easier implementation since no manual handling of recordings is required, making feasible the design of fully automatic birdsong recognition systems.

4. Enhancing the dissimilarity-based multi-instance classification of birdsong recordings

Abstract

As stated in the previous chapter, classification of birdsong recordings can be naturally formulated as a multiple instance problem, where a bag of instances is built from one spectrogram, and each instance corresponds to a region of interest in the spectrogram of its bag. Those instances are detected after a segmentation stage of the audio recordings. In this chapter, 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.

4.1. Introduction[1]

In general, automatic recognition systems require an adequate *representation* of the objects or events to be recognized as well as accurate *classification* rules [87]. In the particular case of bioacoustic applications, we group the options to represent the segmented and preprocessed

[1] This chapter was published as: José Francisco Ruiz-Muñoz, Mauricio Orozco-Alzate, and Germán Castellanos-Domínguez. Enhancing the dissimilarity-based classification of birdsong recordings in Ecological Informatics, Elsevier, Vol. 33, 2016, pp. 75–84.

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* [73], 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 for example the one described in Sec. 4.2.1). It is worth clarifying that the segmentation algorithm may fail —in isolated cases, as indicated by [18]— 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 [25], 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 in [65] 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 [107]. (ii) To optimize or adapt[2] a given dissimilarity measure by using the information from a training set [41]. 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 chapter is as follows: Representation and classification methods are

[2] 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.

Table 4-1.Notations used in this paper.

Notation	*Explanation*
$\boldsymbol{S}_n$	n-th bag of instances
$\boldsymbol{s}_{nm}$	m-th instance of $\boldsymbol{S}_n$
$\boldsymbol{E}\{\cdot\}$	Expectation operator
$d(\cdot,\cdot)$	Any dissimilarity measure
$d_{\min}(\boldsymbol{S}_k,\boldsymbol{S}_l)$	Overall minimum distance
$\mathcal{R}$	Representation set ($\mathcal{R} \subseteq \mathcal{T}$)
$\boldsymbol{R}_l$	l-th bag of the representation set
$d_{\overline{\min}}(\boldsymbol{S}_k,\boldsymbol{S}_l)$	Mean minimum distance
$d_{haus}(\boldsymbol{S}_k,\boldsymbol{S}_l)$	Standard Hausdorff distance
$d_{\overline{haus}}(\boldsymbol{S}_k,\boldsymbol{S}_l)$	Modified Hausdorff distance
$\Vert \cdot \Vert$	Euclidean norm
$\Vert \cdot \Vert_p$	ℓ_p-norm
$\boldsymbol{d}_{\boldsymbol{S}_i}$	Vector of dissimilarities between $\boldsymbol{S}_i$ and elements of $\mathcal{R}$
$\mathcal{T}$	Training set of bags of instances
$\boldsymbol{T}_k$	k-th bag of the training set
$\tilde{d}(\cdot,\cdot)$	Adapted measure
$\tilde{d}_{LANN}(\boldsymbol{S}_i,\boldsymbol{R}_l)$	Dissimilarity measure adapted by LANN
$\tilde{d}_{z-s}(\boldsymbol{S}_i,\boldsymbol{R}_l)$	Dissimilarity measure adapted by z-score
$\tilde{d}_{ESL}(\boldsymbol{S}_i,\boldsymbol{R}_l)$	Dissimilarity measure adapted by ESL
$\tilde{d}_{NLSD}(\boldsymbol{S}_i,\boldsymbol{R}_l)$	Dissimilarity measure adapted by NLSD
$\tilde{\boldsymbol{d}}_{\boldsymbol{S}_i}$	Adapted-dissimilarity vector between $\boldsymbol{S}_i$ and elements of $\mathcal{R}$
ρ	radio estimated by LANN

described in Sec. 4.2. The experiments and obtained results are described in Sec. 4.3 and discussed further in Sec. 4.4. Table **4-1** summarizes the notation used in this chapter.

4.2. 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.

4.2.1. Preprocessing stage

For addressing our classification problem, we use the set of bag-of-instances that are extracted from the data set collected and analyzed by [19]; see their paper for further details about preprocessing and noise reduction stage. Initially, the magnitude spectrogram is estimated by dividing every birdsong recording into frames of 32 ms with 50% overlap, using the discrete Fourier transform windowed by the Hamming function. Afterwards, two iterations of a whitening filter are applied to remove noise in order to improve the contrast of bird sounds from the background. Such a filter consists in normalizing each frequency band by the average of the low-energy frames. Each pixel of the preprocessed spectrograms is automatically labeled as either bird sound or noise. To this end, each single pixel is described by a feature vector composed by the following values from itself and its surrounding: i) coordinate value in the frequency axis associated to the pixel, ii) intensities of pixels in a neighborhood of size 13 × 25 (in time and frequency, respectively) centered in the current pixel and iii) standard deviation of the intensities in the neighborhood. Finally, a supervised classifier is trained with a set of manually-segmented spectrograms. Each large enough and continuous area of the spectrogram labeled as bird sound becomes a single instance. One instance is characterized by a vector holding 38 features grouped into three types: mask descriptors, profile statistics, and the histogram of gradients. Refer again to [19] for a detailed description on the computation of these features. Consequently, the whole set of instances extracted from one spectrogram constitutes a bag. Figure **4-1** illustrates the procedure for extracting a bag of instances from a raw spectrogram.

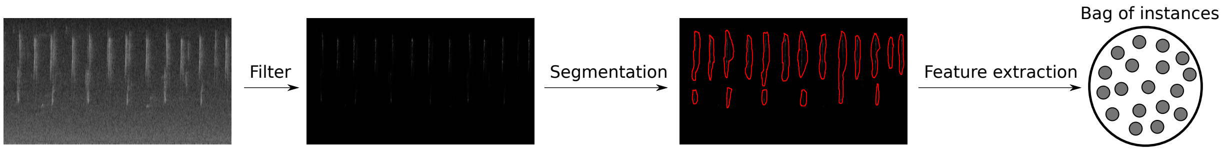


Figure 4-1. Illustration of the procedure for extracting a bag of instances from a raw spectrogram.

4.2.2. Dissimilarity measures between bags

Provided a pair of bags, $\boldsymbol{S}_k = \{\boldsymbol{s}_{km}\}$ and $\boldsymbol{S}_l = \{\boldsymbol{s}_{lq}\}$, each one holding N_k and N_l instances, respectively, we will consider the following pairwise dissimilarity measures between bags:

- **Overall minimum distance**:

$$d_{\min}(\boldsymbol{S}_k, \boldsymbol{S}_l) = \min_{m,q} d(\boldsymbol{s}_{km}, \boldsymbol{s}_{lq}) \tag{4-1}$$

  where $d(\boldsymbol{s}_{km}, \boldsymbol{s}_{lq})$ is the dissimilarity measure between their corresponding instances $\boldsymbol{s}_{km}$ and $\boldsymbol{s}_{lq}$.

- **Mean minimum distance:**

$$d_{\overline{\min}}(\boldsymbol{S}_k, \boldsymbol{S}_l) = \boldsymbol{E}\{\boldsymbol{E}\{\min_{m} d(\boldsymbol{s}_{km}, \boldsymbol{s}_{lq}) : \forall q\}, \boldsymbol{E}\{\min_{q} d(\boldsymbol{s}_{km}, \boldsymbol{s}_{lq}) : \forall m\}\} \tag{4-2}$$

  where $\boldsymbol{E}\{\cdot\}$ stands for expectation operator and it is computed here as an arithmetic average; that is, we assume an equally likely case.

- **Standard Hausdorff distance:**

$$d_{haus}(\boldsymbol{S}_k, \boldsymbol{S}_l) = \max(\max_{q}(\min_{m}\ d(\boldsymbol{s}_{km}, \boldsymbol{s}_{lq})), \max_{m}(\min_{q}\ d(\boldsymbol{s}_{km}, \boldsymbol{s}_{lq}))) \tag{4-3}$$

- **Modified Hausdorff distance:**

$$d_{\overline{haus}}(\boldsymbol{S}_k, \boldsymbol{S}_l) = \max(\boldsymbol{E}\{\min_{m} d(\boldsymbol{s}_{km}, \boldsymbol{s}_{lq}) : \forall q\}, \boldsymbol{E}\{\min_{q} d(\boldsymbol{s}_{km}, \boldsymbol{s}_{lq}) : \forall m\}) \tag{4-4}$$

Due to its easy interpretation, the Euclidean distance is commonly used as the pairwise closeness between instances: $d(\boldsymbol{s}_{km}, \boldsymbol{s}_{lq}) = \|\boldsymbol{s}_{km} - \boldsymbol{s}_{lq}\|$, where notation $\|\cdot\|$ stands for the ℓ_2-norm.

4.2.3. Dissimilarity-based Multiple Instance Classification

In dissimilarity-based MIC, each bag $\boldsymbol{S}_i$ is described by a vector $\boldsymbol{d}_{\boldsymbol{S}_i} = [d(\boldsymbol{S}_i, \boldsymbol{R}_1) \ldots d(\boldsymbol{S}_i, \boldsymbol{R}_P)] \in \mathbb{R}^{1\times P}$ that holds dissimilarities to a representation set composed by P prototype bags $\mathcal{R} = \{\boldsymbol{R}_1, \ldots, \boldsymbol{R}_P\}$, where $d(\cdot,\cdot)$ is a bag-to-bag dissimilarity measure. On the other hand, the training set $\mathcal{T} = \{\boldsymbol{T}_1, \ldots, \boldsymbol{T}_M\}$ contains the points that populate the space in order to define classification boundaries. Note that the representation set can be a subset of the training set, $\mathcal{R} \in \mathcal{T}$, and that the size of the $\mathcal{R}$ set determines the number of entries of the vector representation, which is the dimension of the corresponding dissimilarity space.

Based on the incorporated dissimilarity measure between bags, we further design the classification algorithm so that information from a query test object is given in terms of its dissimilarities to the training set. To this end, we initially employ the baseline k-NN classifier that assigns the most frequently occurring class among the nearest neighbors. Specifically, the 1-NN classifier gives to $\boldsymbol{S}_i$ the label of the prototype that corresponds to the minimum element of $\boldsymbol{d}_{\boldsymbol{S}_i}$. As an alternative dissimilarity-based classification approach, we build a dissimilarity space that maps bags of instances into a vector space in which further conventional statistical training procedures may be applied as suggested in [88]. Axes of this dissimilarity space are associated to prototypes from $\mathcal{R}$, thereby, the larger the cardinality of $\mathcal{R}$, the larger the dimensionality of the space.

In contrast to the instance-level methods, approaches based on dissimilarities often require tuning fewer parameters and allow incorporating global information by taking each bag-of-instances as a whole [72]. Also, the validation of these classifiers is simple once the dissimilarity matrix is estimated, providing similar performance to other state-of-the-art MIC approaches but at lower computational cost [25]. Nonetheless, the computation of distances between bags implies pairwise set comparisons, meaning that Minkowski metrics (that is, ℓ_p-norms between one-dimensional points) are not suitable. Instead, the concept of dissimilarity measure can be adopted by considering each bag-of-instances as a point set in a high-dimensional space.

4.2.4. Metric learning for dissimilarity adaptation

The enhancing (adaption) stage is frequently considered to emphasize particular aspects of the input representation in order to increase the classification performance. Thus, an original distance $d(\boldsymbol{S}_i, \boldsymbol{R}_l)$, which is applied between a test bag and a prototype bag, can be adapted $\tilde{d}(\boldsymbol{S}_i, \boldsymbol{R}_l)$ according to the following strategies [41]:

- *Locally adaptive nearest neighbor distance* (LANN). This adaptation, proposed in [116], searches for appropriate factors to scale dissimilarities between a test object and the training objects so, that reliable training objects are rewarded and, conversely, unreliable ones are penalized:

$$\tilde{d}_{LANN}(\boldsymbol{S}_i, \boldsymbol{R}_l) = d(\boldsymbol{S}_i, \boldsymbol{R}_l)/\rho_l. \tag{4-5}$$

  where $\rho_l = \min_{\forall k:\theta_k \neq \beta_l} d(\boldsymbol{T}_k, \boldsymbol{R}_l)$ is the maximum radius that excludes all bags $\boldsymbol{T}_k \in T$ from classes different to the class of bag $\boldsymbol{R}_l$ (see Fig. **4-2**), and, θ_k and β_l are the class labels associated to $\boldsymbol{T}_k$ and $\boldsymbol{R}_l$, respectively.

- *z-score.* This adaptation is the statistical standardization that scales and shifts each dissimilarity value as follows:

$$\tilde{d}_{z-s}(\boldsymbol{S}_i, \boldsymbol{R}_l) = (d(\boldsymbol{S}_i, \boldsymbol{R}_l) - \mu_l)/\sigma_l, \tag{4-6}$$

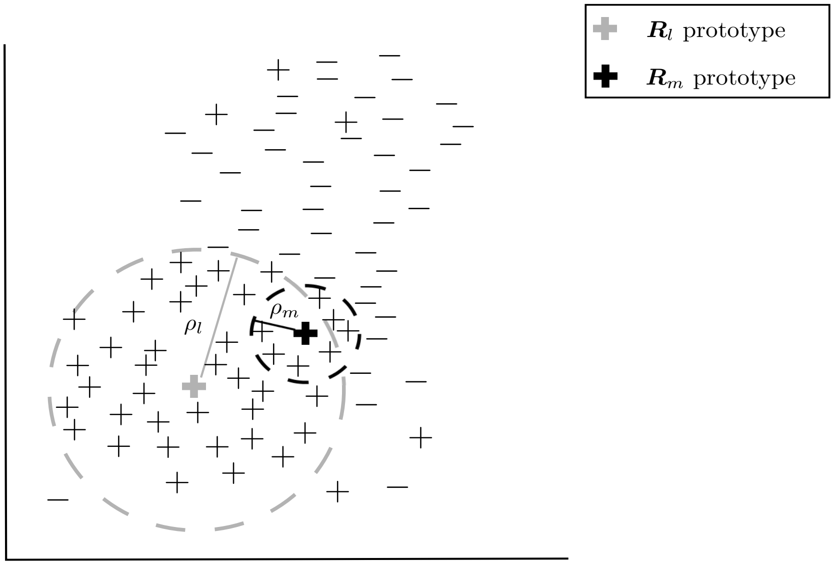


Figure 4-2. Illustration of the scaling radii ρ_l and ρ_m, corresponding to $\boldsymbol{R}_l$ and $\boldsymbol{R}_m$, as defined in (4-5). The radii ρ_l and ρ_m are the distances to the closest bag in the training set T to $\boldsymbol{R}_l$ and $\boldsymbol{R}_m$, respectively, whose labels differ from those of each prototype. Considering the size of the radii, it can be inferred that $\boldsymbol{R}_l$ is more reliable than $\boldsymbol{R}_m$.

where $\mu_l = \boldsymbol{E}\{d(\boldsymbol{T}_k, \boldsymbol{R}_l){:}\forall k\}$ and $\sigma_l = \sqrt{\boldsymbol{E}\{(d(\boldsymbol{T}_k, \boldsymbol{R}_l){:}\forall k\} - \mu_l)^2}$ are the mean and standard deviation, respectively, of the l-th column of the dissimilarity matrix computed between training objects and prototype objects. This adaptation affects the dissimilarity range in such a way that even if the dissimilarity is positive, z-score may produce a negative value, making infeasible to classify by directly using the 1-NN rule.

- *Adaptive distance in EigenSpace* L_p (ESL$_p$). This adaptation is the mapped ℓ_p-norm between each query object $\boldsymbol{S}_i$ and the prototype object $\boldsymbol{R}_l$. It rotates the dissimilarity space to avoid undesired effects of correlations of the dissimilarity data:

$$\tilde{d}_{ESL}(\boldsymbol{S}_i, \boldsymbol{R}_l) = \|\boldsymbol{c}_{\boldsymbol{S}_i} - \boldsymbol{c}_{\boldsymbol{R}_l}\|_p$$

where $\boldsymbol{c}_{\boldsymbol{S}_i} = \boldsymbol{C}\boldsymbol{d}_{\boldsymbol{S}_i}^{\top} \in \mathbb{R}^{P\times 1}$ and $\boldsymbol{c}_{\boldsymbol{R}_l} = \boldsymbol{C}\boldsymbol{d}_{\boldsymbol{R}_l}^{\top} \in \mathbb{R}^{P\times 1}$ are the vectorized versions of each mapped bag, matrix $\boldsymbol{C} = [\boldsymbol{\delta}_1 \dots \boldsymbol{\delta}_P] \in \mathbb{R}^{P\times P}$ is the projection matrix that holds eigenvectors sorted according to the ranked magnitude of the eigenvalues, i.e., it holds that $\lambda_1 \geqslant, \dots, \geqslant \lambda_P$. These eigenvalues and eigenvectors are extracted from the covariance matrix computed as $\boldsymbol{\Sigma} = [\Sigma_{ij}] \in \mathbb{R}^{P\times P}$, where $\Sigma_{ij} = \boldsymbol{E}\{(\boldsymbol{D}(\cdot,i) - \boldsymbol{E}\{\boldsymbol{D}(\cdot,i)\})(\boldsymbol{D}(\cdot,j) - \boldsymbol{E}\{\boldsymbol{D}(\cdot,j)\})^{\top}\}$, $\boldsymbol{D}(\cdot,i)$ denotes the i-th column of $\boldsymbol{D}$, $\boldsymbol{D} = [\boldsymbol{d}_{\boldsymbol{T}_1}^{\top} \cdots \boldsymbol{d}_{\boldsymbol{T}_M}^{\top}]^{\top} \in \mathbb{R}^{M\times P}$ is the dissimilarity matrix of the training set, and $\boldsymbol{d}_{\boldsymbol{T}_m} = [d(\boldsymbol{T}_k, \boldsymbol{R}_1) \dots d(\boldsymbol{T}_k, \boldsymbol{R}_P)] \in \mathbb{R}^{1\times P}$.

- *Nonlinear scaling of dissimilarities* (NLSD). This adaptation non-linearly and monotonically maps the dissimilarity space so that the non-linear mapping does not change the ordering of objects, but it changes the behavior of the classifier in the dissimilarity space:

$$\tilde{d}_{NLSD}(\boldsymbol{S}_i, \boldsymbol{R}_l) = d(\boldsymbol{S}_i, \boldsymbol{R}_l)^{\alpha}, \quad \alpha \in \mathbb{R}^{+} \tag{4-7}$$

  This adaptation reduces the effect of outliers whenever $\alpha<1$ since dissimilarities from distant points are shrunk. So, we choose α as the interval midpoint $(0,1)$, i.e., the square root $(\alpha = 0.5)$.

4.2.5. Classification strategies

Table **4-2** shows the tested strategies for classifying bags of instances from dissimilarity information that are enumerated from the simplest to the most complex one.

The first strategy, noted as *Strategy 1* and also known as template matching, is the simplest strategy that assigns labels by the 1-NN rule. The second one, *Strategy 2*, incorporates the adaptation of the dissimilarity measure in order to improve the performance achieved by template matching. This strategy is the same applied in [41] to classify different data sets by using dissimilarity information; however, they do not consider problems of bag of instances but only data sets already given as dissimilarity matrices. The third strategy, *Strategy 3*, carries out classification in the dissimilarity space, i.e., a vector space where each axis corresponds to a prototype object. Thus, it is possible to take advantage of the dissimilarity information on the training data and use more complex decision rules than template matching, such as linear classifiers, quadratic classifiers or SVM; this strategy is used in [25] to address the problem of classifying birdsong recordings. Finally, we propose the last strategy, *Strategy 4*, which combines *Strategy 2* and *Strategy 3*, aiming to benefit from both the properties of distance measure adaptation and classification in the dissimilarity space; moreover, it also allows using adaptation techniques that do not modify the result of the 1-NN rule such as the monotonic transformations, e.g., NLSD.

Table 4-2.Representation of an object in each classification strategy. Notation $\tilde{d}(\cdot,\cdot)$ stands for adapted measures.

Strategy	*Object representation*	*Notation*
Strategy 1	Original dissimilarities	$d(\boldsymbol{S}_i, \boldsymbol{R}_l)$
Strategy 2	Adapted dissimilarities	$\tilde{d}(\boldsymbol{S}_i, \boldsymbol{R}_l)$
Strategy 3	Coordinates in the dissimilarity space	$\boldsymbol{d}_{\boldsymbol{S}_i} = [d(\boldsymbol{S}_i, \boldsymbol{R}_1) \ldots d(\boldsymbol{S}_i, \boldsymbol{R}_P)]$
Strategy 4	Coordinates in the dissimilarity space	$\tilde{\boldsymbol{d}}_{\boldsymbol{S}_i} = [\tilde{d}(\boldsymbol{S}_i, \boldsymbol{R}_1) \ldots \tilde{d}(\boldsymbol{S}_i, \boldsymbol{R}_P)]$

4.3. Experiments

In order to validate the proposed dissimilarity-based classification of birdsong recordings, we use the methodology displayed in Fig **4-3** that considers the stages described in Sec. 4.2. It is worth noting that two well-known validation models are performed:

a) Leave-one-out cross-validation that we use to compare the discriminant ability of the contrasted dissimilarity measures (second stage). In this case, we employ the 1-NN algorithm classifier since its performance strongly depends on the examined distance and does not demand any parameter tuning.

b) Cross-validation that is carried out for five folds and ten repetitions for comparing the performance of the considered classification strategies.

4.3.1. Birdsong dataset

We employ the public dataset[4] of recordings collected in the H. J. Andrews (HJA) forest, in 2009, that has been used in other previous studies [19, 18, 25]. The raw recordings were acquired by using 13 Wildlife Acoustics Song Meter SM1 devices on six locations with a sample rate of 16 kHz, within the range of 5:00 am to 5:20 am because birds are very active during that period. In order to automatically extract instances from the spectrograms of the recordings, 625 manually-segmented spectrograms were used to train the segmentation algorithm described in Sec. 4.2.1. Such a segmentation algorithm is used to generate bags of instances from the spectrograms of an independent set of 548 10-second recordings, which contains sounds of 13 bird species (Table **4-3**). Ground-truth class labels were assigned by

[4] The same dataset employed in Chapter 3.

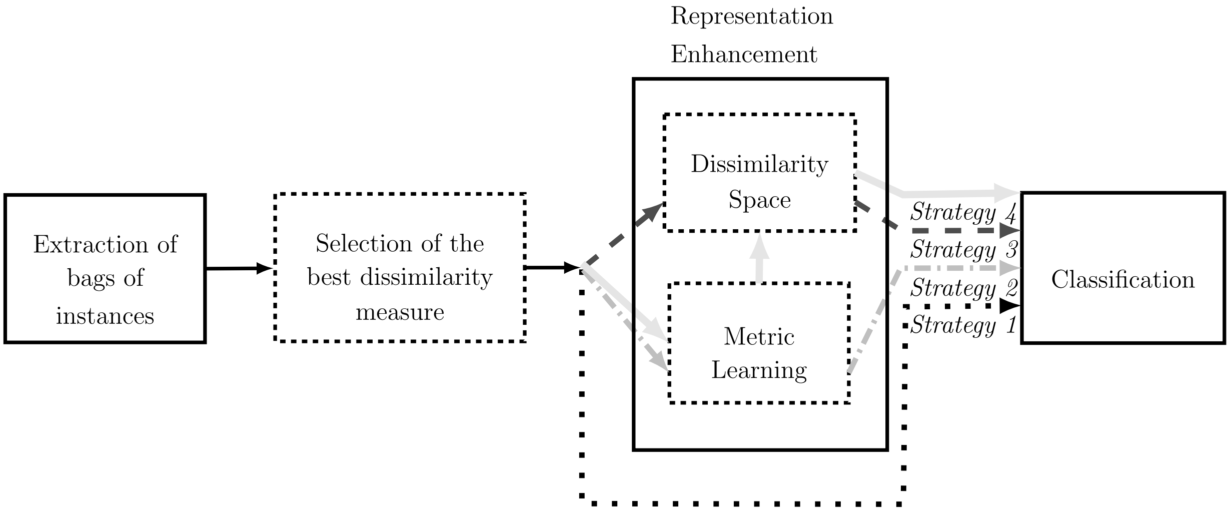


Figure 4-3.Proposed methodology for dissimilarity-based classification of birdsong recordings represented as multiple-instance objects. Four strategies for training are studied. Boxes marked in dashed lines are subjects of investigation.

an expert who inspected the recordings, see acknowledgments in [19, 18]. Each recording can hold sounds vocalized by one up to five species. Consequently, the data set is cast into five two-class MIC data sets by using the one-against-all strategy. Specifically, we alternately designate each one of the classes with more than 100 recordings as the target or positive class, while labeling the remaining classes as negative (see the highlighted cells in Table **4-3**).

4.3.2. Selection of the best dissimilarity measure

We compare the discriminant ability of each dissimilarity measure in terms of the classification performance. In this case, we employ the leave-one-out cross-validation model together with the 1-NN classifier, which requires the selection of a proper dissimilarity measure as indicated in [61]. Particularly, we select the modified Hausdorff distance that reaches the best classifier performance almost for all tested data sets as seen in Table **4-4**; although the mean-minimum distance accomplishes a high classification performance as well. To handle the class imbalance of the tested data sets, we use the F-score as the performance measure.

4.3.3. Comparison of enhanced dissimilarity measures

We estimate the classification performance of the learning representation for the four considered adaptation techniques: LANN, z-score, NLSD, and ESL (as in [41], $p = 1.5$). Refer

Table 4-3.Amount of 10-second birdsong recordings and class labels of the HJA data set.

Abbreviation (Class label)	*Name*	*Latin name*	*Amount of recordings*
BRCR	Brown Creeper	*Certhia americana*	197
WIWR	Winter Wren	*Troglodytes hiemalis*	109
PSFL	Pacific-slope Flycatcher	*Empidonax difficilis*	165
RBNU	Red-breasted Nuthatch	*Sitta canadensis*	82
DEJU	Dark-eyed Junco	*Junco hyemalis*	20
OSFL	Olive-sided Flycatcher	*Contopus cooperi*	90
HETH	Hermit Thrush	*Catharus guttatus*	15
CBCH	Chestnut-backed Chickadee	*Poecile rufescens*	117
VATH	Varied Thrush	*Ixoreus naevius*	89
HEWA	Hermit Warbler	*Setophaga occidentalis*	63
SWTH	Swainson's Thrush	*Catharus ustulatus*	79
HAFL	Hammond's Flycatcher	*Empidonax hammondii*	103
WETA	Western Tanager	*Piranga ludoviciana*	46

Table 4-4.Comparison of the examined distances in terms of F-score.

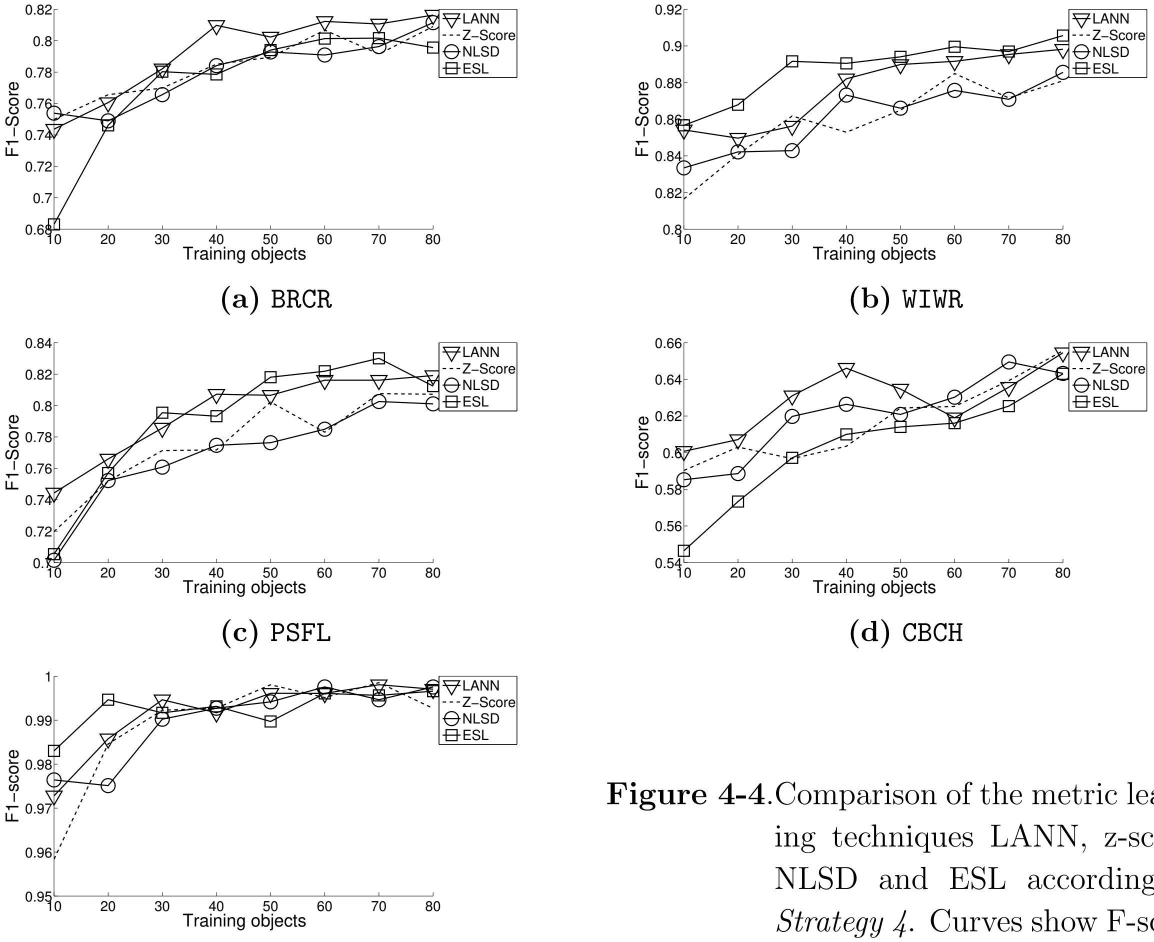
Distance\Class	BRCR	WIWR	PSFL	CBCH	HAFL
Standard Hausdorff	0.7537	0.8722	0.7432	0.6614	0.9394
Mean minimum	**0.7970**	0.9091	0.8318	0.6506	0.9852
Overall minimum	0.6838	0.7636	0.7438	0.5507	0.9756
Modified Hausdorff	0.7921	**0.9099**	**0.8328**	**0.6667**	**0.9852**

again to Fig. **4-3** and notice that the block "Metric learning" is part of *Strategy 2* and *Strategy 4*. However, *Strategy 2* uses 1-NN for classification, whose decision is not affected by monotonic transformations as NLSD. Therefore, in order to compare the four adaptation techniques within the same framework, we only use *Strategy 4* for this experiment. Since HAFL contains the minimum amount of recordings per class (103), the maximum number of training objects belonging to this class is 80 when a 5-fold cross-validation is carried out. Remember that the dimensionality of the dissimilarity space is equal to the number of prototypes (refer back to Sec. 4.2.3) and, therefore, classifiers may suffer from the curse of dimensionality phenomenon[5] if the number of prototypes is too large in comparison with the cardinality of the training set. Consequently, a convenient dimensionality must be fixed in order to avoid that phenomenon. In [61], it is shown an example when a ratio of 1/5 between the dimensionality of the space and the number of training objects is optimal and

[5] A brief introduction to this phenomenon is available at `http://www.37steps.com/2349/curse-of-dimensionality/`.

recommend, in general, a ratio of 1/10 to be in the safe side. In our case, considering the above-mentioned 80 objects from the target class, we set the dimensionality of the dissimilarity space to 20, in such a way that the set of the prototypes is formed by 10 prototypes from the target class and 10 prototypes from the non-target class.

As seen in Figs. **4-4** and **4-5** that display the obtained F-scores depending on the number of training objects and the best recall versus precision point obtained for each adaptation, respectively, all four strategies perform similarly for classification of `BRCR` class. In turn, ESL accomplishes the highest performance for `WIWR` and `PSFL` (see Figs. **4-4(b)** and **4-4(c)**) classes, but it fails for the `CBCH` class (see Fig. **4-4(d)**). However, we choose LANN as the adaptation technique since it evenly exhibits a high performance for all classes.



(a) `BRCR`

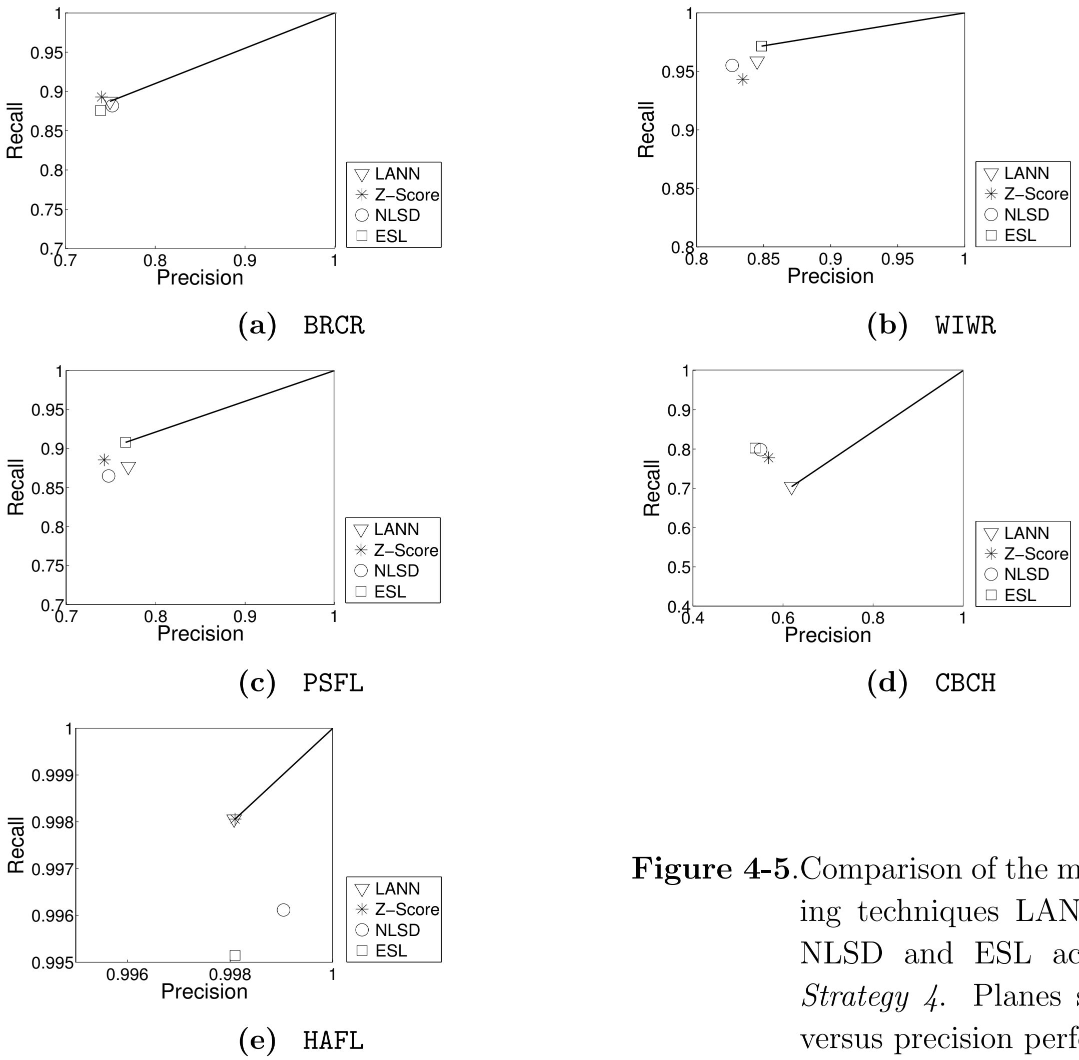
(b) `WIWR`

(c) `PSFL`

(d) `CBCH`

(e) `HAFL`

Figure 4-4. Comparison of the metric learning techniques LANN, z-score, NLSD and ESL according to *Strategy 4*. Curves show F-score versus training set sizes.



(a) BRCR

(b) WIWR

(c) PSFL

(d) CBCH

(e) HAFL

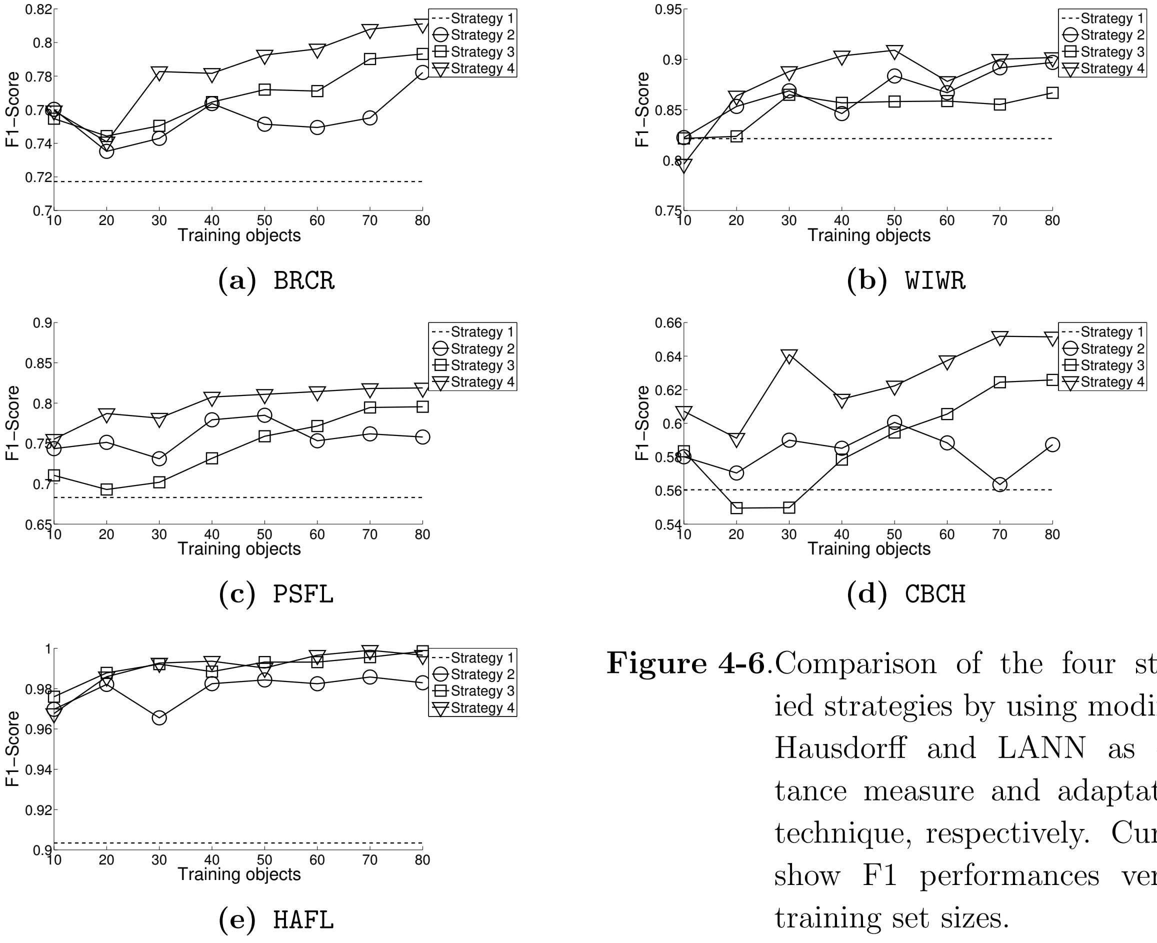
Figure 4-5. Comparison of the metric learning techniques LANN, z-score, NLSD and ESL according to *Strategy 4*. Planes show recall versus precision performance.

4.3.4. Classification and validation

Two classifiers are employed in order to validate all the tested strategies *i*) 1-NN that is directly used in *Strategy 1* and Strategy *2* to infer the labels; and *ii*) linear SVM in dissimilarity spaces carried out by *Strategy 3* and Strategy *4*, as suggested in [42]. For any classifier, we fix the number of prototypes to 20 so that 10 for each class (positive and negative) are randomly chosen in order to avoid the curse of dimensionality, as explained in Sec. 4.3.3. The regularization SVM parameter is heuristically tuned ($C = 10$). In the experiments, different sizes for the training set are tested (from 10 to 80 objects per class). Notice that the set of prototypes (fixed to 10 per class) is a subset of the training set.

Overall, the simplest classification (*Strategy 1*) accomplishes the lowest classification performance as seen in Figs. **4-6** and **4-7**. In comparison to *Strategy 1*, the use of *Strategy*

2 improves the classifier performance for `BRCR`, `PSFL`, and `HAFL` classes (see Figs. **4-6(a)**, **4-6(c)**, and **4-6(e)**, respectively). However, *Strategy 2* fails at some values of the training set for `WIWR` and `CBCH` classes (Figs. **4-6(b)** and **4-6(d)**, respectively), providing irregular performance. In turn, the incorporation of *Strategy 3* improves further the performance achieved for all classes as the size of the training set grows, when the SVM classifier extracts enough information from the training set. Lastly, the best strategy is *Strategy 4* for all tested classes. Moreover, this strategy clearly benefits from the input data information since the provided performance improves as the size of training set increases. Also, it outperforms the baseline *Strategy 1* even when the training set is small (10 objects per class). From Figs. **4-5** and **4-7**, we can observe that the classification system is a bit more sensitive (high recall) than precise.



(a) `BRCR`

(b) `WIWR`

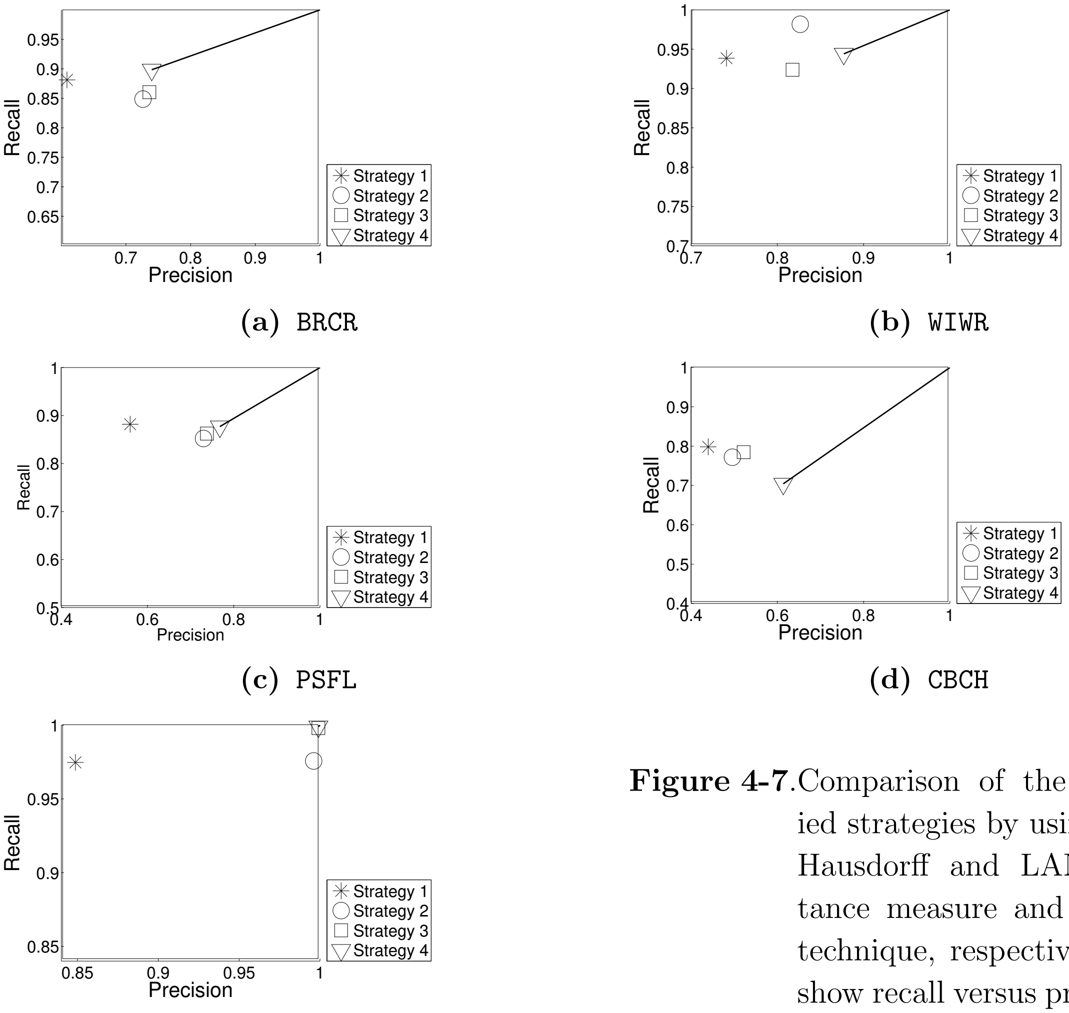
(c) `PSFL`

(d) `CBCH`

(e) `HAFL`

Figure 4-6. Comparison of the four studied strategies by using modified Hausdorff and LANN as distance measure and adaptation technique, respectively. Curves show F1 performances versus training set sizes.

As a result, the insertion of dissimilarity space along with the adaptation stage of dissimilarities allows a larger enhancement, and thus a better performance of the system.



(a) BRCR

(b) WIWR

(c) PSFL

(d) CBCH

(e) HAFL

Figure 4-7. Comparison of the four studied strategies by using modified Hausdorff and LANN as distance measure and adaptation technique, respectively. Planes show recall versus precision performance.

4.4. Discussion

We validate the classification enhancement of birdsong recordings by using metric learning techniques and dissimilarity spaces. From the obtained results on the HJA birdsong data set, the following findings are worth to be mentioned:

Modified Hausdorff and mean-minimum distances provide the best performance among the four compared distances (overall minimum, mean-minimum, standard Hausdorff, and modified Hausdorff). Both of them consider the global distribution of the instances, meaning that they take into account the whole set of sounds in the recording, i.e., vocalizations and environmental noise. Even though, the standard Hausdorff distance also considers the global distribution of instances in each bag, it is too strict and, therefore, outliers can produce mistaken dissimilarity values. Our result coincides with the finding in [40]. Finally, overall minimum distance —particularly in the bioacoustic data set— is affected by the environmental noise, e.g., in case that the rain sound is not completely removed in the segmentation stage, an underestimated low dissimilarity value will be produced.

The next aspect for consideration is the influence of the adaptation of the dissimilarities between bags by using metric learning techniques. After comparing LANN, z-score, ESL and NLSD, the best adaptation technique is LANN, closely followed by ESL. However, the first one is selected due to its simple interpretation.

In the last experiment, the classification strategies are compared. Results show that *Strategy 4* outperforms the others, which is followed, in the given order, by *Strategy 3*, *Strategy 2*, and *Strategy 1*. As it can be seen, the order matches with the complexity of the strategies, in such a way that the more complex the strategy, the better the performance. Notice that despite our methodology can be easily applied to classify more species since we use the one-against-all approach, it is recommended to have a sufficient number of training objects per class. In addition, taking into account that when a fixed size of the representation set is considered, all the strategies take almost the same computation time for classifying a test object. Nevertheless, in order to scale our methodology to much larger data sets, a computationally-tractable size of the set of prototypes must be fixed; however, *Strategy 1* might suffer from that because it does not allow to include additional information to the system that the one provided by the representation set. For this reason, we recommend to use *Strategy 2*, *Strategy 3* or *Strategy 4* when more training objects are available to build the system. Particularly, the last one that takes advantage of the information of the training set in two stages of the process (metric learning and training in dissimilarity spaces) and exhibits the best results in our experiments.

As valuable results, we highlight the following two: i) to consider the dissimilarity information of the training set is relevant, as made by Strategy *2*, Strategy *3*, and Strategy *4* which remarkably outperform the matching technique (*Strategy 1*) and ii) classification in dissimilarity spaces is significantly enhanced by adapting the dissimilarities between bags.

To sum up, we focused on classifying birdsong recordings from dissimilarity information that simplifies its solution as a classical MIC problem. We experimentally verified that metric learning as well as dissimilarity spaces are valuable approaches to take advantage of the dissimilarity information and, furthermore, we found that the learning strategy is more powerful by combining both than separately using each one.

5. Improved multiple instance classification using relative minimum distance between projected bags

Abstract

The one nearest neighbor (1-NN) rule relies on a distance measure between the objects to be classified. In multiple instance classification (MIC) problems, the 1-NN rule assigns to a query bag the label of the closest object in a set of prototype bags or instances. In the past, it has been proved that dissimilarity-based MIC approaches work well for a wide range of datasets, particularly, when the mean-minimum distance is applied. However, most of these approaches mask interpretability of the data because MIC does not require that the dissimilarity between objects behaves as intuitively expected: the closer the objects, the more likely to belong to the same class. Hence, we introduce a methodology for improving the dissimilarity-based MIC, which uses a supervised method to compute a projected space, and a novel *relative minimum distance* to compare bags and prototype bags. Our experiments show that the classification performance, when the proposed methods are applied, is higher than the baseline for most of the studied real-world datasets.

5.1. Introduction

Multiple Instance Classification (MIC) consists in learning from patterns represented as sets of feature vectors (termed *bags*), where each bag holds multiple feature objects (*instances*), so that a label is attached to each bag and not to every instance. Moreover, there might be instances inside a bag, that do not convey any information about its class, or that are more related to other classes of bags, providing confusing information [4]. To deal with the scarcely characterized examples, and assuming that the discriminative information lies at the instance-level, the binary classification setting separates the instances in *positive* and *negative* ones, defining the location of the feature space where positive instances are located as the *concept*.

In order to infer the concept in MIC problems, several techniques try to find the region of the space where positive bags intersect while not containing negative instances, e.g., the Diverse Density method proposed in [76] that looks for an area where there is both high density of positive points and low density of negative points. Likewise, an MIC model is proposed in [99] that requires that a positive bag does not have any points near the so-called "repulsion" points, in addition to having a point near the "target" points. Therefore, it means that positiveness implies that some positive instances have to be present in the bag, and negativeness requires that positive instances must be absent. Much of the recent work in MIC has been concentrated on a relaxed view of the standard assumption, considering alternative assumptions instead, e.g., the collective assumption where all instances in a bag equally contribute to the bag-level labels or bags of words that group classes of instances.

Intending to improve the performance of instance-level MIC algorithms, the dissimilarity-based approaches have been proposed in [25], using pairwise dissimilarity measures that directly compare bags, rather than relying on locating a concept. Thus, a single label is inferred for each query bag, which is represented by a vector that holds the distances between the query bag and a set of prototypes. Therefore, we obtain a dissimilarity representation that is similar to a traditional feature vector space, for which a standard supervised classifier can be employed. With the purpose of comparing the training/test bags against prototypes, there are two type of approaches: i) *Bag-to-instance distances*. These approaches capture the instance-level information, like the Multiple Instance Learning via Embedded Instance Selection, that maps each bag into a high-dimensional feature vector [22]. In this case, however, the classifier has to be appropriately selected and trained to perform well in high-dimensional representations [24]. ii) *Bag-to-bag distances*. This option allows working in a low-dimensional space, but it usually misses instance-level information, i.e., discriminant instances are masked by non-discriminant ones. Nonetheless, it would be desirable that the measure fulfills the following properties: i) it enables to compare a bag and a prototype bag, for building a low-dimensional representation, and ii) it enhances the instances that most likely belong to the concept, intending to benefit from the instance-level information.

To improve the performance of dissimilarity-based MIC algorithms, we propose a dissimilarity measure, termed *relative minimum distance*, that compares bags and prototype bags, considering local instance-level information. To this end, we project the input feature space in advance, to bring closer the instances that most likely represent the concept, yielding a representation space where the closer the bags are, the more likely they belong to the same class. Particularly, relying on the bag labels, we compute a projected space by using Multiple Instance Logistic Discriminant Metric Learning (MildML) [52]. Using a simple one-nearest-neighbor (1-NN) classifier, validation is carried out on several real-world datasets, proving that the proposed dissimilarity measure between bags in a projected space, outperforms the baseline dissimilarities used for MIC.

5.2. Methods

5.2.1. Dissimilarity-based Multiple Instance Classification

MIC is a weakly supervised approach, that represents each object in a feature space by a set of points $\mathcal{S}=\{\boldsymbol{s}_1,\dots,\boldsymbol{s}_N\}$ –termed *bag of instances*– where $\boldsymbol{s}_i\in\mathbb{R}^{n\times 1}$ is the i-th instance [39]. Each bag consists of two subsets of instances, one positive and another one negative: $\mathcal{S}=\overset{\oplus}{\mathcal{S}}\cup\overset{\ominus}{\mathcal{S}}$. Thus, a positive subset holds only positive instances ($\overset{\oplus}{\mathcal{S}}=\{\boldsymbol{s}_1^+,\dots,\boldsymbol{s}_{N^+}^+\}$, with $N^+\leqslant N$), i.e., the instances that are directly related to the concept. In contrast, a negative subset has only negative instances that are not related to the concept ($\overset{\ominus}{\mathcal{S}}=\{\boldsymbol{s}_1^-,\dots,\boldsymbol{s}_{N^-}^-\}$, with $N^-\leqslant N$). Besides, $\mathcal{S}$ is positive if there is, at least, one instance in the bag relating to the concept ($\mathcal{S}^+$), that is, if $\overset{\oplus}{\mathcal{S}}\neq\varnothing$, otherwise, $\mathcal{S}$ is negative ($\mathcal{S}^-$).

Provided a test bag $\mathcal{S}$, the dissimilarity-based MIC –using 1-NN– infers a label according to the closest bag in the representation set $\{\mathcal{R}_1,\dots,\mathcal{R}_Q\}$, composed of Q prototype bags. That is: the label assigned is the one that corresponds to the minimum entry of the distance vector $\boldsymbol{d}_{\mathcal{S}}=[d(\mathcal{S},\mathcal{R}_1)\dots d(\mathcal{S},\mathcal{R}_Q)]$, being $d(\cdot,\cdot)\in\mathbb{R}^+$ a dissimilarity measure. Because of their proved suitability for MIC tasks, in this chapter, we consider the following three dissimilarity measures between bags [25]: *i*) Hausdorff distance (d_{H}), *ii*) Mean-minimum distance ($d_{\overline{\min}}$), and *iii*) Overall-minimum distance ($d_{\min}$), computed, respectively, as follows:

$$\begin{aligned}
d_{\mathrm{H}}(\mathcal{S}_k,\mathcal{S}_l) &=\max(\max_q(\min_m d(\boldsymbol{s}_m^{(k)},\boldsymbol{s}_q^{(l)})),\max_m(\min_q d(\boldsymbol{s}_m^{(k)},\boldsymbol{s}_q^{(l)}))),\\
d_{\overline{\min}}(\mathcal{S}_k,\mathcal{S}_l) &=\boldsymbol{E}\{\boldsymbol{E}\{\min_m d(\boldsymbol{s}_m^{(k)},\boldsymbol{s}_q^{(l)}):\forall q\},\boldsymbol{E}\{\min_q d(\boldsymbol{s}_m^{(k)},\boldsymbol{s}_q^{(l)}):\forall m\}\},\\
d_{\min}(\mathcal{S}_k,\mathcal{S}_l) &=\min_{m,q} d(\boldsymbol{s}_m^{(k)},\boldsymbol{s}_q^{(l)}),
\end{aligned}$$

where $\mathcal{S}_k=\{\boldsymbol{s}_1^{(k)},\dots,\boldsymbol{s}_{N_k}^{(k)}\}$ and $\mathcal{S}_l=\{\boldsymbol{s}_1^{(l)},\dots,\boldsymbol{s}_{N_l}^{(l)}\}$ are the bags of instances that respectively hold N_k and N_l n-dimensional instances, $d(\boldsymbol{s}_m^{(k)},\boldsymbol{s}_q^{(l)})$ is a dissimilarity measure between instances (it is customary to choose the Euclidean distance), and notation $\boldsymbol{E}\{\cdot\}$ stands for the expectation operator. Note that the 1-NN rule demands the following asymptotic behavior of dissimilarities between bags: $d(\mathcal{S}^+,\mathcal{R}^+)\approx 0$, $d(\mathcal{S}^+,\mathcal{R}^-)\gg 0$, $d(\mathcal{S}^-,\mathcal{R}^+)\gg 0$, and $d(\mathcal{S}^-,\mathcal{R}^-)\approx 0$.

5.2.2. A relative minimum distance in a projected space

To associate the concept with a particular region, we adopt a projection that maps the input space, such that at least one instance of each positive training bag is brought as close as possible to the others, while instances of the negative bags must be driven apart. Specifically, we use the *MildML* method that maximizes the following loss function [52]:

$$\max_{\boldsymbol{L},b}\ \mathcal{L} = \max_{\boldsymbol{L},b}\ \sum\nolimits_{k,l} \lambda_{kl} \log p_{kl} + (1-\lambda_{kl})\log(1-p_{kl}) \tag{5-1}$$

where λ_{kl} is 1 if both ($\mathcal{S}_k$ and $\mathcal{S}_l$) are positive bags, otherwise, it is 0, $p_{kl} \in \mathbb{R}[0,1]$ is the probability that $\mathcal{S}_k$ and $\mathcal{S}_l$ are together positive, and is calculated by the *sigmoid* function as follows:

$$p_{kl} = \left(1+\exp\left(-b+d(\widetilde{\mathcal{S}}_k,\widetilde{\mathcal{S}}_l)^2\right)\right)^{-1}$$

where $b \in \mathbb{R}$ is a shifting parameter, $d(\cdot,\cdot) \in \mathbb{R}^+$ is a dissimilarity measure between the projected bags $\widetilde{\mathcal{S}} = \{\tilde{\boldsymbol{s}}_n^{\boldsymbol{L}} : n \in N\}$, which hold the instances in the mapped space through a projection matrix $\boldsymbol{L} \in \mathbb{R}^{n\times n}$, that is, $\tilde{\boldsymbol{s}}_n^{\boldsymbol{L}} : \mathbb{R}^{n\times 1} \to \mathbb{R}^{n\times 1}$.

The optimization task in (5-1) can be solved iteratively by the conjugate gradient algorithm [37], employing the update rule:

$$\boldsymbol{L}_{k+1} = \boldsymbol{L}_k + \alpha_k \boldsymbol{P}_k,$$

where $\alpha_k \in \mathbb{R}$ is the step length, and $\boldsymbol{P}_k \in \mathbb{R}^{n\times n}$ is the search direction that is fixed to $\boldsymbol{P}_k = -\boldsymbol{G}_k + \beta_k \boldsymbol{P}_{k-1}$. In the case of the Polak-Ribiére update, the scalar-valued step $\beta_k \in \mathbb{R}$ is computed as $\beta_k = (\Delta\boldsymbol{\gamma}_{k-1}^\top \boldsymbol{\gamma}_k)/(\boldsymbol{\gamma}_{k-1}^\top \boldsymbol{\gamma}_{k-1})$, where $\Delta\boldsymbol{\gamma}_k \in \mathbb{R}^{n^2\times 1}$ denotes the vectorized version of the gradient $\Delta\boldsymbol{G}_k = \boldsymbol{G}_{k+1} - \boldsymbol{G}_k$ (with $\Delta\boldsymbol{G}_k \in \mathbb{R}^{n\times n}$). Likewise, $\boldsymbol{\gamma}_k \in \mathbb{R}^{n^2\times 1}$ is the vectorized version of the loss function gradient $\boldsymbol{G}_k = \partial\mathcal{L}/\partial\boldsymbol{L}$ (with $\boldsymbol{G}_k \in \mathbb{R}^{n\times n}$), defined for the MildML method in terms of $\boldsymbol{L}$ as below:

$$\boldsymbol{G}_k = \boldsymbol{L}\sum\nolimits_{k,l}(\lambda_{k,l} - p_{k,l})(\boldsymbol{s}_k - \boldsymbol{s}_l)(\boldsymbol{s}_k - \boldsymbol{s}_l)^\top. \tag{5-2}$$

Therefore, the maximizing framework in (5-1) minimizes the distance $d(\tilde{\mathcal{S}}_k^+, \tilde{\mathcal{S}}_l^+)$ evaluated between the projected positive bags, maximizing at the same time the distance $d(\tilde{\boldsymbol{s}}_k^{\boldsymbol{L}}, \tilde{\boldsymbol{s}}_l^{\boldsymbol{L}})$ between the projected instances $\tilde{\boldsymbol{s}}_k^{\boldsymbol{L}} \in \mathbb{R}^{n\times 1}$ and $\tilde{\boldsymbol{s}}_l^{\boldsymbol{L}}$ that belong to the projected negative bags $\tilde{\mathcal{S}}_k^-$ and $\tilde{\mathcal{S}}_l^-$, respectively. Consequently, it allows the concept being confined, within a narrower region of the projected space than in the original space, since mapping reduces the distances between positive instances, while enlarges the distances between negative instances.

Although the concept is associated with a single region in the projected space, the distance between instances fulfills just the following trade-off conditions: $d(\boldsymbol{s}_k^+, \boldsymbol{s}_l^+) \approx 0$, while $d(\boldsymbol{s}_m^+, \boldsymbol{s}_q^-) \gg 0$. Moreover, the value $d(\boldsymbol{s}_m^-, \boldsymbol{s}_q^-)$ is not restricted, since negative instances can be present either in positive bags or negative bags, causing the considered distances between bags fail, in the following cases:

- $d_H(\mathcal{S}^+,\mathcal{R}^+)\gg 0$ or $d_H(\mathcal{S}^-,\mathcal{R}^-)\gg 0$, if there is at least an instance $\boldsymbol{s}$ such that $\min_{\boldsymbol{r}} d(\boldsymbol{s},\boldsymbol{r})\gg 0$ ($\boldsymbol{s}\in\mathcal{S}$ and $\boldsymbol{r}\in\mathcal{R}$), or at least an $\boldsymbol{r}$ such that $\min_{\boldsymbol{s}} d(\boldsymbol{s},\boldsymbol{r})\gg 0$.

- $d_{\overline{min}}(\mathcal{S}^+,\mathcal{R}^+)\gg 0$ or $d_{\overline{min}}(\mathcal{S}^-,\mathcal{R}^-)\gg 0$, if there are several instances $\boldsymbol{s}$ such that $\min_{\boldsymbol{r}} d(\boldsymbol{s},\boldsymbol{r})\gg 0$, or several $\boldsymbol{r}$ such that $\min_{\boldsymbol{s}} d(\boldsymbol{s},\boldsymbol{r})\gg 0$. In addition, $d_{\overline{min}}(\mathcal{S}^\pm,\mathcal{R}^\mp)\approx 0$, if there are several instances $\boldsymbol{s}$ such that $\min_{\boldsymbol{r}} d(\boldsymbol{s},\boldsymbol{r})\approx 0$, or several $\boldsymbol{r}$ such that $\min_{\boldsymbol{s}} d(\boldsymbol{s},\boldsymbol{r})\approx 0$.

- $d_{\min}(\mathcal{S}^\pm,\mathcal{R}^\mp)\approx 0$, if there is at least one pair of instances $\boldsymbol{s}$ and $\boldsymbol{r}$ such that $\min_{\boldsymbol{s},\boldsymbol{r}} d(\boldsymbol{s},\boldsymbol{r})\approx 0$. Also, $d_{\min}(\mathcal{S}^-,\mathcal{R}^-)\gg 0$, if $\min_{\boldsymbol{s},\boldsymbol{r}} d(\boldsymbol{s},\boldsymbol{r})\gg 0$.

Aiming to avoid the above-mentioned drawbacks in the distances between bags, we propose the relative minimum distance that, provided a test bag $\mathcal{S}$ and either case of prototypes $\mathcal{R}_r^*$, is defined as follows:

$$d_{\mathrm{rel}}(\mathcal{S},\mathcal{R}_r^*) = \begin{cases} \dfrac{1}{Q^+-1}\dfrac{\sum_{i\neq r} d_{\min}(\mathcal{S},\mathcal{R}_i^+)}{\sum_{i\neq r} d_{\min}(\mathcal{R}_r^+,\mathcal{R}_i^+)}, & \text{if}\,\mathcal{R}_r^* = \mathcal{R}_r^+ \\ \dfrac{1}{Q^+}\dfrac{\sum_i d_{\min}(\mathcal{R}_r^-,\mathcal{R}_i^+)}{\sum_i d_{\min}(\mathcal{S},\mathcal{R}_i^+)}, & \text{if}\,\mathcal{R}_r^* = \mathcal{R}_r^- \end{cases}$$

where Q^+ is the number of positive prototypes.

So, a measured value $d_{\mathrm{rel}}(\mathcal{S},\mathcal{R}_r^+)<1$ implies that $\mathcal{S}$ becomes "more positive" than the r-th positive prototype. In contrast, $d_{\mathrm{rel}}(\mathcal{S},\mathcal{R}_r^-) < 1$ yields a "less positive" $\mathcal{S}$ than the r-th negative prototype. Consequently, the lower the value of $d_{\mathrm{rel}}(\cdot,\cdot)$ – the higher the probability of belonging to the class of the corresponding prototype. Then, the problematic issue of dissimilarity measures between bags is avoided for the 1-NN rule.

5.3. Experiments

For appraising the effectiveness of the tested dissimilarity measures, we estimate the F-score as the classification performance, using a leave-one-out validation. In particular, the 1-NN rule is applied to each bag of the dataset, assuming the complete input data as the prototype set, but removing the query bag. In the experimental setup, validation is accomplished on the real-world datasets described in Table **5-1** (publicly available on `http://www.miproblems.org/`). Note that the performance evaluation is carried out on both the original and projected spaces as seen in Table **5-2**.

Table 5-1. Real-world datasets described by the number of positive bags ($\mathcal{S}^+$), number of negative bags ($\mathcal{S}^-$), number of features per instance (Dimension), total number of instances and minimum and maximum (min and max, respectively) number of instances per bag.

Application	*Dataset*	$\mathcal{S}^+$	$\mathcal{S}^-$	*Dimension*	*Instances*	min	max
Molecule activity	`Musk 1`	47	45	166	476	2	4
	`Musk 2`	39	63	166	65598	1	1044
	`Mutagenesis1`	125	63	7	10486	28	88
	`Mutagenesis2`	13	29	7	2132	26	86
Images	`Fox`	100	100	230	1302	2	13
	`Tiger`	100	100	230	1220	1	13
	`Elephant`	100	100	230	1391	2	13
Audio classification	`Brown Creeper`	197	351	38	10232	2	43
	`Winter Wren`	109	439	38	10232	2	43
	`Pacific-slope Flycatcher`	165	383	38	10232	2	43
	`Red-breasted Nuthatch`	82	466	38	10232	2	43
	`Dark-eyed Junco`	20	528	38	10232	2	43
	`Olive-sided Flycatcher`	90	458	38	10232	2	43
	`Hermit Thrush`	15	533	38	10232	2	43
	`Chestnut-backed Chickadee`	117	431	38	10232	2	43
	`Varied Thrush`	89	459	38	10232	2	43
	`Hermit Warbler`	63	485	38	10232	2	43
	`Swainson's Thrush`	79	469	38	10232	2	43
	`Hammond's Flycatcher`	103	445	38	10232	2	43
	`Western Tanager`	46	502	38	10232	2	43

5.3.1. Multiple Instance Classification in the original space

Firstly, we directly apply the 1-NN classifier to the dissimilarities between bags, computed in the original space. Note that $d_{\overline{\min}}$ exhibits the best performance for most of the datasets. An advantage of this distance is that it considers the overall distribution of the bag of instances but is not so sensitive to outliers as d_H or $d_{\min}$. However, $d_{\overline{\min}}$ fails for some of the classification problems, e.g., with the `Hermit Thrush` dataset of the audio application, for which the number of negative instances may be remarkably higher than the number of positive instances in the positive bags. Also, d_{rel} fails in most of the cases, since it is not guaranteed that the concept is located in a particular region of the original space.

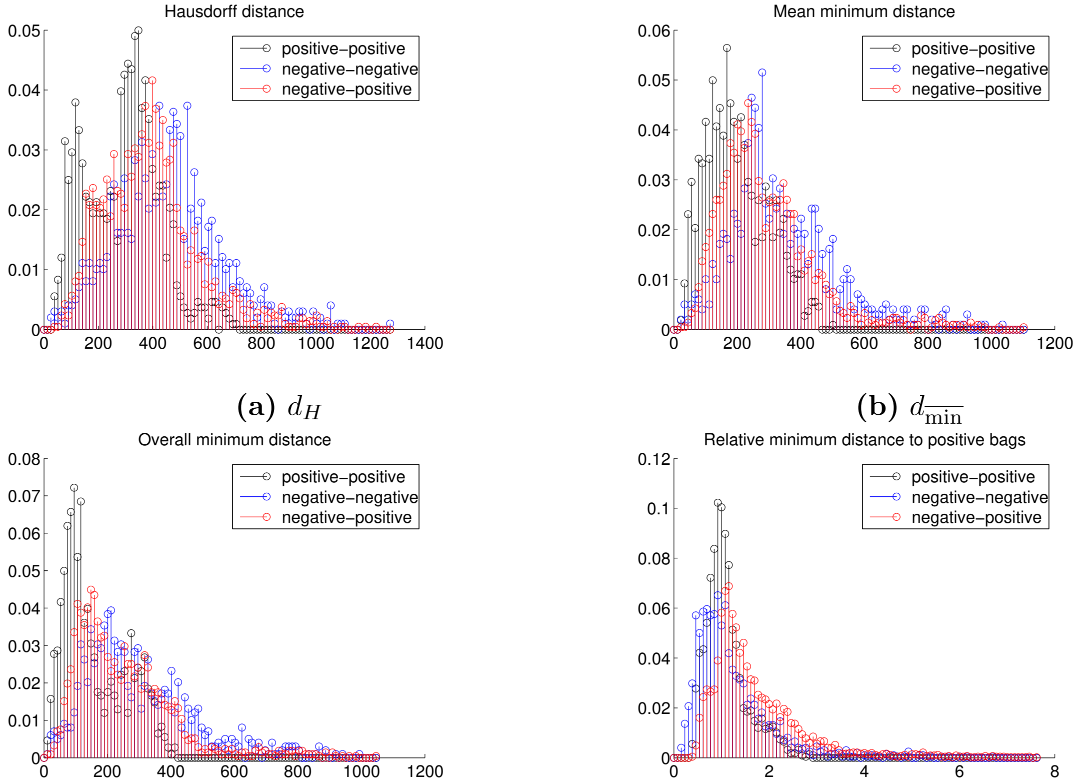
Table 5-2.1-NN leave-one-out test (F-score performance) applying the four studied dissimilarity measures between bags in the original space and the projected space. Best performances are highlighted in boldface.

Dataset	*Original space*				*Projected space*			
	d_H	$d_{\overline{\min}}$	$d_{\min}$	d_{rel}	d_H	$d_{\overline{\min}}$	$d_{\min}$	d_{rel}
Musk1	86.60	88.89	86.32	72.38	86.54	91.26	91.26	**100**
Musk2	80.00	71.91	70.45	60.00	78.26	80.43	88.64	**100**
Mutagenesis1	86.31	**87.45**	81.25	78.60	84.77	**87.45**	80.93	79.72
Mutagenesis2	43.48	45.45	52.17	40.91	41.67	**61.54**	50	59.26
Fox	56.60	62.50	53.27	68.77	59.16	68.14	77.87	**98.49**
Tiger	70.71	76.04	73.10	64.31	66.03	80.89	87.72	**100**
Elephant	76.78	81.65	79.07	69.50	75.00	82.70	80.65	**100**
Brown Creeper	75.37	**80.20**	68.38	68.95	73.32	75.40	68.38	72.05
Winter Wren	87.23	91.32	76.36	79.41	88.59	**94.12**	78.93	83.94
Pacific-slope Flycatcher	74.32	**83.79**	74.38	69.85	66.86	76.58	78.91	81.94
Red-breasted Nuthatch	67.50	75.47	57.65	49.66	54.88	67.01	70.39	**100**
Dark-eyed Junco	25.53	37.50	20.00	26.67	24.00	40.82	30.30	**97.44**
Olive-sided Flycatcher	63.59	**82.61**	56.97	57.75	70.72	72.36	64.82	67.15
Hermit Thrush	7.14	25.00	11.11	14.29	18.18	20.69	26.32	**96.55**
Chestnut-backed Chickadee	66.14	65.35	55.07	48.52	60.56	70.54	77.42	**94.42**
Varied Thrush	95.40	**99.44**	87.18	73.59	96.00	96.63	83.96	98.88
Hermit Warbler	68.75	75.76	68.12	77.78	71.43	76.12	70.06	**93.94**
Swainson's Thrush	77.03	86.53	52.50	30.13	65.31	69.51	65.74	**92.62**
Hammond's Flycatcher	93.94	98.52	97.56	34.77	78.92	98.08	97.63	**100**
Western Tanager	54.95	85.06	49.44	47.15	57.14	78.43	55.56	**96.70**

5.3.2. Multiple Instance Classification in the projected space

Table **5-2** shows that the proposed dissimilarity measure, d_{rel}, in the projected space remarkably outperforms the rest of the compared measures for most of the tested datasets. Nonetheless, the distance $d_{\overline{\min}}$ exhibits the best performance in some cases. When $d_{\overline{\min}}$ outperforms others, the bags of instances might mainly hold positive instances.

As mentioned in Sec. 5.2.1, some distances for multiple instance datasets are prone to behave in a non-intuitive way, since they may produce high values between correctly labeled bags of the same class, and small values between bags that belong to different classes. To illustrate the overall behavior of each estimated measure, we plot the normalized histograms obtained from the considered dissimilarity measures —in the original space (see Fig. **5-1**) as well as in the projected space (see Fig. **5-2**)— between the following cases of bags in the `Musk1` dataset: positive-positive bags, negative-negative bags, and negative-positive bags. Note that the histogram of a good dissimilarity measure should contain most of the distances between bags of the same class, positive-positive or negative-negative, on the left of the horizontal axis (i.e., smaller dissimilarities), and most of the distances between bags of different classes, negative-positive, on the right of the horizontal axis (i.e., larger dissimilarities).



(a) d_H

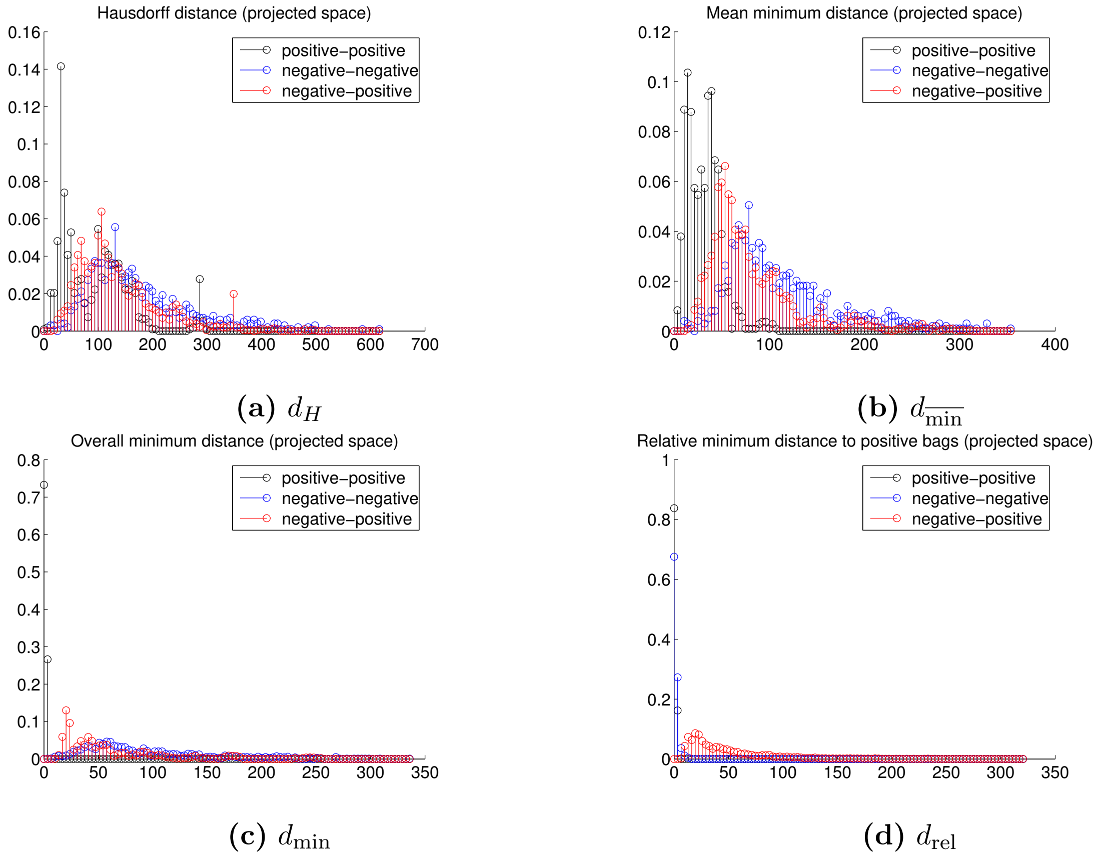
(b) $d_{\overline{\min}}$

(c) $d_{\min}$

(d) d_{rel}

Figure 5-1.Normalized histograms of distances in the original feature space (Musk1 dataset).

Figure **5-1** shows that, in the original space, there is no clear difference in the distribution of the dissimilarities between neither bags of the same class, nor between bags of different classes. However, the mean of the dissimilarities calculated between positive-positive bags (for d_H, $d_{\overline{\min}}$ and $d_{\min}$) seems to be slightly smaller, than the dissimilarities between negative-negative and negative-positive bags. For d_{rel}, the dissimilarities between negative-negative bags are located on the left of those between positive-positive and negative-positive bags. On the other hand, Fig. **5-2** shows that applying the projection makes the distributions of the dissimilarities to behave in the above-mentioned intuitive way. Particularly, for $d_{\min}$, the dissimilarities between positive-positive bags are smaller, than those between negative-negative and negative-positive bags. For d_{rel}, the distances between bags of the same class are clearly lower, than the ones between bags of different classes. It coincides with the outstanding results for this dataset reported in Table **5-2**.



(a) d_H

(b) $d_{\overline{\min}}$

(c) $d_{\min}$

(d) d_{rel}

Figure 5-2.Normalized histograms of distances in the projected space (Musk1 dataset).

5.4. Discussion

We introduce the relative minimum distance for positive bags, denoted by d_{rel}, that globally compares a query bag to a prototype bag. This bag-to-bag distance is appropriate to build low-dimensional dissimilarity-based representations. Furthermore, we benefit from instance-level information, by previously computing a supervised projection, using MildML. This projection estimates a space, where the most likely positive instances are concentrated in the same region and pushes apart the most likely negative instances. Therefore, it allows inducing the desired behavior of the dissimilarities, for the 1-NN rule, i.e., the distance between bags of the same class should be small, and the distance between bags of different classes should be high. Moreover, the proposed algorithm is fully automated, so the tuning of all the parameters is included in the framework. In our experiments, we classify datasets of molecule activity, images, and audio. Our results show that applying d_{rel} in the projected space improves the classification performance in most of the studied datasets, in comparison with conventional distance measures between bags applied in either, the original or the projected space. Our algorithm could fail if the projected space is not estimated as expected, e.g., if an insufficient number of iterations for MildML are computed or if increasing the distance between negative instances is not possible, because it is zero for many of them.

Part II.

Dictionary Learning

6. Dictionary learning for bioacoustic monitoring with applications to species classification

Abstract

This chapter deals with the application of the convolutive version of dictionary learning (DL) to analyze in-situ audio recordings for bioacoustic monitoring. We propose an efficient learning approach that uses a sparse convolutive model to represent a collection of spectrograms. In this approach, we identify repeated bioacoustic patterns, e.g., bird syllables, as time-frequency patterns or words, and represent new spectrograms using these words, and their corresponding activation signals. Moreover, we propose a supervised DL approach in the multiple-label setting to support a multi-label classification of unlabeled spectrograms. Our approach relies on a random projection for reduced computational complexity. As a consequence, the non-negativity requirement on the dictionary words and activations is relaxed. Furthermore, the proposed approach is well-suited for a collection of discontinuous spectrograms. We evaluate our approach on synthetic examples and two real-world datasets consisting of multiple birds audio recordings. Additionally, we successfully apply our DL approach to spectrogram denoising and species classification.

6.1. Introduction[1]

One of the challenges in bioacoustics is to extract a robust representation to animal vocalizations from noisy recordings [105]. In this chapter, we are interested in analyzing a collection

[1] This chapter was published as: José Francisco Ruiz-Muñoz, Zeyu You, Raviv Raich, and Xiaoli Z. Fern. Dictionary Learning for Bioacoustics Monitoring with Applications to Species Classification in Journal of Signal Processing Systems, Springer US, 2016, pp. 1–15.

of audio recordings of bioacoustic data. In particular, our focus lies in learning vocalization models for bird species.

Machine learning techniques are usually applied in order to ease the analysis of a large collection of data. Specifically, dictionary learning (DL) is commonly used to obtain a concise mathematical representation to data for further processing. Previously, dictionary learning have been proposed for analyzing speech and music signals, but only a few attempts have been made for bioacoustic applications.

The DL approach has been used for searching time-varying patterns of audio signals [7, 74], e.g., to detect basic acoustic units as phonemes in speech recognition [83]. In [82], convolutive non-negative matrix factorization (CNMF) [118] is proposed as a DL method. In CNMF approach, a signal is represented by a set of atoms and their associated sparse activation patterns [60]. One of the advantages of CNMF is the simplicity of factor dependencies [11] because each recording is recovered by a linear combination of shifted dictionary words.

In this chapter, we propose a DL approach —based on CNMF— particularly focused on bioacoustic signal processing. Although CNMF has been already applied for analyzing time-series signals, a few challenges arise when applied to the bioacoustic setting: (i) the high computational requirement of CNMF [114] makes it difficult to be applied to large amounts of bioacoustic signals [105]; (ii) CNMF is typically used in a single spectrogram setting, where bioacoustic signals usually contain a collection of discontinuous recordings; and (iii) it is often assumed that the length of the activation signal is the same as the length of the spectrogram in the time domain but it is possible that recordings register only part of a vocalization at the beginning or the end. In this case, a longer activation signal should allow for representing syllable parts in the beginning and the end of the spectrogram.

In this study, we adapt CNMF for a collection of potentially discontinuous spectrograms in which vocalizations may occur prior to the beginning of the recording such that only part of them is observed. The proposed modification is designed to better suit the convolutive DL approach to bioacoustic audio recordings. To illustrate the merit in this approach, we compare our approach against a standard CNMF approach. To address challenges with computational complexity, we propose a randomly projected dictionary learning approach. Additionally, we describe a framework for classifying birdsong recordings based on feature extracted from the sparse representation of spectrograms. In experiments, we present an application of the proposed approach for (i) denoising spectrograms, which are corrupted by rain noise, (ii) unsupervised bird syllable discovery, and (iii) supervised classification of birdsong recordings.

This chapter is organized as follows. Section 6.2 reviews a convolutive DL model using

CNMF. Section 6.3 presents a random matrix projection approach. Section 6.4 develops a two-step update equations for DL and activation signal extraction and analyzes the proposed algorithm for computational complexity. Section 6.5 describes a DL based approach for classifying spectrograms. Finally, Section 6.6 evaluates the randomly projected approach and the classification framework.

6.2. Background and problem formulation

In this section, we first review previous work on the application of CNMF to speech or audio analysis before we introduce our approach. Then, we present the mathematical formulation of the DL approach considered in this chapter.

6.2.1. Background on convolutive NMF

In CNMF [82, 83], a series of non-negative matrices $\boldsymbol{W}_w \in \mathbb{R}^{F\times K}$ $(w = 1, \dots, W)$ and a non-negative matrix $\boldsymbol{H} \in \mathbb{R}^{K\times T}$ are used to approximate a matrix $\boldsymbol{V} \in \mathbb{R}^{F\times T}$ in a convolutive way. Based on this model, an observed spectrogram $\boldsymbol{V}$ can be written as:

$$\boldsymbol{V} \approx \sum_{w=1}^{W} \boldsymbol{W}_w \overset{w\rightarrow}{\boldsymbol{H}}, \tag{6-1}$$

where $\overset{w\rightarrow}{\boldsymbol{H}}$ is the matrix $\boldsymbol{H}$ shifted w columns to the right with the leftmost columns zero filled, W is the length of each word, and each entry of $\boldsymbol{V}$ is

$$\boldsymbol{V}(f,t) = \sum_{w=1}^{W} \sum_{k=1}^{K} \boldsymbol{W}_w(f,k) \overset{w\rightarrow}{\boldsymbol{H}}(k,t).$$

Since this approach requires positive factorized matrices, the Kullback-Leibler (KL) divergence is used in [70] for solving this decomposition problem. In [83], the solution with sparseness constraint is achieved by repeatedly alternating between updating $\boldsymbol{W}_w$ and $\boldsymbol{H}$ as:

$$\begin{aligned} \boldsymbol{H} &= \boldsymbol{H} \otimes \frac{\boldsymbol{W}_w^{\top} \overset{w\rightarrow}{[\frac{\boldsymbol{V}}{\boldsymbol{\Lambda}}]}}{\boldsymbol{W}_w^{\top} \boldsymbol{\Xi} + \lambda \boldsymbol{\Xi}}, \text{ and,} && (6\text{-}2) \\ \boldsymbol{W}_w &= \boldsymbol{W}_w + \gamma_{\mathrm{KL}} \big[\frac{\boldsymbol{V}}{\boldsymbol{\Lambda}} \overset{w\rightarrow}{\boldsymbol{H}}^{\top} - \boldsymbol{\Xi} \overset{w\rightarrow}{\boldsymbol{H}}^{\top} \big], && (6\text{-}3) \end{aligned}$$

where $\boldsymbol{\Lambda} = \sum_{w=1}^{W} \boldsymbol{W}_w \overset{w\rightarrow}{\boldsymbol{H}}$, γ_{KL} has to be small enough to reduce the cost function determined by the Kullback-Leibler divergence, $\otimes$ is the element-wise product, the division is also element-wise and $\boldsymbol{\Xi}$ is an $F \times T$ matrix with all its entries equal to 1 [114].

Two limitations of this approach with respect to bioacoustics applications are: 1) the non-negativity assumptions on both factorized matrices may be invalid, especially when we consider a random projection approach, and 2) the activation signal may occur before the time of the first observation, which requires the length of the activations to be greater than the length of the observation signals. To address these issues, we present a convolutive DL model for random projected spectrograms.

6.2.2. Notation

In this part, we use lower case to denote indexes (e.g. $1 \leqslant m \leqslant M$), upper case to denote fixed scalars (e.g., $M \in \mathbb{N}$), boldfaced lower case to denote vectors (e.g., $\boldsymbol{v} \in \mathbb{R}^{M\times 1}$), $\overleftarrow{\boldsymbol{v}}$ to denote the reversal of $\boldsymbol{v}$ (i.e., $\overleftarrow{\boldsymbol{v}}(m) = \boldsymbol{v}(M - m + 1)$ where $\overleftarrow{\boldsymbol{v}}(m)$ is the m-th entry of $\overleftarrow{\boldsymbol{v}} \in \mathbb{R}^{M\times 1}$), boldfaced upper case to denote matrices (e.g., $\boldsymbol{V} \in \mathbb{R}^{M\times N}$), and calligraphic font to denote sets (e.g. the set of matrices $\mathcal{V} = \{\boldsymbol{V}^{(1)}, \ldots, \boldsymbol{V}^{(L)}\}$ where $\boldsymbol{V}^{(l)} \in \mathbb{R}^{M\times N}$ for $1 \leqslant l \leqslant L$).

We consider two types of discrete convolution operations, denoted by $\star$ (regular convolution) and $\ast$ (full convolution). Taking two vectors $\boldsymbol{u} \in \mathbb{R}^{N\times 1}$ and $\boldsymbol{v} \in \mathbb{R}^{M\times 1}$ where $N \geqslant M$, regular convolution $(\boldsymbol{u} \star \boldsymbol{v})$ returns a third vector $\boldsymbol{z} \in \mathbb{R}^{(N-M+1)\times 1}$ following

$$\boldsymbol{z}(t) = (\boldsymbol{u} \star \boldsymbol{v})(t) = \sum_{m=1}^{M} \boldsymbol{u}(t - m + M)\boldsymbol{v}(m)$$

for $1 \leqslant t \leqslant N - M + 1$. On the other hand, full convolution $(\boldsymbol{u} \ast \boldsymbol{v})$ returns a third vector $\boldsymbol{z} \in \mathbb{R}^{(N+M-1)\times 1}$, which is greater than the one returned by regular convolution. Its entries are computed as follows

$$\boldsymbol{z}(t) = (\boldsymbol{u} \ast \boldsymbol{v})(t) = \sum_{m=1}^{M} \boldsymbol{u}(t - m + 1)\boldsymbol{v}(m)$$

for $1 \leqslant t \leqslant N + M - 1$. Notice that $\boldsymbol{u}(t)$ is considered for $-M \leqslant t \leqslant N + M - 1$, and $\boldsymbol{v}(t)$ for $1 \leqslant t \leqslant M$. Those unavailable values, $\boldsymbol{u}(t)$ for $t < 1$ and $N < t$, are assigned as zeros.

Furthermore, regular convolution can be expressed as a matrix multiplication operation by $\boldsymbol{u} \star \boldsymbol{v} = T_{\boldsymbol{u}}\boldsymbol{v}$ or $\boldsymbol{u} \star \boldsymbol{v} = T_{\boldsymbol{v}}\boldsymbol{u}$ where $T_{\boldsymbol{u}} = \textsc{toeplitz}(\boldsymbol{u}, M, N, M) \in \mathbb{R}^{(N-M+1)\times M}$, and $T_{\boldsymbol{v}} = \textsc{toeplitz}(\boldsymbol{v}, M, N, N) \in \mathbb{R}^{(N-M+1)\times N}$ (Algorithm 2 explains the construction of a Toeplitz matrix). Similarly, full convolution can be expressed as a matrix multiplication operation by $\boldsymbol{u} \ast \boldsymbol{v} = \overset{*}{T}_{\boldsymbol{u}}\boldsymbol{v}$ or $\boldsymbol{u} \ast \boldsymbol{v} = \overset{*}{T}_{\boldsymbol{v}}\boldsymbol{u}$ where $\overset{*}{T}_{\boldsymbol{u}} = \textsc{toeplitz}(\boldsymbol{u}, 1, N+M-1, M) \in \mathbb{R}^{(N+M-1)\times M}$, and $\overset{*}{T}_{\boldsymbol{v}} = \textsc{toeplitz}(\boldsymbol{v}, 1, N+M-1, N) \in \mathbb{R}^{(N+M-1)\times N}$.

Algorithm 2 Toeplitz matrix

```
$\boldsymbol{x} \in \mathbb{R}^{m\times 1}$
$\tau, R, C \in \mathbb{N}$
function TOEPLITZ($\boldsymbol{x}, \tau, R, C$)
    $$\boldsymbol{T} \leftarrow \begin{bmatrix} \phi(\boldsymbol{x},\tau) & \phi(\boldsymbol{x},\tau-1) & \cdots & \phi(\boldsymbol{x},\tau-C+1) \\ \phi(\boldsymbol{x},\tau+1) & \phi(\boldsymbol{x},\tau) & \cdots & \phi(\boldsymbol{x},\tau-C+2) \\ \vdots & \cdots & \cdots & \vdots \\ \phi(\boldsymbol{x},R) & \phi(\boldsymbol{x},R-1) & \cdots & \phi(\boldsymbol{x},R-C+1) \end{bmatrix}$$
    return $\boldsymbol{T} \in \mathbb{R}^{(R-\tau+1)\times C}$
end function
function $\phi(\boldsymbol{x}, t)$
    $$\tilde{x} \leftarrow \begin{cases} \boldsymbol{x}(t), & \text{if } 1 \leqslant t \leqslant \textsc{dim}(\boldsymbol{x}) \\ 0, & \text{otherwise} \end{cases}$$
    return $\tilde{x}$
end function
```

6.2.3. Problem formulation

We assume that the spectrogram $\boldsymbol{Y} \in \mathbb{R}^{F\times T}$ is composed of a sequence of successive time-frequency units called dictionary words that are activated at certain time instants (see Fig. **6-1**). This approximation is expressed by the discrete convolution operation as follows

$$\boldsymbol{y}_f \approx \sum_{k=1}^{K} \boldsymbol{a}_k \star \boldsymbol{d}_{kf} \tag{6-4}$$

where $\boldsymbol{y}_f \in \mathbb{R}^{T\times 1}$ is a transpose row of $\boldsymbol{Y} = [\boldsymbol{y}_1 \ldots \boldsymbol{y}_F]^\top$, corresponding to the f-th frequency-band, $\boldsymbol{a}_k \in \mathbb{R}^{L\times 1}$ is a column of the matrix of activations $\boldsymbol{A} = [\boldsymbol{a}_1 \ldots \boldsymbol{a}_K] \in \mathbb{R}^{L\times K}$, $\boldsymbol{d}_{kf} \in \mathbb{R}^{W\times 1}$ is the time pattern at frequency f of the time-frequency pattern $\boldsymbol{D}_k = [\boldsymbol{d}_{k1} \ldots \boldsymbol{d}_{kF}]^\top \in \mathbb{R}^{F\times W}$, and $\boldsymbol{D} \in \mathbb{R}^{K\times F\times W}$ is the full dictionary —which is built by stacking all $\boldsymbol{D}_k$. According to this decomposition, K is the number of time-frequency patterns, W is the length of each time-frequency pattern, and $L = T + W - 1$ is the length of each activation signal.

As explained at the beginning of this section, the convolution in (6-4) is equivalent to a matrix multiplication as follows

$$\sum_{k=1}^{K} \boldsymbol{a}_k \star \boldsymbol{d}_{kf} = \sum_{k=1}^{K} \boldsymbol{T}_{\boldsymbol{a}_k} \boldsymbol{d}_{kf} \tag{6-5}$$

where $\boldsymbol{T}_{\boldsymbol{a}_k} = \text{TOEPLITZ}(\boldsymbol{a}_k, W, L, W) \in \mathbb{R}^{T \times W}$; this is the expression that we use to update the dictionary. Likewise, (6-4) can be expressed as follows

$$\sum_{k=1}^{K} \boldsymbol{a}_k \star \boldsymbol{d}_{kf} = \sum_{k=1}^{K} \boldsymbol{T}_{\boldsymbol{d}_{kf}} \boldsymbol{a}_k \tag{6-6}$$

where $\boldsymbol{T}_{\boldsymbol{d}_{kf}} = \text{TOEPLITZ}(\boldsymbol{d}_{kf}, W, L, L) \in \mathbb{R}^{T \times L}$; this is the expression that we use to update the activations.

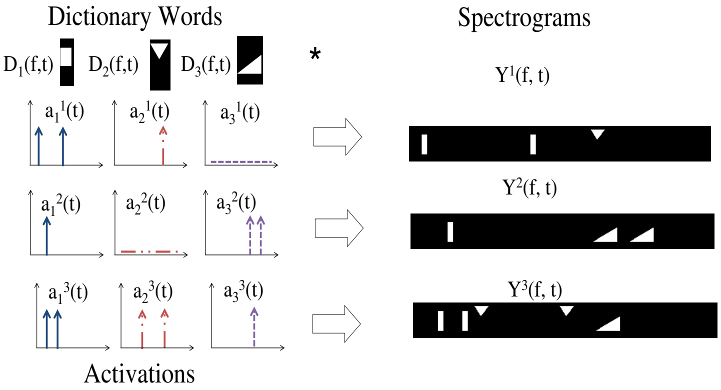


Figure 6-1. Reproduction of a convolutive model for dictionary learning [97]. This illustration shows how the elements $Y^i(f,t)$ of a spectrogram are computed by applying the convolution operation between the elements of the dictionary words $d_1(t,f)$, $d_2(t,f)$ and $d_3(t,f)$, and the activation signals $a_1^i(t)$, $a_2^i(t)$ and $a_3^i(t)$.

To minimize the reconstructed error between a set of N stacked spectrograms $\mathcal{Y} = \{\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(N)}\}$ and their approximated spectrograms, we propose a convolutive dictionary model formulated as the following optimization problem:

$$\begin{gathered} \min_{\boldsymbol{D},\boldsymbol{A}} \ell(\mathcal{Y}, \boldsymbol{D}, \mathcal{A}) \\ \ell(\mathcal{Y}, \boldsymbol{D}, \mathcal{A}) := \frac{1}{2} \sum_{n=1}^{N} \left(\sum_{f=1}^{F} ||\boldsymbol{y}_f^{(n)} - \tilde{\boldsymbol{y}}_f^{(n)}||^2 + 2\lambda \sum_{k=1}^{K} \sum_{t=1}^{L} |\boldsymbol{a}_k^{(n)}(t)| \right) \\ \text{subject to } \sum_{f=1}^{F} \sum_{t=1}^{W} |\boldsymbol{d}_{kf}(t)|^2 \leqslant 1, \forall 1 \leqslant k \leqslant K. \end{gathered} \tag{6-7}$$

where $\mathcal{A} = \{\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(N)}\}$. The constraints are introduced to prevent $\boldsymbol{D}$ from being arbitrary large, which would lead to arbitrary small values of activation signals. Therefore, the proposed convolutive model provides a natural representation for a set of N discontinuous spectrograms of bird vocalizations.

To facilitate the optimization of the objective in (6-7), we use Toeplitz matrices [106] and an iterative approach that alternates between the estimation of the dictionary $\boldsymbol{D}$ and the activations $\mathcal{A}$. Therefore, in the p-th iteration, for the *dictionary update*, $\mathcal{A}$ is held fixed, and the dictionary $\boldsymbol{D}$ is updated following: $\boldsymbol{D}_{(p)} = \arg\min_{\boldsymbol{D}} \ell(\boldsymbol{Y}, \boldsymbol{D}_{(p-1)}, \boldsymbol{A})$. To this end, we approximate each frequency-band of the spectrogram by $\tilde{\boldsymbol{y}}_f^{(n)} = \sum_{k=1}^{K} \boldsymbol{T}_{\boldsymbol{a}_k}^{(n)} \boldsymbol{d}_{kf}$. Similarly, to update $\mathcal{A}_{(p)} = \arg\min_{\mathcal{A}} \ell(\boldsymbol{Y}, \boldsymbol{D}, \mathcal{A}_{(p-1)})$, we approximate each frequency-band of the spectrogram by $\tilde{\boldsymbol{y}}_f^{(n)} = \sum_{k=1}^{K} \boldsymbol{T}_{\boldsymbol{d}_{kf}} \boldsymbol{a}_k^{(n)}$.

Constructing Toeplitz matrices is memory inefficient and solving the above alternating quadratic programming problem with matrix inversion is time-consuming. To reduce the computational complexity and the memory issue of the convolutive model, in this chapter, we propose a random projected convolutive model with modified gradient descent algorithm that utilizes the convolution operator.

6.3. Random projected dictionary learning

In order to facilitate the reduction in computational complexity, we consider a compressive sampling approach. Since bird vocalizations are mostly concentrated in a small range of frequencies, spectrograms of bird vocalization tend to have sparse columns. Therefore, we apply a random matrix transformation (with less rows than the number of frequency bins) to both the spectrogram width and dictionary word width. In such a way, the computational complexity of obtaining dictionary words and activations is decreased by reducing the size of the spectrograms and hence, the number of unknowns (see Sec. 6.4.3). Accordingly, the proposed new formulation of the dictionary learning is

$$\begin{gathered}\min_{\boldsymbol{D},\boldsymbol{A}} \ell(\mathcal{Y}^Q, \boldsymbol{D}^Q, \mathcal{A}) \\ \ell(\mathcal{Y}^Q, \boldsymbol{D}^Q, \mathcal{A}) := \frac{1}{2}\sum_{n=1}^{N}\left(\sum_{r=1}^{R} ||\boldsymbol{y}_r^{Q(n)} - \tilde{\boldsymbol{y}}_r^{Q(n)}||^2 + 2\lambda \sum_{k=1}^{K}\sum_{t=1}^{L} |\boldsymbol{a}_k^{(n)}(t)|\right) \\ \text{subject to } \sum_{r=1}^{R}\sum_{t=1}^{W} |\boldsymbol{d}_{kr}^Q(t)|^2 \leqslant 1, \forall 1 \leqslant k \leqslant K.\end{gathered} \tag{6-8}$$

where $\boldsymbol{Q} \in \mathbb{R}^{R\times F}$ is a transformation matrix, $\mathcal{Y}^Q = \{\boldsymbol{Y}^{Q(1)}, \ldots, \boldsymbol{Y}^{Q(N)}\}$ is a set of transformed spectrograms such that $\boldsymbol{Y}^{Q(n)} = \boldsymbol{Q}\boldsymbol{Y}^{(n)} \in \mathbb{R}^{R\times T}$, $\boldsymbol{y}_r^Q \in \mathbb{R}^{T\times 1}$ is a transpose row of $\boldsymbol{Y}^Q = [\boldsymbol{y}_1^Q \ldots \boldsymbol{y}_R^Q]^\top$, and $\boldsymbol{D}_k^Q = \boldsymbol{Q}\boldsymbol{D}_k$ is the k-th time-frequency pattern transformed, such that $\boldsymbol{d}_{kr}^{Q} \in \mathbb{R}^{W\times 1}$ is the time pattern at frequency r of the time-frequency pattern $\boldsymbol{D}_k = [\boldsymbol{d}_{k1}^{Q} \dots \boldsymbol{d}_{kR}^{Q}]^\top \in \mathbb{R}^{R\times W}$.

If the intensities at several frequency bins are compressed into a single coefficient using Mel-Frequency-Coefficients (MFC), it is difficult to recover their value from the single coefficient but it might be prevented by applying a compressive transformation with a random matrix [115]. By relying on the sparsity of the signal and the compressive approach, the recovery of the original spectrograms or dictionary words can be implemented using a linear programming approach [8, 93].

6.4. Solution approach for dictionary learning and extraction of activations

Considering the DL problem in (6-7), we analyze two cases:

$$\begin{gathered}\min_{\boldsymbol{D}} \ell_{\boldsymbol{D}}(\mathcal{Y}, \boldsymbol{D}, \mathcal{A}) \\ \ell_{\boldsymbol{D}}(\mathcal{Y}, \boldsymbol{D}, \mathcal{A}) := \frac{1}{2}\sum_{n=1}^{N}\sum_{f=1}^{F} ||\boldsymbol{y}_f^{(n)} - \sum_{k=1}^{K} \boldsymbol{T}_{\boldsymbol{a}_k}^{(n)} \boldsymbol{d}_{kf}||^2 \\ \text{subject to } \sum_{f=1}^{F}\sum_{t=1}^{W} |\boldsymbol{d}_{kf}(t)|^2 \leqslant 1, \forall 1 \leqslant k \leqslant K,\end{gathered} \tag{6-9}$$

and

$$\begin{gathered}\min_{\boldsymbol{A}} \ell_{\mathcal{A}}(\mathcal{Y}, \boldsymbol{D}, \mathcal{A}) \\ \ell_{\mathcal{A}}(\mathcal{Y}, \boldsymbol{D}, \mathcal{A}) := \frac{1}{2}\sum_{n=1}^{N}\left(\sum_{f=1}^{F} ||\boldsymbol{y}_f^{(n)} - \sum_{k=1}^{K} \boldsymbol{T}_{\boldsymbol{d}_{kf}} \boldsymbol{a}_k^{(n)}||^2 + 2\lambda \sum_{k=1}^{K}\sum_{t=1}^{L} |\boldsymbol{a}_k^{(n)}(t)|\right).\end{gathered} \tag{6-10}$$

Several solution approaches have been established for solving both of the problems written above, (6-9) and (6-10). For (6-9), one of the current state-of-art methods is least square solution with normalization on dictionary words and projected Newton descent method. For (6-10), the current state-of-art method is Least Angle Regression (LARS) algorithm [43] or feature-sign sparse coding algorithm [71]. However, these algorithms require a large matrix inversion to obtain an efficient and exact solution. To solve our proposed problem directly, computing $\boldsymbol{T}_{\boldsymbol{a}_k}^{(n)\top} \boldsymbol{T}_{\boldsymbol{a}_k}^{(n)}$ and $\boldsymbol{T}_{\boldsymbol{d}_{kf}}^{\top} \boldsymbol{T}_{\boldsymbol{d}_{kf}}$ require a computational complexity of the order $\mathcal{O}(NKT\log T)$ and $\mathcal{O}(FKT\log T)$, respectively, and computing their inverse requires $\mathcal{O}((KW)^3)$ and $\mathcal{O}((KL)^3)$, respectively. It strongly limits the practical applicability of the approach. Hence, we follow the work in [120] on majorization-minimization [59] for DL in a non-convolutive case and consider a similar optimization transfer approach to minimize both (6-9) and (6-10).

In majorization-minimization, a surrogate function $g(x, x')$ satisfying i) $f(x) \leqslant g(x, x'),\ \forall x, x'$ and ii) $f(x') = g(x', x'),\ \forall x'$ is considered as a replacement to the original objective $f(x)$. The update iteration $x^{(j+1)} = \arg\min_x g(x, x')$ guarantees $f(x^{(j+1)}) \leqslant f(x^{(j)})$.

6.4.1. Dictionary learning

Since we consider an optimization transfer approach —majorization-minimization [59]— to facilitate an iterative minimization of the objective in (6-9), we need to identify an efficient surrogate function. For this purpose, we use the following inequality

$$\begin{aligned}
\frac{1}{2}\sum_{n=1}^{N}\|\boldsymbol{y}_f^{(n)} - \sum_{k=1}^{K}\boldsymbol{T}_{\boldsymbol{a}_k}^{(n)}\boldsymbol{d}_{kf}\|^2 &= \frac{1}{2}\sum_{n=1}^{N}\|\sum_{k=1}^{K}\boldsymbol{T}_{\boldsymbol{a}_k}^{(n)}(\boldsymbol{d}_{kf}-\boldsymbol{d}'_{kf}) - (\boldsymbol{y}_f^{(n)} - \sum_{k=1}^{K}\boldsymbol{T}_{\boldsymbol{a}_k}^{(n)}\boldsymbol{d}'_{kf})\|^2 \\
&\leqslant \frac{\gamma_f}{2}\sum_{k=1}^{K}\|\boldsymbol{d}_{kf}-\boldsymbol{d}'_{kf}\|^2 - \sum_{n=1}^{N}\sum_{k=1}^{K}\boldsymbol{d}_{kf}^{\top}\boldsymbol{T}_{\boldsymbol{a}_k}^{(n)\top}(\boldsymbol{y}_f^{(n)} - \sum_{k=1}^{K}\boldsymbol{T}_{\boldsymbol{a}_k}^{(n)}\boldsymbol{d}'_{kf}) + \text{const.} \qquad \text{(6-11)} \\
&= \frac{\gamma_f}{2}\sum_{k=1}^{K}\|\boldsymbol{d}_{kf} - (\boldsymbol{d}'_{kf} + \frac{1}{\gamma_f}\sum_{n=1}^{N}\boldsymbol{T}_{\boldsymbol{a}_k}^{(n)\top}(\boldsymbol{y}_f^{(n)} - \sum_{k=1}^{K}\boldsymbol{T}_{\boldsymbol{a}_k}^{(n)}\boldsymbol{d}'_{kf}))\|^2 + \text{const.},
\end{aligned}$$

which provides a surrogate to

$$\frac{1}{2}\sum_{n=1}^{N}\|\boldsymbol{y}_f^{(n)} - \sum_{k=1}^{K}\boldsymbol{T}_{\boldsymbol{a}_k}^{(n)}\boldsymbol{d}_{kf}\|^2.$$

To satisfy the inequality, we choose

$$\gamma_d = \max_f \gamma_f \geqslant \max_f \sum_{k=1}^{K}\frac{\|\sum_{n=1}^{N}\boldsymbol{T}_{\boldsymbol{a}_k}^{(n)}(\boldsymbol{d}_{kf}-\boldsymbol{d}'_{kf})\|^2}{\|\boldsymbol{d}_{kf}-\boldsymbol{d}'_{kf}\|^2}.$$

Replacing the objective using the surrogate in (6-11) and minimizing with respect to the dictionary yields

$$\begin{aligned}
&\min_{\boldsymbol{d}_{kf}} \sum_{f=1}^{F}\frac{\gamma_d}{2}\|\boldsymbol{d}_{kf} - \boldsymbol{g}_{kf}\|^2 \\
&\text{subject to } \sum_{f=1}^{F}\|\boldsymbol{d}_{kf}\|^2 \leqslant 1, \forall 1 \leqslant k \leqslant K,
\end{aligned} \qquad \text{(6-12)}$$

where

$$\boldsymbol{g}_{kf} = \boldsymbol{d}'_{kf} + \frac{1}{\gamma_d}\sum_{n=1}^{N}\boldsymbol{v}^{(n)}_{\boldsymbol{d}_{kf}} \tag{6-13}$$

and

$$\boldsymbol{v}^{(n)}_{\boldsymbol{d}_{kf}} = \boldsymbol{T}^{(n)\top}_{\boldsymbol{a}_k}(\boldsymbol{y}^{(n)}_f - \sum_{k=1}^{K}\boldsymbol{T}^{(n)}_{\boldsymbol{a}_k}\boldsymbol{d}'_{kf}).$$

To solve (6-12), we form the Lagrangian

$$L(\boldsymbol{D},\beta) = \sum_{k=1}^{K}\sum_{f=1}^{F}\frac{\gamma_d}{2}\|\boldsymbol{d}_{kf} - \boldsymbol{g}_{kf}\|^2 + \sum_{k=1}^{K}\beta_k(\sum_{f=1}^{F}\|\boldsymbol{d}_{kf}\|^2 - 1). \tag{6-14}$$

Minimizing the Lagrangian with respect to $\boldsymbol{d}_{kf}$ results in

$$\boldsymbol{d}_{kf} = \frac{\gamma_d}{\gamma_d + 2\beta_k}\boldsymbol{g}_{kf}. \tag{6-15}$$

Substituting $\boldsymbol{d}_{kf}$ back into (6-14) yields the dual function

$$G(\beta_k) = \sum_{k=1}^{K}\big[\beta_k(\frac{\gamma_d}{\gamma_d + 2\beta_k}\sum_{f=1}^{F}\|\boldsymbol{g}_{kf}\|^2 - 1)\big].$$

Maximizing the dual objective with respect to β_k subject to $\beta_k \geqslant 0$ yields

$$\operatorname*{argmin}_{\beta_k} G(\beta_k) = \begin{cases} 0, & \text{if } \sum_{f=1}^{F}\|\boldsymbol{g}_{kf}\|^2 \leqslant 1; \\ \frac{\gamma_d}{2}(\sqrt{\sum_{f=1}^{F}\|\boldsymbol{g}_{kf}\|^2} - 1), & \text{otherwise.} \end{cases} \tag{6-16}$$

To compute the optimal $\boldsymbol{d}_{kf}$, we replace (6-16) back into (6-15) and obtain

$$\boldsymbol{d}_{kf} = \begin{cases} \boldsymbol{g}_{kf}, & \text{if } \sqrt{\sum_{f=1}^{F}\|\boldsymbol{g}_{kf}\|^2} \leqslant 1; \\ \frac{\boldsymbol{g}_{kf}}{\sqrt{\sum_{f=1}^{F}\|\boldsymbol{g}_{kf}\|^2}}, & \text{otherwise.} \end{cases} \tag{6-17}$$

Finally, replacing (6-13) into (6-17) yields

$$\boldsymbol{d}^{(j+1)}_{kf} = \begin{cases} \boldsymbol{d}^{(j)}_{kf} + \frac{1}{\gamma_d}\sum_{n=1}^{N}\boldsymbol{v}^{(n)}_{\boldsymbol{d}_{kf}}, & \text{if } \sqrt{\sum_{f=1}^{F}\|\boldsymbol{d}^{(j)}_{kf} + \frac{1}{\gamma_d}\sum_{n=1}^{N}\boldsymbol{v}^{(n)}_{\boldsymbol{d}_{kf}}\|^2} \leqslant 1; \\ \frac{\boldsymbol{d}^{(j)}_{kf} + \frac{1}{\gamma_d}\sum_{n=1}^{N}\boldsymbol{v}^{(n)}_{\boldsymbol{d}_{kf}}}{\sqrt{\sum_{f=1}^{F}\|\boldsymbol{d}^{(j)}_{kf} + \frac{1}{\gamma_d}\sum_{n=1}^{N}\boldsymbol{v}^{(n)}_{\boldsymbol{d}_{kf}}\|^2}}, & \text{otherwise.} \end{cases} \tag{6-18}$$

Step-size selection for the DL update

To determine the step size γ_d, we consider two cases: i) when the updated dictionary words satisfies the constraint that

$$\sum_{f=1}^{F} \|\boldsymbol{g}_k^f\|^2 \leqslant 1,$$

the optimal step-size is

$$\gamma_d = \frac{\|\sum\limits_{n=1}^{N} \boldsymbol{T}_{\boldsymbol{a}_k^{(n)}} \boldsymbol{v}_{\boldsymbol{d}_{kf}^{(n)}}\|^2}{\|\sum\limits_{n=1}^{N} \boldsymbol{v}_{\boldsymbol{d}_{kf}^{(n)}}\|^2};$$

and, ii) when the constraints are not satisfied, the optimal step-size has no closed-form solution. Setting

$$\gamma_d = \lambda_{\max}(\sum_{n=1}^{N} {\boldsymbol{T}_{\boldsymbol{a}_k^{(n)}}}^\top \boldsymbol{T}_{\boldsymbol{a}_k^{(n)}})$$

ensures that

$$\|\sum_{n=1}^{N} \boldsymbol{T}_{\boldsymbol{a}_k^{(n)}} \boldsymbol{v}_{\boldsymbol{d}_{kf}^{(n)}}\|^2 \leqslant \gamma_d \|\sum_{n=1}^{N} \boldsymbol{v}_{\boldsymbol{d}_{kf}^{(n)}}\|^2$$

for any $\sum\limits_{n=1}^{N} \boldsymbol{v}_{\boldsymbol{d}_{kf}^{(n)}}$. This conservative approach results in a small step size $1/\gamma_d$, which leads to a slow convergence rate. To improve this, we consider the following tighter bound on γ_d. We rely on maximizing first individual frequency-band and take the maximum over all of them.

From (6-18), we have

$$\boldsymbol{d}_{kf} - \boldsymbol{d}'_{kf} = \frac{\boldsymbol{d}'_{kf} + \frac{1}{\gamma_d} \sum\limits_{n=1}^{N} \boldsymbol{v}_{\boldsymbol{d}_{kf}}^{(n)}}{\sqrt{\sum\limits_{f=1}^{F} \|\boldsymbol{d}'_{kf} + \frac{1}{\gamma_d} \sum\limits_{n=1}^{N} \boldsymbol{v}_{\boldsymbol{d}_{kf}}^{(n)}\|^2}} - \boldsymbol{d}'_{kf} = \alpha_1 \boldsymbol{d}_{kf} + \alpha_2 \sum_{n=1}^{N} \boldsymbol{v}_{\boldsymbol{d}_{kf}}^{(n)}.$$

Since $\boldsymbol{d}_{\cdot f} - \boldsymbol{d}'_{\cdot f} \in \operatorname{span}\{\boldsymbol{d}_{\cdot f}, \boldsymbol{v}_{\boldsymbol{d}_{\cdot f}}^{(\cdot)}\}$ where $\boldsymbol{d}_{\cdot f} = [\boldsymbol{d}_{1f}^\top \dots \boldsymbol{d}_{Kf}^\top]^\top \in \mathbb{R}^{KW \times 1}$, $\boldsymbol{d}'_{\cdot f} = [{\boldsymbol{d}'_{1f}}^\top \dots {\boldsymbol{d}'_{Kf}}^\top]^\top \in \mathbb{R}^{KW \times 1}$, and $\boldsymbol{v}_{\boldsymbol{d}_{\cdot f}}^{(\cdot)} = [\sum\limits_{n=1}^{N} {\boldsymbol{v}_{\boldsymbol{d}_{1f}}^{(n)}}^\top \dots \sum\limits_{n=1}^{N} {\boldsymbol{v}_{\boldsymbol{d}_{Kf}}^{(n)}}^\top] \in \mathbb{R}^{KW \times 1}$, we can further restrict γ_d without violating the bound on γ_d. Using Gram-Schmidt orthogonalization, we obtain the orthogonal basis for $\{\boldsymbol{v}_{\boldsymbol{d}_{\cdot f}}^{(\cdot)}, \boldsymbol{d}'_{\cdot f}\}$ as $\boldsymbol{u}_1 = \boldsymbol{v}_{\boldsymbol{d}_{\cdot f}}^{(\cdot)} / \|\boldsymbol{v}_{\boldsymbol{d}_{\cdot f}}^{(\cdot)}\|$ and $\boldsymbol{u}_2 = \tilde{\boldsymbol{d}}'_{\cdot f} / \|\tilde{\boldsymbol{d}}'_{\cdot f}\|$, where $\tilde{\boldsymbol{d}}'_{\cdot f} = \boldsymbol{d}'_{\cdot f} - ({\boldsymbol{d}'_{\cdot f}}^\top \boldsymbol{u}_1)\boldsymbol{u}_1$. For every value of (α_1, α_2) in the representation of $\boldsymbol{d}_{\cdot f} - \boldsymbol{d}'_{\cdot f} = \alpha_1 \boldsymbol{d}'_{\cdot f} + \alpha_2 \boldsymbol{v}_{\boldsymbol{d}_{\cdot f}}^{(\cdot)}$ there is a (β_1, β_2) in the equivalent representation of $\boldsymbol{d}_{\cdot f} - \boldsymbol{d}'_{\cdot f} = \beta_1 \boldsymbol{u}_1 + \beta_2 \boldsymbol{u}_2$. Hence, we can find γ_f

by maximizing the following with respect to (β_1, β_2):

$$\frac{\|\sum_{n=1}^{N} \boldsymbol{T}_{\boldsymbol{A}}^{(n)}(\boldsymbol{d}_{\cdot f} - \boldsymbol{d}'_{\cdot f})\|^2}{\|\boldsymbol{d}_{\cdot f} - \boldsymbol{d}'_{\cdot f}\|^2} = \frac{\|\sum_{n=1}^{N} \boldsymbol{T}_{\boldsymbol{A}}^{(n)}[\boldsymbol{u}_1, \boldsymbol{u}_2][\beta_1, \beta_2]^\top\|^2}{\|[\beta_1, \beta_2]^\top\|^2} \tag{6-19}$$

where $\boldsymbol{T}_{\boldsymbol{A}}^{(n)} = [\boldsymbol{T}_{\boldsymbol{a}_1^{(n)}} \ldots \boldsymbol{T}_{\boldsymbol{a}_K^{(n)}}] \in \mathbb{R}^{T \times KW}$.

Accordingly, we can bound (6-19) by

$$\gamma_f = \lambda_{\max}([\boldsymbol{u}_1, \boldsymbol{u}_2]^\top (\sum_{n=1}^{N} \boldsymbol{T}_{\boldsymbol{A}}^{(n)} \boldsymbol{T}_{\boldsymbol{A}}^{(n)\top})[\boldsymbol{u}_1, \boldsymbol{u}_2]).$$

Notice that although $\sum_{n=1}^{N} \boldsymbol{T}_{\boldsymbol{A}}^{(n)} \boldsymbol{T}_{\boldsymbol{A}}^{(n)\top}$ is independent of the frequency f, it is fairly large and its associated eigendecomposition may be computationally intensive. Instead, we replace it with the eigendecomposition of F 2×2 frequency-dependent matrices $[\boldsymbol{u}_1, \boldsymbol{u}_2]^\top (\sum_{n=1}^{N} \boldsymbol{T}_{\boldsymbol{A}}^{(n)} \boldsymbol{T}_{\boldsymbol{A}}^{(n)\top})[\boldsymbol{u}_1, \boldsymbol{u}_2]$. To ensure that the bound holds for every f, we select the step size $\gamma_d = \max_f \gamma_f$.

6.4.2. Extraction of activations

Similarly to DL, we consider an optimization transfer approach to facilitate an iterative rule to the minimization of the objective in (6-10) with respect to the activations. Therefore, a bounding technique yields

$$\begin{gathered}\frac{1}{2}\sum_{f=1}^{F} ||\boldsymbol{y}_f^{(n)} - \sum_{k=1}^{K} \boldsymbol{T}_{\boldsymbol{d}_{kf}} \boldsymbol{a}_k^{(n)}||^2 \leqslant \\ \frac{\gamma_a^{(n)}}{2} \sum_{k=1}^{K} \|\boldsymbol{a}_k^{(n)} - (\boldsymbol{a'}_k^{(n)} + \frac{1}{\gamma_a^{(n)}} \sum_{f=1}^{F} \boldsymbol{T}_{\boldsymbol{d}_{kf}}{}^\top (\boldsymbol{y}_f^{(n)} - \sum_{k=1}^{K} \boldsymbol{T}_{\boldsymbol{d}_{kf}} \boldsymbol{a}_k^{(n)}))\|^2 + \text{const.},\end{gathered} \tag{6-20}$$

such that

$$\gamma_a^{(n)} \geqslant \sum_{k=1}^{K} \frac{\|\sum_{f=1}^{F} \boldsymbol{T}_{\boldsymbol{d}_{kf}}(\boldsymbol{a}_k^{(n)} - \boldsymbol{a'}_k^{(n)})\|^2}{\|\boldsymbol{a}_k^{(n)} - \boldsymbol{a'}_k^{(n)}\|^2}. \tag{6-21}$$

Therefore, the surrogate problem for extracting activations is defined as:

$$\min_{\boldsymbol{a}_k^{(n)}} \frac{1}{2}||\boldsymbol{a}_k^{(n)} - \boldsymbol{h}_k^{(n)}||^2 + \lambda \sum_{k=1}^{K} \sum_{t=1}^{L} |\boldsymbol{a}_k^{(n)}(t)| \tag{6-22}$$

where

$$\boldsymbol{h}_k^{(n)} = \boldsymbol{a'}_k^{(n)} + \frac{1}{\gamma_a^{(n)}}\boldsymbol{v}_{\boldsymbol{a}_k^{(n)}}$$

and

$$\boldsymbol{v}_{\boldsymbol{a}_k^{(n)}} = \sum_{f=1}^{F}\boldsymbol{T}_{\boldsymbol{d}_{kf}}{}^{\top}(\boldsymbol{y}_f^{(n)} - \sum_{k=1}^{K}\boldsymbol{T}_{\boldsymbol{d}_{kf}}\boldsymbol{a}_k^{(n)}).$$

Notice that the objective in (6-22) is separable since

$$\frac{1}{2}||\boldsymbol{a}_k^{(n)} - \boldsymbol{h}_k^{(n)}||^2 + \lambda\sum_{k=1}^{K}\sum_{t=1}^{L}|\boldsymbol{a}_k^{(n)}(t)| = \sum_{t=1}^{L}\left(\frac{1}{2}||\boldsymbol{a}_k^{(n)}(t) - \boldsymbol{h}_k^{(n)}(t)||^2 + \lambda\sum_{k=1}^{K}|\boldsymbol{a}_k^{(n)}(t)|\right)$$

Therefore, the solution to (6-22) can be obtained by solving element-wise for every $\boldsymbol{a}_k^{(n)}(t)$. The resulting updating rule for extracting the activation follows

$$\boldsymbol{a}_k^{(n)}(t)^{(j+1)} = \begin{cases} \boldsymbol{a}_k^{(n)}(t)^{(j)} + \frac{1}{\gamma_a^{(n)}}(\boldsymbol{v}_{\boldsymbol{a}_k^{(n)}}(t) - \lambda), & \text{if } \boldsymbol{a}_k^{(n)}(t)^{(j)} + \frac{1}{\gamma_a^{(n)}}(\boldsymbol{v}_{\boldsymbol{a}_k^{(n)}}(t) - \lambda) > 0 \\ \boldsymbol{a}_k^{(n)}(t)^{(j)} + \frac{1}{\gamma_a^{(n)}}(\boldsymbol{v}_{\boldsymbol{a}_k^{(n)}}(t) + \lambda), & \text{if } \boldsymbol{a}_k^{(n)}(t)^{(j)} + \frac{1}{\gamma_a^{(n)}}(\boldsymbol{v}_{\boldsymbol{a}_k^{(n)}}(t) + \lambda) < 0 \\ 0, & \text{otherwise.} \end{cases} \tag{6-23}$$

Step-size selection for updating activations

Since the optimal step-size for activation updates must satisfy (6-21), we can bound $\gamma_a^{(n)}$ by $\lambda_{\max}(\sum_{f=1}^{F}\boldsymbol{T}_{\boldsymbol{d}_{kf}}^{\top}\boldsymbol{T}_{\boldsymbol{d}_{kf}})$, which is the largest eigenvalue of the matrix $\sum_{f=1}^{F}\boldsymbol{T}_{\boldsymbol{d}_{kf}}^{\top}\boldsymbol{T}_{\boldsymbol{d}_{kf}}$. Computing this eigenvalue is computational expensive. Instead, we use the fast Fourier transform (FFT). According to the Parseval theorem and Cauchy-Schwartz inequality, we derive the following bound

$$\sum_{k=1}^{K}\frac{\|\sum_{f=1}^{F}\boldsymbol{T}_{\boldsymbol{d}_{kf}}(\boldsymbol{a}_k^{(n)} - \boldsymbol{a'}_k^{(n)})\|^2}{\|\boldsymbol{a}_k^{(n)} - \boldsymbol{a'}_k^{(n)}\|^2} \leqslant \sum_{f=1}^{F}\max_{\omega}\sum_{k=1}^{K}|\hat{\boldsymbol{d}}_{kf}(\omega)|^2$$

where $\hat{\boldsymbol{d}}_{kf} \in \mathbb{R}^{W\times 1}$ contains the magnitude values of the discrete Fourier transform of $\boldsymbol{d}_{kf}$. Hence, $\sum_{k=1}^{K}\|\sum_{f=1}^{F}\boldsymbol{T}_{\boldsymbol{d}_{kf}}(\boldsymbol{a}_k^{(n)} - \boldsymbol{a'}_k^{(n)})\|^2/\|\boldsymbol{a}_k^{(n)} - \boldsymbol{a'}_k^{(n)}\|^2$ can be upper bounded by

$$\gamma_a^{(n)} = \sum_{f=1}^{F}\max_{\omega}\sum_{k=1}^{K}|\hat{\boldsymbol{d}}_{kf}(\omega)|^2.$$

6.4.3. Solution approach for random projection model

For the random projection approach, we simply replace $\mathcal{Y}$ with $\mathcal{Y}^Q$ and $\boldsymbol{D}$ with $\boldsymbol{D}^Q$. Thus, we obtain an efficient algorithm for practical dictionary extraction. This algorithm consists of three main parts: (i) transforming the original input spectrograms using a random projection matrix, (ii) alternatively updating dictionary words and activations until a convergence criterion is met, and (iii) recovering the uncompressed domain dictionary words by solving the optimization problem with the extracted activations and the original data $\mathcal{Y}$.

Using the FFT and inverse fast Fourier transform (IFFT), the computational complexity for each convolution block with size L is $\mathcal{O}(L \log L)$. For the iterative procedure in (6-9), calculating $\sum_{k=1}^{K} \boldsymbol{T}_{\boldsymbol{a}_k}^{(n)} \boldsymbol{d}_{kf}$, $\boldsymbol{v}_{\boldsymbol{d}_{kf}}^{(n)}$ and γ_d, requires $\mathcal{O}(NKL \log L)$. Therefore, the overall computational complexity is $\mathcal{O}(FNKL \log L)$. Updating the activations requires the same computational complexity as $\mathcal{O}(NFKL \log L)$. The total computational complexity for the algorithm without random projection is $\mathcal{O}(FNKL \log L)$. With random projection, the computational complexity is $\mathcal{O}(RNKL \log L)$, which is proportional to the original computation complexity. Thus, if the reduced frequency band R is 20% of the original frequency band F, the running time will be five times faster than the uncompressed dictionary learning algorithm. Therefore, using random projection makes the convolutive dictionary learning method more efficient and practical.

6.5. Dictionary-based classification framework

Dictionary learning is not limited to spectrogram reconstruction or denoising. Here, we present a dictionary-based classification step that aims to use the learned sparse representation for classifying bioacoustic recordings. The proposed scheme is inspired by the framework used in music analysis [121]. In Fig. **6-2**, we present the classification framework in two parts: training and test. In the first part, the dictionary words and activation signals are estimated from the training set, a set of features is extracted from the activations signals, which is used for training an SVM classifier. In the second part, the activations signals corresponding to the test set are estimated using the dictionary previously learned. Then features are extracted based on the activations. Finally, the features are provided as an input tor the SVM classifier. The supervised dictionary learning adaptation, feature extraction and SVM for training and classification are explained below.

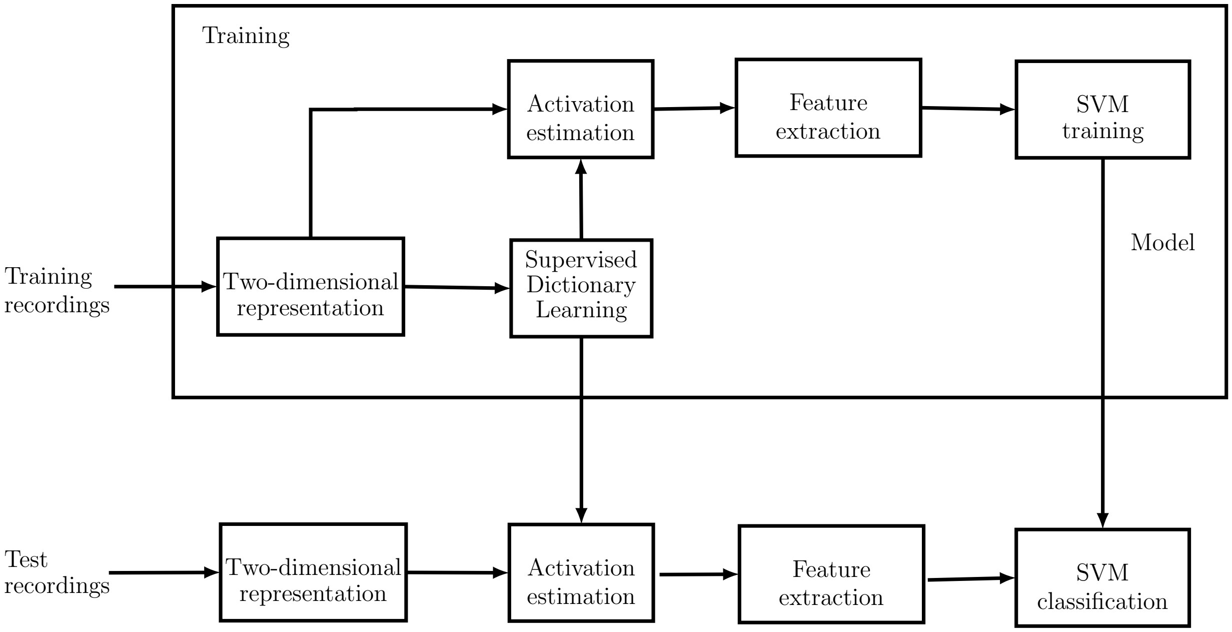


Figure 6-2.Diagram of dictionary-based classification.

6.5.1. Supervised dictionary learning

For classification, we consider the case in which each recording may contain dictionary words from multiple classes. In our application, vocalizations in the same recording may come from from multiple bird species. This classification framework has been considered for species recognition of in-situ recordings [19]. This setting is often referred to as the multiple label setting. In the multiple label setting, spectrogram $\boldsymbol{Y}^{(n)}$ is associated with a label vector $\boldsymbol{l}^{(n)} \in \mathbb{R}^{1\times N_c}$ such that $\boldsymbol{l}^{(n)}(l) \in \{0,1\}$ for $1 \leqslant l \leqslant N_c$, and N_c is the number of classes. Those binary entries indicate the presence (by 1) or the absence (by 0) of the l-th species in the n-th recording. Therefore, the label information can be summarize using the matrix $\boldsymbol{L} = [\boldsymbol{l}^{(1)} \dots \boldsymbol{l}^{(N)}]^\top \in \mathbb{R}^{N\times N_c}$.

To adapt our dictionary learning approach to this setting, we assume that the dictionary consists of N_c sub-dictionaries (one for each class). The l-th sub-dictionary consists of K_l words and the total number of dictionary words is K such that $\sum_{l=1}^{N_c} K_l = K$. We assume that a sub-dictionary can only be used to construct a given spectrogram if it contains its corresponding class. Alternatively, if the class is absent in the n-th spectrogram, the activations associated with its dictionary words are assigned as zeros.

We consider using a summarization of the activations as a feature vector that provide information about the presence or absence of a given class in a recording. To this end, we map the set of activations of the n-th spectrogram $\boldsymbol{A}^{(n)}$ to a vector where its dimension is the number of estimated features. Therefore, we compute a vector $\boldsymbol{x}^{(n)} \in \mathbb{R}^{K\times 1}$ where

$$\boldsymbol{x}^{(n)}(k) = \frac{\sum_{t=1}^{L} |\boldsymbol{a}_k^{(n)}(t)|}{\sum_{k=1}^{K}\sum_{t=1}^{L} |\boldsymbol{a}_k^{(n)}(t)|}.$$

Notice that for each activation, the entire activation time series is first replaced with its l_1 norm. Then, the l_1 norms are scaled by the sum of l_1 norms to make $\boldsymbol{x}^{(n)}$ sum to one. We use the set of feature extracted from activation signals as input of a support vector machine (SVM) classifier. For training this SVM classifier, we use a linear kernel [57]: $K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = {\boldsymbol{x}^{(i)}}^{\top} \boldsymbol{x}^{(n)}$.

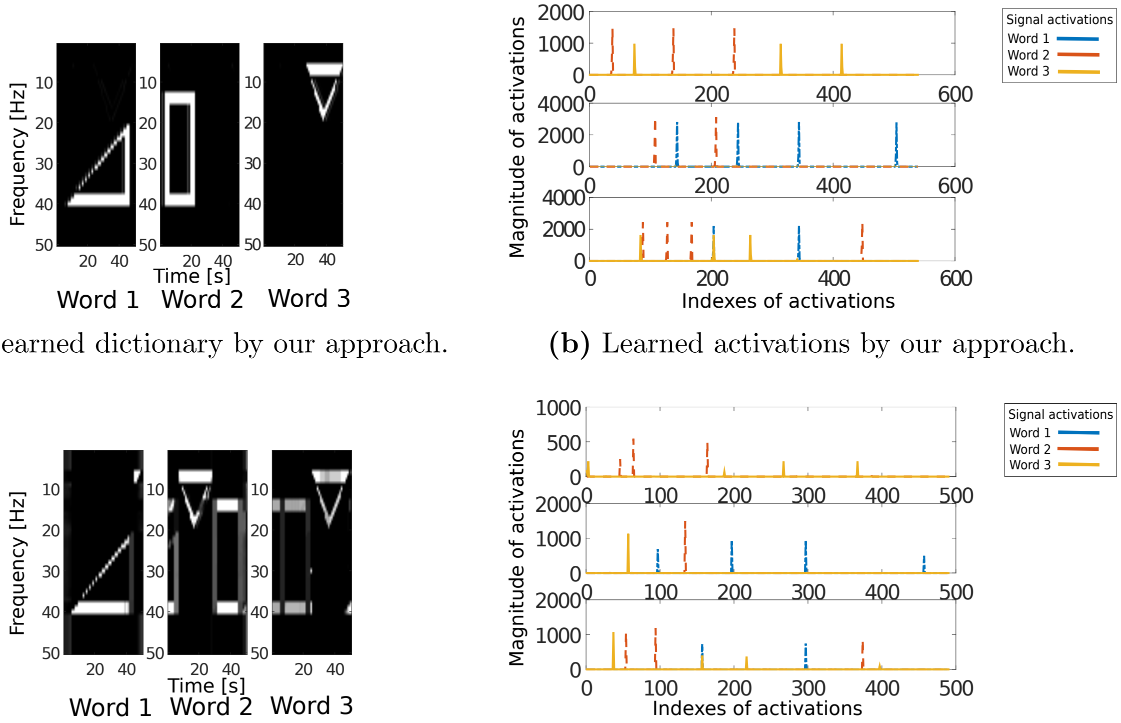
6.6. Experiments and discussion

In this section, we empirically evaluate the proposed random projected dictionary learning approach on both synthetic and real data. First, we compare how the boundary effect is addressed by our approach and CNMF. Additionally, we evaluate the proposed approach for the problems of denoising, dictionary discovery and classification of birdsong recordings.

6.6.1. Analysis on synthetic data

In this case, we use three spectrograms synthetically generated with three dictionary words and their corresponding sparse activation signals. The dimensions of each spectrogram are fixed to $F = 50 \times T = 500$, and the dimensions of each dictionary word are $F = 50 \times W = 50$.

The learned dictionary words for these three spectrograms and activations using our approach and CNMF [101] are shown in Fig. **6-3**. The number of iterations is $10,000$ in both cases. We observe the proposed approach accurately recovers the dictionary words (see Fig. **6-3**(a)) and the spectrograms (see Fig. **6-4**(a)) despite the boundary effect in the first spectrogram of Fig. **6-4**(c). However, CNMF learns each dictionary word as a mixture of the original dictionary words (see Fig. **6-3**(c) and Fig. **6-4**(b)) including the part of the dictionary word appearing in the beginning of the first spectrogram. As it can be seen, our model is more robust to boundary effects than CNMF.



(a) Learned dictionary by our approach.

(b) Learned activations by our approach.

(c) Learned dictionary by CNMF.

(d) Learned activations by CNMF.

Figure 6-3.Comparison of activations between our approach and CNMF [101].

6.6.2. Analysis on real-world data

In order to apply our random projected convolutive dictionary learning approach for birdsong analysis tasks, we use two real-world data sets:

- **MLSP 2013[2] dataset:** it contains 645 recordings of 19 different bird species (see Table **7-1**).
- **H. J. Andrews (HJA) dataset** [19]**:** it contains a total of 548 recordings with six different locations $PC1$, $PC4$, $PC7$, $PC8$, $PC13$, and $PC15$ (see Table **3-1**).

We convert each recording into a two-dimensional spectrogram with $F = 247$ and $T = 2497$ and examine four aspects of the proposed approach: (i) spectrogram denoising (ii) optimal parameter selection, (iii) dictionary learning, and (iv) species classification.

[2] https://www.kaggle.com/c/mlsp-2013-birds

Table 6-1.Number of recordings per species of MLSP2013 dataset.

Abbreviation (Class label)	Class name	# recordings
BRCR	Brown Creeper	14
PAWR	Pacific Wren	81
PSFL	Pacific-slope Flycatcher	46
RBNU	Red-breasted Nuthatch	9
DEJU	Dark-eyed Junco	20
OSFL	Olive-sided Flycatcher	14
HETH	Hermit Thrush	47
CBCH	Chestnut-backed Chickadee	40
VATH	Varied Thrush	61
HEWA	Hermit Warbler	53
SWTH	Swainson's Thrush	103
HAFL	Hammond's Flycatcher	28
WETA	Western Tanager	33
BHGB	Black-headed Grosbeak	9
GCKI	Golden Crowned Kinglet	37
WAVI	Warbling Vireo	17
MGWA	MacGillivray's Warbler	6
STJA	Stellar's Jay	10
CONI	Common Nighthawk	26

Spectrogram denoising

We use the proposed dictionary learning approach for spectrogram denoising. To this end, we learn a dictionary from a clean set of recordings (they are selected from the HJA dataset) and use it for recovering a rain corrupted spectrogram. The result in Fig. **6-5** shows that after running the dictionary learning algorithm, the rain artifact that appears as a long vertical line has been significantly reduced in the reconstructed spectrogram.

Parameter selection for dictionary learning

The model parameters that affect the performance of dictionary learning are the number of dictionary words K and sparsity of the activations λ. To show the relationship between the model parameters and the dictionary learning performance, we present the reconstruction error

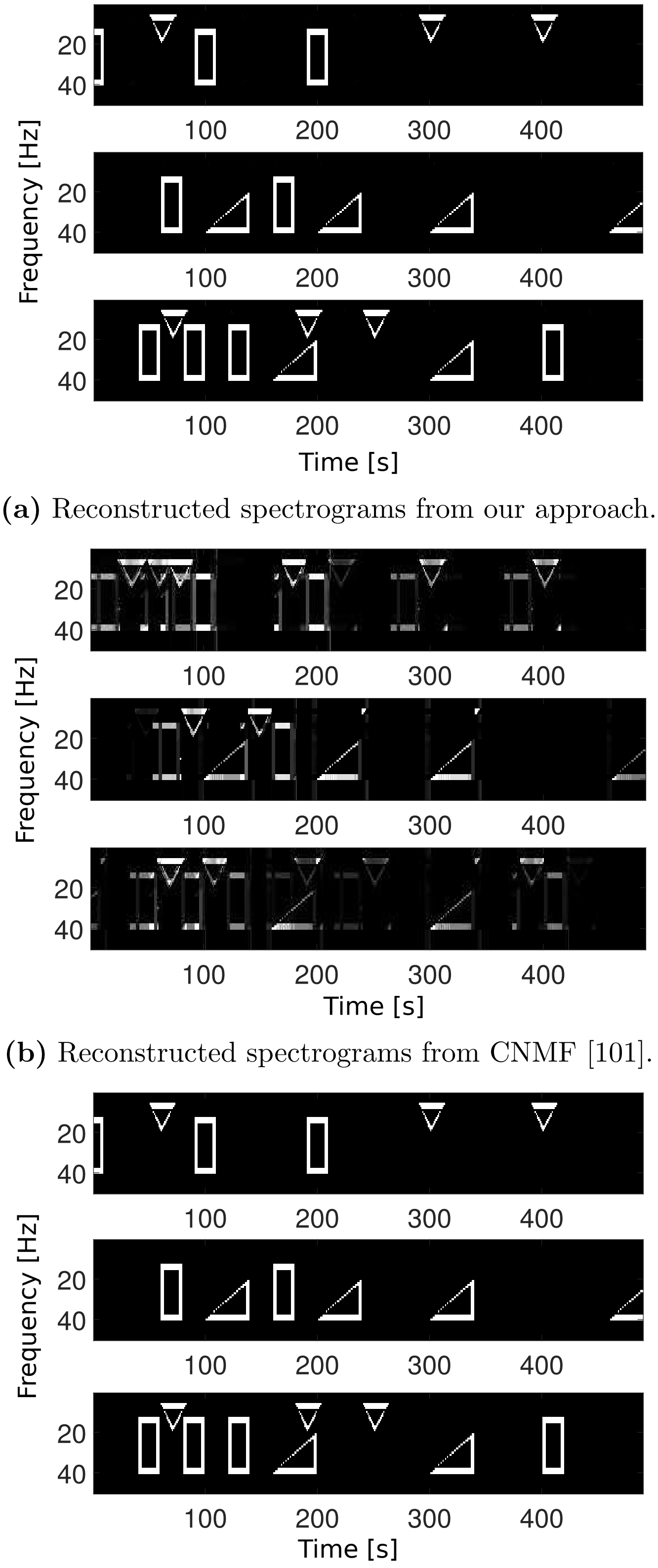
$$\sum_{n=1}^{N}\sum_{f=1}^{F}||\boldsymbol{y}_f^{(n)} - \sum_{k=1}^{K}\boldsymbol{T}_{\boldsymbol{a}_k}^{(n)}\boldsymbol{d}_{kf}||^2$$

against a practical approximation of the L_0 norm of the activations (number of the elements in $\mathcal{A}$ that are greater than $\epsilon = 10^{-2}$). During the training phase, we select 8 spectrograms from location $PC15$ of the HJA dataset and run the proposed algorithm to extract the dictionary words for each of the following parameter values $K = \{5, 10, 15\}$ and $\lambda = \{1, 5, 10, 15, 30, 50\}$. We apply the learned dictionary words in the validation phase to independent three test spectrograms, the performance curves are shown in Fig. **6-6(a)** (a) and (b). Results show that the reconstruction error decreases with decreasing value of λ and/or increasing the value of K. The L_0 norm of the activations increases with decreasing the value of λ. For a large λ, the dictionary concentrates on high energy words and low energy words are not discovered. For a small λ, the L_0 norm of activations increases significantly even though the reconstruction error decreases.

We select the optimal set of parameters ($\lambda = 10, K = 15$) to balance the reconstruction error and the sparseness of the activation in the validation set. We show the extracted dictionary words in the Fig. **6-7**.

Extracted dictionary words on MLSP2013 dataset

We select four or five rich-of-syllable spectrograms from each species to learn the bird song dictionary and show the discovered dictionary words of all 19 species in Fig. **6-8** by using randomly projected dictionary learning with $R = 10\%F$ and setting $W = 200$ for all species.

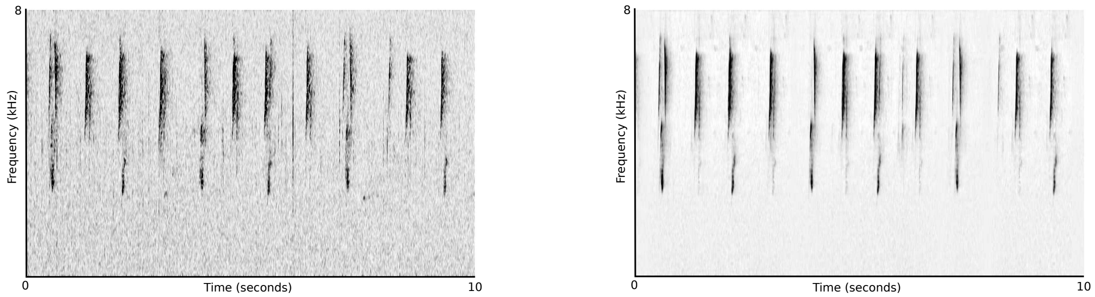


(a) Reconstructed spectrograms from our approach.

(b) Reconstructed spectrograms from CNMF [101].

(c) Original spectrograms.

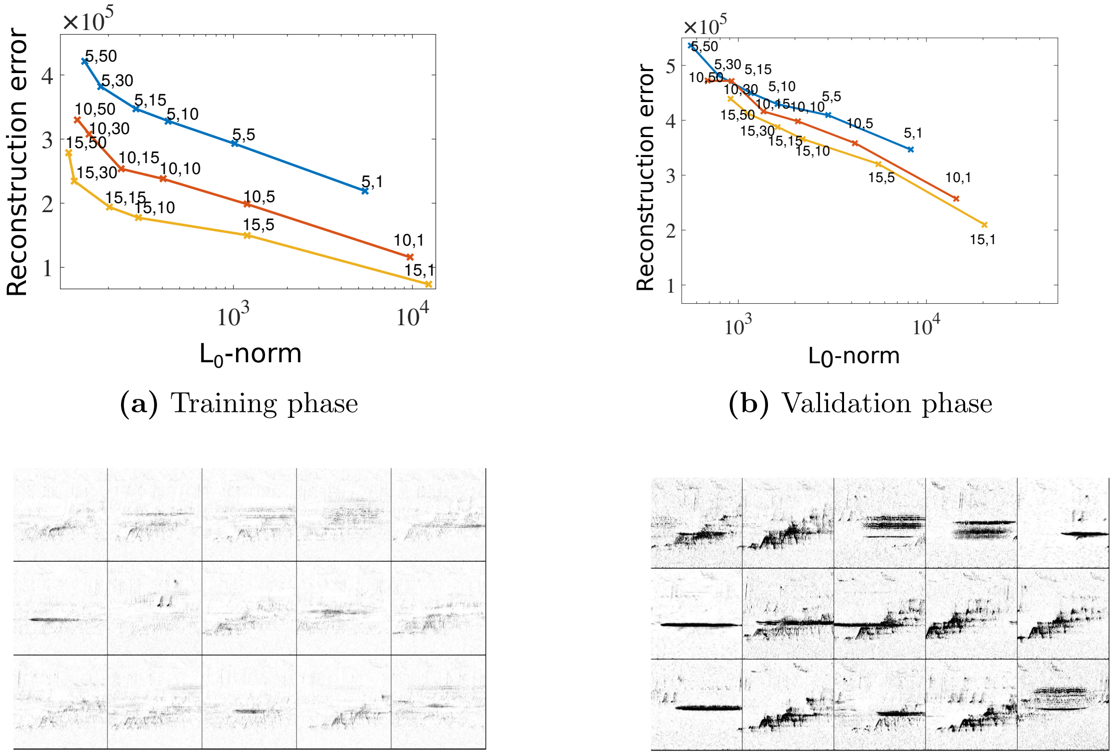
Figure 6-4.Comparison of original and recovered spectrograms



(a) Original spectrogram

(b) Reconstructed spectrogram

Figure 6-5.Examples of rain denoising on test spectrogram



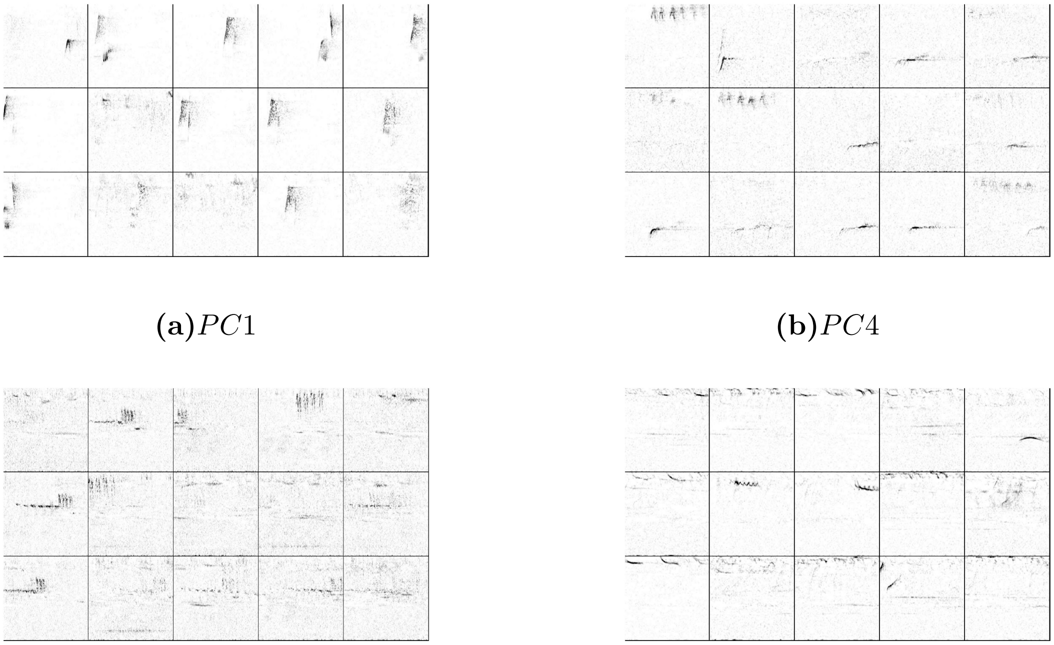
(a) Training phase

(b) Validation phase

(c) Learned dictionary

(d) Learned dictionary

Figure 6-6.Parameter selection: (a) training phase reconstruction error vs. L_0 norm of activations for $PC15$ (the first number for each point represents K and the second number for each point represents λ); (b) validation phase reconstruction error vs. L_0 norm of activations for $PC15$; (c) set of 15 learned dictionary words with $\lambda = 10$ for $PC15$; (d) set of 15 learned dictionary words with $\lambda = 50$ for $PC15$.



(a)$PC1$ **(b)**$PC4$

(c)$PC7$ **(d)**$PC13$

Figure 6-7.Learned dictionary words for recordings of HJA dataset on locations (a) PC1, (b) PC4 (c) PC7 and (d) PC13.

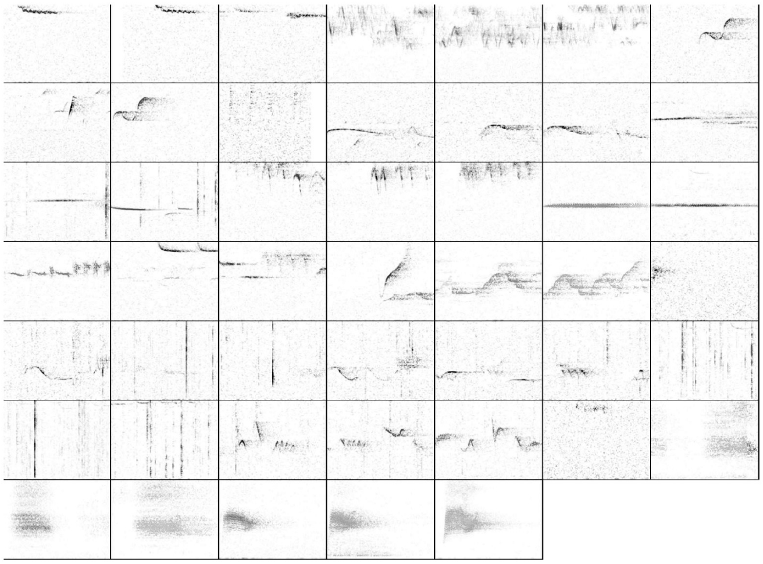


Figure 6-8.Learned dictionary words for MLSP2013 dataset

6.6.3. Classification experiments

In order to test the discriminative information provided by the learned dictionary, we formulate the problem of bird species recognition in recordings of the MLSP2013 dataset and HJA dataset (Tables **7-1** and **3-1**, respectively). In each classification experiment, we perform a five-fold cross-validation with ten repetitions. Since our experiments are carried out on multi-label datasets (each recording can be associated with multiple labels simultaneously), we report the results by evaluating each class label separately [123], i.e., the multi-label problem is seen as N_c binary-classification problems, where N_c is the number of classes, so that, each class is taken once as target class (labeled as '1') and the others as non-target class (labeled as '0'). Additionally, we choose area under the curve (AUC) as performance measure because it is regarded as appropriate for evaluating learning algorithms [58] in this setting in which label imbalance may occur. This measure estimates the probability that a classifier produce a higher output for an object of the target class than an object from the non-target class.

Random projection is applied with $R = 12$. The following are the used parameters: $\lambda = 0.1$ (heuristically fixed), $K = 2(N_c + 1)$ (i.e., two dictionary words are estimated for each class plus two words of a class which all the recordings belong, in order to find common patterns), and, 5000 iterations for estimating dictionary words from the training set and 10000 for estimating activations from the test set. The parameter C, which controls the trade off between errors of the SVM and margin maximization is selected among $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$.

In order to evaluate the performance of the proposed approach against other methods, we compared the proposed classification method that extracts features from activations using random projection (Act-RP) against classification methods whose features are extracted from: random projected spectrograms (Spc-RP), Mel-Frequency-Coefficients (MFC, reduction of the original frequency scale to a logarithmic scale with 12 coefficients) dimension reduced spectrograms and original spectrograms (Spc). Tables **6-2** and **6-3** show the classification performance (AUC mean and standard deviation) from HJA dataset and MLSP2013 dataset, respectively.

Table **6-2** shows that most of the species from the HJA dataset can be well-classified by using both the proposed method based on dictionary learning and extracting features directly from the spectrograms. Therefore, it means that the spectral information is highly relevant for classifying those recordings. Additionally, the good performance of the classification performed based on MFC features confirms that. The species that benefits mostly of our approach is `HETCH`. However, the baseline outperforms our method when `RBNU` is classified. We suspect that possibly highly relevant frequency information from this class is lost when the random projection is applied. Nevertheless, in comparison to the poor performance of classification from random-projected spectrograms without using DL, our DL approach provides a significant increase in performance by recovering the hidden activations and using them as features for classification.

Table 6-2. Classification results obtained with the HJA dataset and feature representation extracted from i) activations using the random projection (Act-RP); ii) the random projected spectrograms (Spc-RP); iii) MFC dimension reduced spectrograms (MFC) and iv) original spectrograms (Spc).

Class	Act-RP	Spc-RP	MFC	Spc
BRCR	**97.10** (0.70)	70.21 (2.53)	88.91 (1.67)	95.92 (0.70)
WIWR	**99.15** (0.36)	83.46 (1.68)	98.72 (0.50)	**99.08** (0.27)
PSFL	**94.54** (0.86)	72.45 (1.74)	91.10 (0.74)	**93.61** (1.20)
RBNU	90.04 (2.31)	43.26 (3.84)	90.05 (2.51)	**96.70** (1.38)
DEJU	**96.69** (1.13)	47.24 (4.06)	78.10 (3.05)	94.96 (2.11)
OSFL	**98.26** (0.56)	60.19 (3.17)	92.92 (0.64)	**98.18** (0.57)
HETH	**94.63** (2.81)	80.34 (4.37)	83.74 (4.15)	91.26 (2.25)
CBCH	**92.80** (2.03)	66.99 (1.97)	89.60 (1.03)	**92.89** (1.49)
VATH	**100** (0)	70.60 (3.29)	**100** (0)	**100** (0)
HEWA	**97.53** (0.49)	64.98 (3.39)	**98.23** (0.69)	**97.70** (0.80)
SWTH	**99.44** (0.24)	52.19 (4.75)	**99.01** (0.48)	**98.57** (0.82)
HAFL	**100** (0)	81.00 (3.22)	**100** (0)	**100** (0)
WETA	**97.02** (0.89)	89.39 (1.69)	87.38 (1.69)	**97.46** (1.17)

Classification experiments carried out with the MLSP2013 dataset behave differently. In this case, direct use of the spectrograms produces relatively low AUC results. Generally speaking the MLSP dataset is more challenging than the HJA dataset due to a larger number of species and the presence of rain that may overlap with bird vocalizations. It can be observed that our method outperforms other methods for almost all of the species with the exception of the BHGN species. As mentioned above, it is possible that vocalizations of this species contain discriminative information that is lost when the dimension reduction techniques are applied.

Table 6-3.Classification results obtained with the MLSP2013 dataset and feature representation extracted from i) activations using the random projection (Act-RP); ii) the random projected spectrograms (Spc-RP); iii) MFC dimension reduced spectrograms (MFC) and iv) original spectrograms (Spc).

Class	Act-RP	Spc-RP	MFC	Spc
BRCR	**97.86** (0.90)	62.78 (7.26)	89.84 (2.57)	86.51 (8.97)
PAWR	**88.71** (1.32)	66.03 (2.65)	86.69 (1.07)	85.92 (2.21)
PSFL	**89.37** (1.73)	43.68 (6.34)	71.24 (3.21)	71.94 (3.58)
RBNU	**68.89** (14.29)	21.99 (6.55)	53.55 (14.10)	58.18 (10.11)
DEJU	**83.05** (6.05)	62.78 (7.48)	65.90 (8.24)	68.81 (3.71)
OSFL	**93.97** (3.49)	69.70 (9.36)	78.60 (5.10)	72.78 (8.53)
HETH	**85.66** (3.65)	49.96 (3.09)	81.00 (2.77)	77.46 (4.58)
CBCH	**74.46** (4.77)	55.71 (6.87)	67.36 (4.10)	61.30 (7.06)
VATH	**87.29** (2.57)	60.20 (4.82)	81.59 (3.28)	78.02 (4.33)
HEWA	**90.41** (2.53)	63.05 (5.14)	78.20 (2.08)	74.10 (2.93)
SWTH	**89.24** (1.50)	50.15 (3.87)	85.74 (2.69)	84.01 (2.84)
HAFL	**85.54** (5.03)	64.93 (5.26)	75.58 (6.80)	69.33 (3.77)
WETA	**91.08** (2.72)	57.06 (6.82)	81.68 (4.45)	77.96 (3.86)
BHGB	48.87 (15.47)	49.78 (13.47)	49.05 (15.53)	**68.49** (14.19)
GCKI	**91.47** (2.44)	64.41 (5.59)	76.98 (3.52)	68.54 (5.91)
WAVI	**96.42** (1.69)	47.52 (8.26)	85.28 (2.65)	73.80 (4.76)
MGWA	**50.00** (15.81)	14.64 (8.48)	43.20 (13.56)	37.51 (14.30)
STJA	**92.86** (4.55)	58.26 (8.32)	71.62 (9.99)	64.24 (7.86)

7. Online learning of time-frequency patterns

Abstract

We present an online method to learn recurrent time-frequency patterns from spectrograms. Our method relies on a convolutive decomposition that estimates sequences of spectra into time-frequency patterns and their corresponding activation signals. This method processes one spectrogram at a time such that, in comparison with a batch method, the computational cost is reduced proportionally to the number of considered spectrograms. We use a first-order stochastic gradient descent and show that a monotonically decreasing learning-rate works appropriately. Furthermore, we suggest a framework to classify spectrograms based on the estimated set of time-frequency patterns. Results, on a set of synthetically generated spectrograms and a real-world dataset, show that our method finds meaningful time-frequency patterns and that it is suitable to handle a large amount of data.

7.1. Introduction[1]

Learning time-frequency patterns is helpful for both supervised and unsupervised analyses of acoustic signals. For this purpose, the mathematical model known as dictionary learning (DL) has been used. Estimation of such a model is usually formulated as a constrained optimization problem that includes a data fit term between the signal and a combination of a set of patterns —called *dictionary*— and their corresponding coefficients for weighting those patterns —called *activations*.

Depending on the problem, a physical meaning can be attributed to patterns and coefficients [66]. For example, for bioacoustic signals, dictionary patterns can be associated with

[1] This chapter was accepted for presentation on *The 42nd IEEE International Conference on Acoustics, Speech and Signal Processing* ICASSP 2017, March 5-9, 2017, New Orleans, USA.

different sound sources, e.g., bird species vocalization, and coefficients can be related to the time when the vocalizations are emitted. For later analysis, a DL algorithm should appropriately recover the original signal and satisfy the constraints, e.g., norm-constraints or non-negativity. Nevertheless, those algorithms are usually computationally expensive; therefore, to scale up and allow handling a large amount of data, it is important to consider complexity and memory requirements [60].

One approach for DL, which has been widely applied in machine learning and digital signal processing, is based on nonnegative matrix factorization (NMF) [86]. Particularly, NMF allows extracting meaningful information from audio recordings that contain mixtures of sounds [31, 103]. In order to apply NMF, the audio signal is usually represented by its spectrogram [7, 46, 98]. NMF has been successfully applied to various audio applications including automatic transcription, music analyses and blind source separation [30, 74]. NMF is formulated as an optimization problem (sparsity constraints are often added) that minimizes the least-squares error or the generalized Kullback-Leibler divergence [70] between the measured signal and its decomposition.

Using NMF a spectrum is decomposed into a product of two matrices: one corresponding to a collection of 1-D spectra (which forms the dictionary) and another corresponding to their activations in time. An alternative model is the convolutive non-negative matrix factorization (cNMF) in which each pattern of the dictionary is a matrix that corresponds to a sequence of 1-D spectra (time-frequency pattern) [83, 34]. The resulting time-frequency patterns provide useful information related to relevant temporal structures contained in the recordings [102]. Nevertheless, when dealing with large data (e.g., in bioacoustics), traditional cNMF algorithms become computationally expensive and demand large memory resources. To reduce the computational complexity and memory consumption, low-rank approximations are applied [124]. However, this approach generally results in information loss. An alternative approach to alleviate the processing requirements is using online algorithms. For instance, in [75], an algorithm for learning 1-D patterns using stochastic gradient descent is proposed and in [113], an online version of the cNMF algorithm proposed in [117] is introduced.

In this chapter, we propose an unsupervised online version of the algorithm originally presented in [95]. For this purpose, we use a first-order stochastic gradient descent approach. Our algorithm progressively updates the dictionary with each incoming spectrogram. Additionally, we propose a scheme for classifying audio signals based on features extracted from the convolutive decomposition of the spectrograms. We evaluate and compare the proposed approach on synthetic and real-world datasets.

7.2. Online dictionary learning

In Chapter 6, an iterative rule for updating the k-th time-frequency pattern is proposed, which can be rewritten as follows:

$$\boldsymbol{D}_{k(p)} = \Pi(\boldsymbol{D}_{k(p-1)} + \eta_d \nabla_{\boldsymbol{D}_k} \ell(\boldsymbol{Y}, \boldsymbol{D}_{k(p-1)}, \boldsymbol{A})), \tag{7-1}$$

where (p) denotes the current iteration, $\eta_d = 1/\gamma_d$ (γ_d is the step-size described in Chapter 6), and the projection Π is defined as

$$\Pi(\boldsymbol{D}_k) = \begin{cases} \boldsymbol{D}_k \text{ if } ||\boldsymbol{D}_k|| \leqslant 1 \\ \frac{\boldsymbol{D}_k}{||\boldsymbol{D}_k||} \text{ otherwise} \end{cases} \quad \forall k,$$

the projection projection $\Pi : \mathbb{R}^{N \times M} \rightarrow \mathbb{R}^{N \times M}$ is defined as

$$\Pi(\boldsymbol{M}) = \begin{cases} \boldsymbol{M} \text{ if } \|\boldsymbol{M}\|_F \leqslant 1 \\ \frac{\boldsymbol{M}}{\|\boldsymbol{M}\|_F} \text{ otherwise} \end{cases}$$

such that $\| \cdot \|_F$ is the Frobenius norm that is computed for any arbitrary matrix $\boldsymbol{M}$ by

$$\|\boldsymbol{M}\|_F = \sqrt{\sum_{n=1}^{N} \sum_{m=1}^{M} \boldsymbol{M}(n, m)^2},$$

. The gradient of the loss function wrt $\boldsymbol{D}_k$ is

$$\nabla_{\boldsymbol{D}_k} \ell(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{A}) = [\boldsymbol{v}_{\boldsymbol{d}_{k1}} \ldots \boldsymbol{v}_{\boldsymbol{d}_{kF}}]^\top \in \mathbb{R}^{F \times W} \tag{7-2}$$

where $\boldsymbol{v}_{\boldsymbol{d}_{kf}} = \boldsymbol{T}_{\boldsymbol{a}_k}^\top [\boldsymbol{y}_f - \sum_{k=1}^{K} \boldsymbol{T}_{\boldsymbol{a}_k} \boldsymbol{d}_{kf}^{(p-1)}] \in \mathbb{R}^{W \times 1}$.

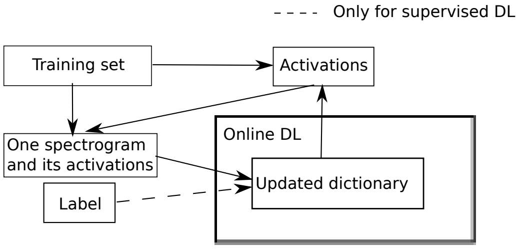

---- Only for supervised DL
Training set
Activations
Online DL
One spectrogram and its activations
Updated dictionary
Label

Figure 7-1.Diagram of the online method for learning time-frequency patterns. The dashed line is only enabled for the supervised online DL.

We propose an online version of the DL method (see Fig. **7-1**), which updates the dictionary by considering the current spectrogram and the ones processed in the past. In order to learn a dictionary from a set of N stacked spectrograms $\mathcal{Y} = \{\boldsymbol{Y}^{(1)} \dots \boldsymbol{Y}^{(N)}\}$, in [95], the updating rule of (7-1) is applied as follows:

$$\boldsymbol{D}_{k(p)} = \Pi(\boldsymbol{D}_{k(p-1)} + \eta_d \sum_{n=1}^{N} \nabla_{\boldsymbol{D}_k} \ell(\boldsymbol{Y}^{(n)}, \boldsymbol{D}_{(p-1)}, \boldsymbol{A}^{(n)})). \tag{7-3}$$

Alternatively, we propose an online algorithm that updates the time-frequency patterns according to the current spectrogram and the ones observed in the past. Therefore, we define the following loss function:

$$g_N(\boldsymbol{D}) := \frac{1}{N} \sum_{n=1}^{N} \ell(\boldsymbol{Y}^{(n)}, \boldsymbol{D}, \boldsymbol{A}^{(n)})$$

where $\boldsymbol{A}^{(n)}$ is the estimated activation matrix that corresponds to the n-th spectrogram $\boldsymbol{Y}^{(n)}$. Hence, the dictionary learning task consists in minimizing the expected cost

$$g(\boldsymbol{D}) := \mathbb{E}_{\boldsymbol{Y}}[\ell(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{A})] := \lim_{N \to \infty} g_N(\boldsymbol{D}).$$

For this purpose, we update $\boldsymbol{D}_k$ by using the first-order stochastic gradient descent algorithm [2, 75] as follows:

$$\boldsymbol{D}_{k(p)} = \Pi(\boldsymbol{D}_{k(p-1)} + \mu_p \eta_d \nabla_{\boldsymbol{D}_k} \ell(\boldsymbol{Y}^{(n)}, \boldsymbol{D}_{(p-1)}, \boldsymbol{A}^{(\boldsymbol{n})})) \tag{7-4}$$

where

$$n = \begin{cases} N & \text{if } \mathrm{mod}(p, N) = 0 \\ \mathrm{mod}(p, N) & \text{otherwise,} \end{cases}$$

and μ_p is the factor for scaling the gradient, also known as learning-rate. Notice that one iteration of (7-3) requires computing N times the gradient $\nabla_{\boldsymbol{D}_k} \ell(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{A})$ but (7-4) requires computing this gradient only once.

According to [28], two learning-rate schedules commonly used in matrix factorization are:

1. **Fixed Schedule (FS):** the learning rate $\mu_p = \alpha \; \forall p$ is fixed throughout the online learning process.

2. **Monotonically Decreasing Schedule (MDS):** the learning rate monotonically decreases each time that a new spectrogram is observed. Two options are: i) MDS1: $\mu_p = \frac{\alpha}{p}$, and ii) MDS2: $\mu_p = \frac{\alpha}{\sqrt{p}}$.

Our DL process, which aims to compute $\mathcal{A} = \{\boldsymbol{A}^{(1)} \dots \boldsymbol{A}^{(N)}\}$ and $\boldsymbol{D}$, alternatively updates both of them. Therefore, in the p-th iteration, the algorithm updates $\boldsymbol{A}^{(n)}$ for $1 \leqslant n \leqslant N$ by (6-23), and $\boldsymbol{D}$ by (7-4). Note that due to the non-convex nature of the problem convergence to a global optimum is not guaranteed.

7.3. Classifying spectrograms

Our classification task consists of mapping the vector representation of a spectrogram $\boldsymbol{x} \in \mathbb{R}^K$ –which is computed by using the learned dictionary– to a categorical (class) label $y \in \{-1, 1\}$. The label in the binary classification setting indicates the presence ($y = 1$) or absence ($y = -1$) of the target class in a given spectrogram. For this purpose, we divide the experiments into two stages: training and test. In the training stage, the dictionary is estimated by using the proposed online DL method, which receives a sequence of spectrograms. The learned dictionary words are used to extract the feature vector $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{iK}] \in \mathbb{R}^K$ for the i-th spectrogram $\boldsymbol{Y}_i \in \mathbb{R}^{F \times T}$ in a training set, in such a way that each one of the entries of $\boldsymbol{x}_i$ corresponds to the point of maximum correlation of a dictionary word and the spectrogram. So, it is computed as follows: $x_{ik} = \max\limits_t |\sum\limits_f \boldsymbol{h}_{kf}^{(i)}(t)|$ where $\boldsymbol{h}_{kf}^{(i)} = \overleftarrow{\boldsymbol{d}}_{kf} * \boldsymbol{y}_f^{(i)} \in \mathbb{R}^{(T+W-1)\times 1}$, and $\overleftarrow{\cdot}$ denotes the vector in reversed order.

7.4. Experiments and discussion

7.4.1. Experiments on an artificial dataset

Initially, we perform experiments in a collection of 1000 synthetically generated spectrograms containing some of six different time-frequency patterns. The three-dimensional binary label vector of each spectrogram $\boldsymbol{Y} \in \mathbb{R}^{16 \times 30}$ indicates the presence or absence of each class in the spectrogram. We generate the training spectrograms by randomly combining the words (an adding Gaussian noise) of an “original dictionary” formed by six basic time-frequency patterns (see Fig. **7-4(a)**). In the original dictionary, two types of time-frequency patterns of length 10 correspond to each class.

The free parameters in the proposed online DL method are: length of window W, number of dictionary words K, learning-rate μ, and ℓ_1-norm regularization parameter λ. We fix $W = 10$, since this parameter is known beforehand, and $K = 8$ (we over-estimate the size of the dictionary in order to avoid missing a time-frequency pattern). Estimation of the remaining parameters is described below.

We compare the schedules of μ described in Sec. 7.2 and tune the parameter α. Figure **7-2** contains the reconstruction error of a test set of 30 spectrograms and the actual sparsity of their activations (rate of non-zero entries) as a function of the number of observed spectrograms for a set of different values of α (for $\lambda = 0.1$). According to this experiment, the FS schedule works well for moderate values of α trading off initial instability at a large value of α with slow convergence for a small value of α.

Figure **7-3** shows the reconstruction error and the rate of non-zero entries in function of λ after observing 1000 spectrograms (the learning-rate is MDS1 for a set of different values of α). Results confirm the trade-off in the objective function between the reconstruction error and the ℓ_1-norm constraint. Figures **7-4(a)** and **7-4(b)** show the original set of time-frequency patterns and the estimated ones (with MDS1, $\alpha = 100$ and $\lambda = 0.1$), respectively.

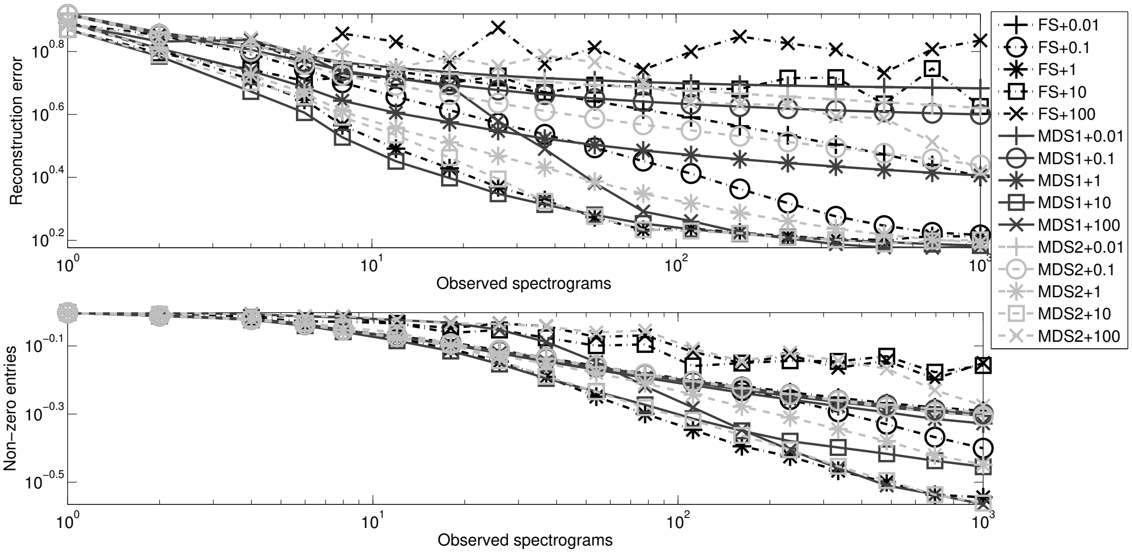


Figure 7-2. Comparison of the studied learning-rate schedules, in a test set of 30 spectrograms, for a set of different values of α (FS+α, MDS1+α and MDS2+α) and $\lambda = 0.1$.

7.4.2. Experiments on real-world datasets

To validate the proposed method, we perform experiments on the MLSP 2013 Bird Classification Challenge dataset,[3] which was collected in the H. J. Andrews (HJA) Long-Term Experimental Research Forest in Oregon (USA). Table **7-1** shows the number of recordings and classes of this dataset.

The classification experiments consider the following: i) for each class a binary (presence/absence) classification problem is considered; ii) the dataset is randomly divided into 50% for training and 50% for test (with 20 repetitions); iii) spectrograms are computed with

[3] `https://www.kaggle.com/c/mlsp-2013-birds`

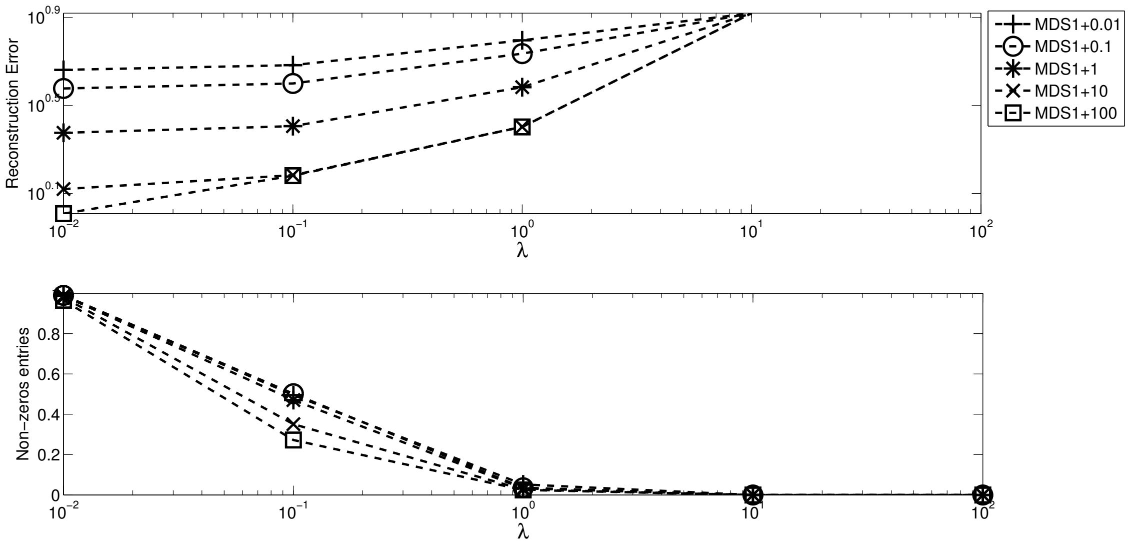
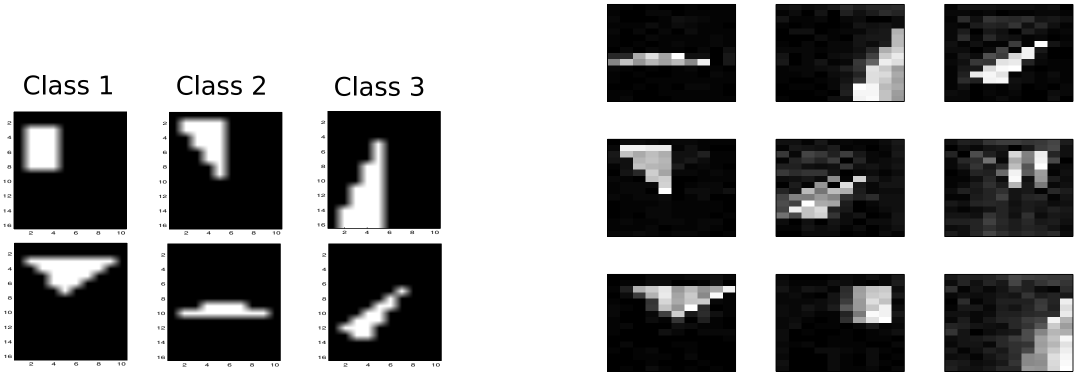


Figure 7-3.Reconstruction error and rate of non-zero entries in function of λ after observing 1000 spectrograms (the learning-rate is MDS1 for a set of different values of α).

dimensions $F = 80$ and $T = 250$ (corresponding to 10 seconds); iv) the parameters of DL are: $W = 25$ (window length of 1 sec.), $K = 6$, λ is tuned for $\{0.01, 0.1, 1, 10\}$, learning-rate MDS1 where α is tuned for $\{1, 10, 100\}$, and 1000 iterations (due to the small size of the dataset, online DL cycles through the spectrograms several times to allow for a number of iterations that is greater than the number of spectrograms available in the dataset); v) feature representation as indicated in Sec. 7.3; vi) linear SVM classifier (the best regularization parameter C is searched using a cross-validation in the training set, such that $C \in [1 \times 10^{-2}, 1 \times 10^{2}]$); and, vii) performance is reported by the F-score.

Table **7-1** shows the classification performance of three methods: 1) *Wang et. al. (2013)* that considers the proposed classification framework but learns the dictionary by the online method proposed in [113] (using our own implementation); 2) *Online DL* that applies the proposed online DL method and classification framework; and, 3) *Frequency* that applies the proposed classification framework, but the feature representation is directly extracted from the spectrograms, i.e., the feature vector corresponds to the normalized average spectra. According to our results, the proposed *Online DL* outperforms the others in 12 of the 19 classes. *Frequency* outperforms the others when classifying `BRCR`, `VATH`, `BHGB`, and `MGWA`. Among these classes, the performance is remarkably high for `BHGB`, this suggests that the frequency band is enough to distinguish this species. *Wang et al. (2013)* produces the best performance when classifying `OSFL`, `CBCH`, and `WETA`.

Both types of experiments, on the artificial dataset and on the real-world dataset, show that



(a) Original dictionary

(b) Estimated dictionary (with MDS1, $\alpha = 100$ and $\lambda = 0.1$)

Figure 7-4.Original and estimated sets of time-frequency patterns.

the method recovers appropriately the original spectrograms and finds meaningful time-frequency patterns for classification outperforming a the baseline DL method and the classification based on the raw frequency information.

Table 7-1. Number of 10-second recordings per species of MLSP2013 dataset. Size of the traing/test datasets. F-score performance of classification experiments (boldface indicates the highest result per species).

Label	#	F-score performance		
	recordings	Wang et al.	Online DL	Frequency
BRCR	14	70.8 ± 7.5	92.8 ± 3.7	**94.1** ± **0.7**
PAWR	81	80.1 ± 3.4	**84.9** ± **1.3**	79.4 ± 2.4
PSFL	46	75.1 ± 5.3	**84.0** ± **3.4**	77.8 ± 2.4
RBNU	9	53.4 ± 10.8	**83.9** ± **8.7**	79.7 ± 7.6
DEJU	20	87.8 ± 4.3	**89.9** ± **4.7**	80.9 ± 2.4
OSFL	14	**90.0** ± 5.0	79.7 ± 7.3	88.6 ± 4.6
HETH	47	70.6 ± 5.5	**80.5** ± **5.1**	78.0 ± 2.5
CBCH	40	**83.9** ± **4.5**	74.7 ± 6.0	81.7 ± 1.5
VATH	61	74.7 ± 4.5	83.6 ± 3.0	**84.1** ± **0.6**
HEWA	53	75.5 ± 4.8	**80.7** ± **5.0**	77.7 ± 2.7
SWTH	103	70.0 ± 4.0	**82.4** ± **4.3**	77.2 ± 2.4
HAFL	28	81.6 ± 5.1	**85.8** ± **4.3**	74.1 ± 2.5
WETA	33	**88.1** ± **4.2**	75.3 ± 6.6	86.4 ± 0.8
BHGB	9	70.2 ± 9.1	67.5 ± 9.9	**95.5** ± **0.5**
GCKI	37	67.8 ± 5.6	**85.4** ± **6.0**	83.0 ± 1.4
WAVI	17	83.4 ± 4.9	**92.7** ± **6.8**	89.0 ± 1.1
MGWA	6	42.3 ± 10.6	77.3 ± 8.7	**86.3** ± **6.5**
STJA	10	86.7 ± 6.5	**94.3** ± **7.8**	93.6 ± 0.9
CONI	26	86.0 ± 3.5	**89.1** ± **4.6**	85.6 ± 1.4

7.4.3. Computational cost: Batch Learning vs Online Learning

In order to show the computational benefits of our method (online learning), we compare it against a batch learning approach. We call batch learning to the DL method that updates the time-frequency patterns by (7-3), which requires the whole set of spectrograms to estimate the gradient. These experiments were carried out on a CPU with Processor 2.20GHz × 8 and Memory 3.8 GB.

Figure **7-5** compares the time needed to reconstruct 20 (randomly selected) spectrograms from the MLSP 2013 dataset by batch learning and online learning. Note that since there are more than 20 iterations, the spectrograms are observed several times in the online case. We observe that the error is not monotonically decreasing at the beginning for online learning. However, the reconstruction error for both the online and batch methods converges to a similar value after several iterations. Furthermore, as expected, the online learning is faster than the batch learning by a factor of the number of spectrograms reconstructed at each iteration. Since the proposed method handles better the computational resources than its batch counterpart, it might be preferred for analyzing large datasets.

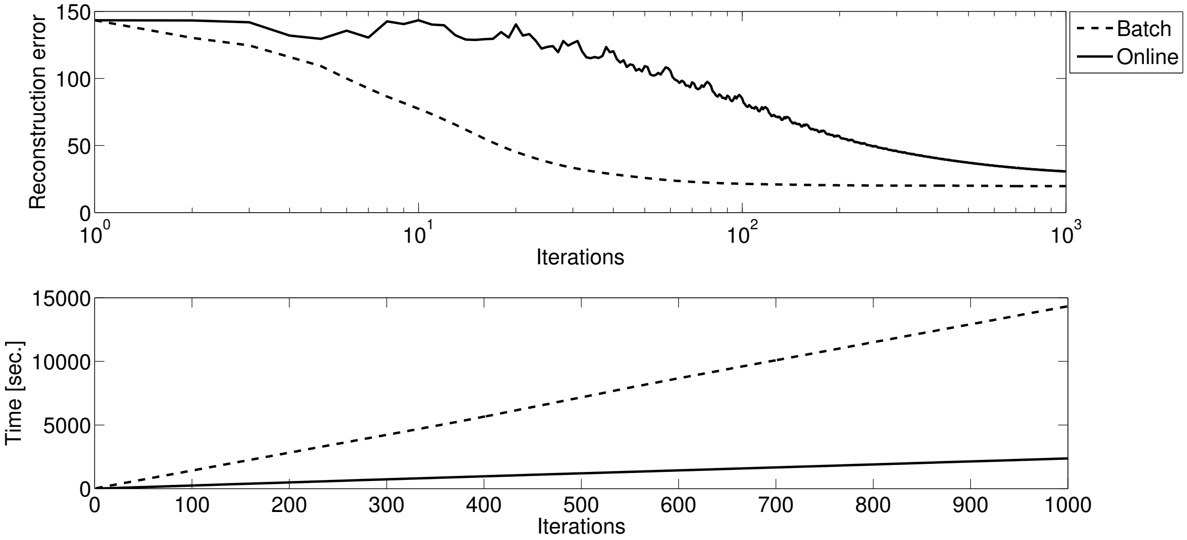


Figure 7-5. Comparison of reconstruction error (top) and computational cost (bottom) between batch learning and online learning.

8. Conclusion

8.1. Concluding remarks

The technological tools currently available to acquire, store and process information can be used to support the collection of environmental information. It facilitates tasks such as counting species, rapid assessment of biodiversity with large spatial and temporal coverage, the enrichment of information studies of ecology and conservation, development of automated monitoring activities, monitoring of rare or endangered species, invasive species detection, estimation of the impact of climate change, and analysis of the effects of anthropological activity.

Particularly, this thesis provides the description and improvements of two state-of-the-art approaches of PR and DSP: multiple instance learning (MIL) and dictionary learning (DL). Regarding MIL, we described the application of this weakly supervised learning approach in bioacoustic problems and its advantages. In the first part of this document, it is proposed an unsupervised recording segmentation method of audio birdsong recordings that improves species classification with the benefit of easier implementation since no manual handling of recordings is required. Afterward, we especially focused on the dissimilarity-based MIC. Our results showed 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 and confirmed 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. Furthermore, we proposed a method for improving the 1-NN classification on multiple instance datasets, which uses a supervised projection of the original feature space and a novel relative minimum distance between test bags and prototype bags.

In the second part, a novel convolutive DL method for learning a representative dictionary from a collection of multi-labeled audio dataset is proposed. About this topic, we proposed a dictionary learning approach for randomly projected spectrograms. This approach combines the power of estimating time-frequency patterns given by the convolutive model and the computational complexity reduction associated with the random projection approach. Additionally, we address the boundary effect arising in a collection of discontinuous spectrograms. Furthermore, we introduce a step-size selection criterion to improve convergence rate when updating the activations and the dictionary words. The DL method was successfully applied to spectrogram denoising and species classification. Finally, we presented an efficient online version of the DL method –based on stochastic gradient descent– that outperforms other state-of-the-art batch and online methods, in both, computational cost and quality of the discovered patterns.

8.2. Recommendations

Environmental acoustic signals are not acquired under controlled conditions. Therefore, automated recognition systems must have robust methods against noise, variations in intensity and not demand high computational requirements —taking into account the amount of data required to inspect. Moreover, given the random and non-stationary nature of bioacoustic signals (temporal and spatial variations of its parameters), these systems should allow the eventual inclusion of corrections suggested by experts.

In order to extend the studied multi-instance approaches to another applications or datasets, the instance-extraction stage has to be implemented taking into account the nature of the data and any prior knowledge that is available.

We recommend using dissimilarity-based MIC —with the proposed improvements— in bioacoustic recognition tasks (and in general for PR) because it reduces the required information for training and exhibits a good performance. Also, we suggest using the MildML projected space to classify instances and reduce dimensionality. Besides, we encourage to apply DL for unsupervised analysis of raw data, feature extraction, classification, and denoising tasks.

8.3. Future work

As future work, with respect to MIL, we consider that more studies must be undertaken on improving the feature extraction stage. Besides, the use of classifiers demanding less input training examples remains an important issue since, in practice, collecting labeled data is very costly. Regarding DL, supervised and semi-supervised analyses could improve the information provided by the learned dictionary. Moreover, further research in MIL and DL can be carried out by taking into account multi-labeled or unlabeled data.

Bibliography

[1] Acevedo, Miguel A.; Corrada-Bravo, Carlos J.; Corrada-Bravo, Héctor; Villanueva-Rivera, Luis J.; Aide, T. M.: Automated classification of bird and amphibian calls using machine learning: A comparison of methods. In: *Ecological Informatics* . 4 (2009), Nr. 4, S. 206 – 214

[2] Aharon, Michal; Elad, Michael: Sparse and Redundant Modeling of Image Content Using an Image-Signature-Dictionary. In: *SIAM Journal on Imaging Sciences* . 1 (2008), Nr. 3, S. 228–247

[3] Aide, T. Mitchell; Corrada-Bravo, Carlos; Campos-Cerqueira, Marconi; Milan, Carlos; Vega, Giovany; Alvarez, Rafael: Real-time bioacoustics monitoring and automated species identification. In: *PeerJ* . 1 (2013), S. e103

[4] Amores, Jaume: Multiple instance classification: Review, taxonomy and comparative study. In: *Artificial Intelligence* . 201 (2013), August, S. 81–105

[5] Armitage, David W.; Ober, Holly K.: A comparison of supervised learning techniques in the classification of bat echolocation calls. In: *Ecological Informatics* . 5 (2010), Nr. 6, S. 465 – 473

[6] Ayadi, Moataz E.; Kamel, Mohamed S.; Karray, Fakhri: Survey on speech emotion recognition: Features, classification schemes, and databases. In: *Pattern Recognition* . 44 (2011), Nr. 3, S. 572 – 587

[7] Badeau, Roland; Plumbley, M: Multichannel high resolution NMF for modelling convolutive mixtures of non-stationary signals in the time-frequency domain. In: *Transactions on Audio, Speech and Language Processing* . 22 (2013), Nr. 11, S. 1670–1680

[8] Baraniuk, Richard: Compressive sensing. In: *IEEE signal processing magazine* . 24 (2007), Nr. 4

[9] Bardeli, Rolf: Similarity Search in Animal Sound Databases. In: *IEEE Transactions on Multimedia* . 11 (2009), jan, Nr. 1, S. 68 –76

[10] Bardeli, Rolf; Wolff, Daniel; Kurth, F.; Koch, M.; Tauchert, K. H.; Frommolt, Karl-Heinz: Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. In: *Pattern Recognition Letters* . 31 (2010), September, Nr. 12, S. 1524–1534

[11] Bengio, Yoshua; Courville, Aaron; Vincent, Pascal: Representation Learning: A Review and New Perspectives. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* . 35 (2013), Nr. 8, S. 1798–1828

[12] Blumstein, Daniel T.; Mennill, Daniel J.; Clemins, Patrick; Girod, Lewis; Yao, Kung; Patricelli, Gail; Deppe, Jill L.; Krakauer, Alan H.; Clark, Christopher; Cortopassi, Kathryn A.; Hanser, Sean F.; McCowan, Brenda; Ali, Andreas M.; Kirschel, Alexander N. G.: Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. In: *Journal of Applied Ecology* . 48 (2011), Nr. 3, S. 758–767

[13] van den Boomgaard, Rein; van Balen, Richard: Methods for Fast Morphological Image Transforms Using Bitmapped Binary Images. In: *Graphical Models and Image Processing* . 54 (1992), Nr. 3, S. 252–258. – ISSN 1049–9652

[14] Botero, Jorge E.; Arbeláez, Daniel; Lentijo, Gloria M.: Métodos para estudiar las aves. In: *Biocarta* . (2005), Juli, Nr. 8, S. 1–4

[15] Bramer, Max: *Principles of Data Mining.* Second. Springer, 2013

[16] Brandes, Scott T.: Automated sound recording and analysis techniques for bird surveys and conservation. In: *Bird Conservation International* . 18 (2008), S. S163–S173

[17] Briggs, Forrest; Fern, Xiaoli; Raich, Raviv: Acoustic Classification of Bird Species from Syllables: an Empirical Study. / Oregon State University. 2009. – Forschungsbericht

[18] Briggs, Forrest; Fern, Xiaoli Z.; Raich, Raviv; Lou, Qi: Instance Annotation for Multi-Instance Multi-Label Learning. In: *ACM Transactions on Knowledge Discovery from Data* . 7 (2013), Nr. 3

[19] Briggs, Forrest; Lakshminarayanan, Balaji; Neal, Lawrence; Fern, Xiaoli Z.;

Raich, Raviv; Hadley, Sarah J K.; Hadley, Adam S.; Betts, Matthew G.: Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. In: *The Journal of the Acoustical Society of America* . 131 (2012), Nr. 6, S. 4640–50. – ISSN 1520–8524

[20] Catchpole, Clive K.; Slater, Peter J. B.: *Bird Song - Biological Themes and Variations*. Cambridge University Press, 1995

[21] Caycedo-Rosales, Paula C.; Ruiz-Muñoz, José F.; Orozco-Alzate, Mauricio: Reconocimiento automatizado de señales bioacústicas: Una revisión de métodos y aplicaciones. In: *Ingeniería y Ciencia* . 9 (2013), Nr. 18, S. 171–195

[22] Chen, Yixin; Bi, Jinbo; Wang, James Z.: MILES: Multiple-Instance Learning via Embedded Instance Selection. In: *IEEE transactions on Pattern Analysis and Machine Intelligence* . 28 (2006), Nr. 12, S. 1931–1947

[23] Chen, Zhixin; Maher, Robert C.: Semi-automatic classification of bird vocalizations using spectral peak tracks. In: *The Journal of the Acoustical Society of America* . 120 (2006), Nr. 5, S. 2974–2984

[24] Cheplygina, Veronika; Tax, David M J.; Loog, Marco: Dissimilarity-Based Ensembles for Multiple Instance Learning. In: *IEEE Transactions on Neural Networks and Learning Systems* . (2015), S. 1–13

[25] Cheplygina, Veronika; Tax, David M. J.; Loog, Marco: Multiple instance learning with bag dissimilarities. In: *Pattern Recognition* . 48 (2015), Nr. 1, S. 264 – 275

[26] Chesmore, David: The Automated Identification of Taxa: Concepts and Applications. In: MacLeod, Norman (Hrsg.): *Automated Taxon Identification in Systematics: Theory, Approaches and Applications* Bd. 74. Boca Raton, FL : CRC Press, 2008, Kapitel 6, S. 83–100

[27] Chesmore, David; Frommolt, Karl-Heinz; Wolff, Daniel; Bardeli, Rolf; Huebner, Sebastian: Computational Bioacoustics: New Tools for Assessing Biological Diversity, 2008. – Side Event at the ninth meeting of the Conference of the Parties (COP 9). Bonn, Germany

[28] Chin, Wei S.; Zhuang, Yong; Juan, Yu C.; Lin, Chih J.: A learning-rate schedule for stochastic gradient methods to matrix factorization. In: *Advances in Knowledge Discovery and Data Mining. 19th Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam, May 19-22, 2015, Proceedings, Part I* Bd. 9077, 2015, S. 442–

455

[29] Chu, Wei; Blumstein, D. T.: Noise robust bird song detection using syllable pattern-based hidden Markov models. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011*, 2011, S. 345 –348

[30] Cichocki, Andrzej; Zdunek, Rafal; Amari, Shun-ichi: New algorithms for non-negative matrix factorization in applications to blind source separation. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings.* Bd. 5. IEEE, 2006, S. V–621–V–624

[31] Cichocki, Andrzej; Zdunek, Rafal; Phan, Anh H.; Amari, Shun-ichi: *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation.* John Wiley & Sons, 2009

[32] Cugler, Daniel C.; Medeiros, Claudia B.; Toledo, Luis F.: Managing animal sounds-some challenges and research directions. In: *Proceedings V eScience Workshop-XXXI Brazilian Computer Society Conference*, 2011

[33] Dalal, N.; Triggs, B.: Histograms of oriented gradients for human detection. In: *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* Bd. 1, 2005, S. 886–893

[34] De Fréin, Ruairí; Rickard, Scott T.: Learning speech features in the presence of noise: Sparse convolutive robust non-negative matrix factorization. In: *16th International Conference on Digital Signal Processing.* IEEE, 2009, S. 1–6

[35] Dennis, Jonathan; Tran, Huy D.; Li, Haizhou: Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions. In: *Signal Processing Letters, IEEE* . 18 (2011), Nr. 2, S. 130–133

[36] Depraetere, Marion; Pavoine, Sandrine; Jiguet, Fréderic; Gasc, Amandine; Duvail, Stéphanie; Sueur, Jerôme: Monitoring animal diversity using acoustic indices: Implementation in a temperate woodland. In: *Ecological Indicators* . 13 (2012), Nr. 1, S. 46 – 54

[37] Diene, Oumar; Bhaya, Amit: Conjugate gradient and steepest descent constant modulus algorithms applied to a blind adaptive array. In: *Signal Process.* . 90 (2010), Nr. 10, S. 2835–2841

[38] Dietterich, Thomas G.; Lathrop, Richard H.; Lozano-Pérez, Tomás: Solving

the multiple instance problem with axis-parallel rectangles. In: *Artificial Intelligence* . 89 (1997), S. 31–71

[39] Du, Ruo; Wu, Qiang; He, Xiangjian; Yang, Jie: MIL-SKDE: Multiple-instance learning with supervised kernel density estimation. In: *Signal Process.* . 93 (2013), Nr. 6, S. 1471–1484

[40] Dubuisson, M.-P.; Jain, A. K.: A modified Hausdorff distance for object matching. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* Bd. 1, 1994, S. 566–568

[41] Duin, Robert P. W.; Bicego, Manuele; Orozco-Alzate, Mauricio; Kim, Sang-Woon; Loog, Marco: Metric Learning in Dissimilarity Space for Improved Nearest Neighbor Performance. In: Fränti, Pasi (Hrsg.); Brown, Gavin (Hrsg.); Loog, Marco (Hrsg.); Escolano, Francisco (Hrsg.); Pelillo, Marcello (Hrsg.): *Structural, Syntactic, and Statistical Pattern Recognition* Bd. 8621. Springer Berlin Heidelberg, 2014, S. 183–192

[42] Duin, Robert P. W.; Pekalska, Elżbieta: The dissimilarity space: Bridging structural and statistical pattern recognition. In: *Pattern Recognition Letters* . 33 (2012), Nr. 7, S. 826 – 832

[43] Efron, Bradley; Hastie, Trevor; Johnstone, Iain; Tibshirani, Robert [u. a.]: Least angle regression. In: *The Annals of statistics* . 32 (2004), Nr. 2, S. 407–499

[44] Fagerlund, Seppo: Bird Species Recognition Using Support Vector Machines. In: *EURASIP Journal on Advances in Signal Processing* . 2007 (2007), Nr. 1, S. 64–64. – ISSN 1110–8657

[45] Farnsworth, Andrew; Russell, Robert W.: Monitoring flight calls of migrating birds from an oil platform in the northern Gulf of Mexico. In: *Journal of Field Ornithology* . 78 (2007), Nr. 3, S. 279–289

[46] Févotte, Cédric; Kowalski, Matthieu: Low-rank time-frequency synthesis. In: *Advances in Neural Information Processing Systems*, 2014, S. 3563–3571

[47] Foulds, James; Frank, Eibe: A review of multi-instance learning assumptions. In: *The Knowledge Engineering Review* . 25 (2010), Nr. 01, S. 1–25

[48] Frommolt, Karl-Heinz; Tauchert, Klaus-Henry: Applying bioacoustic methods for long-term monitoring of a nocturnal wetland bird. In: *Ecological Informatics* . 21

(2014), S. 4 – 12. – Ecological Acoustics

[49] Ganchev, Todor D.; Jahn, Olaf; Marques, Marinez I.; De Figueiredo, Josiel M.; Schuchmann, Karl L.: Automated acoustic detection of Vanellus chilensis lampronotus. In: *Expert Systems with Applications* . 42 (2015), Nr. 15-16, S. 6098–6111

[50] Gaunt, Sandra; McCallum, Archibald: Birdsong and Conservation. In: *Nature's music: the science of birdsong.* 2004, Kapitel 12, S. 343–362

[51] Grant, B. Rosemary; Grant, Peter R.: Hybridization and speciation in Darwin's finches: the role of sexual imprinting on a culturally transmitted trait. In: *Endless Forms: Species and Speciation* . (1998), S. 404–422

[52] In: Guillaumin, Matthieu; Verbeek, Jakob; Schmid, Cordelia: *Multiple Instance Metric Learning from Automatically Labeled Bags of Faces.* Berlin, Heidelberg : Springer Berlin Heidelberg, 2010, S. 634–647

[53] Hao, Yuan; Campana, Bilson; Keogh, Eamonn: Monitoring and Mining Animal Sounds in Visual Space. In: *Journal of Insect Behavior* . 26 (2013), Nr. 4, S. 466–493

[54] Härmä, Aki: Automatic identification of bird species based on sinusoidal modeling of syllables. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2003* Bd. 5, 2003, S. 545–548

[55] Haussler, David: Learning Conjunctive Concepts in Structural Domains. In: *Machine Learning* . 4 (1989), Nr. 1, S. 7–40

[56] Hobson, Keith A.; Rempel, Robert S.; Greenwood, Hamilton; Turnbull, Brian; Van Wilgenburg, Steven L.: Acoustic surveys of birds using electronic recordings: new potential from an omnidirectional microphone system. In: *Wildlife Society Bulletin* . (2002), S. 709–720

[57] Hsu, Chih-Wei; Chang, Chih-Chung; Lin, Chih-Jen: A Practical Guide to Support Vector Classification. / Department of Computer Science, National Taiwan University. 2003. – Forschungsbericht

[58] Huang, Jin; Ling, Charles X.: Using AUC and accuracy in evaluating learning algorithms. In: *IEEE Transactions on Knowledge and Data Engineering* . 17 (2005), Nr. 3, S. 299–310

[59] Hunter, David R.; Lange, Kenneth: A tutorial on MM algorithms. In: *The Amer-*

ican Statistician . 58 (2004), Nr. 1, S. 30–37

[60] Jafari, Maria G.; Plumbley, Mark D.: Fast Dictionary Learning for Sparse Representations of Speech Signals. In: *IEEE Journal of Selected Topics in Signal Processing* . 5 (2011), S. 1025–1031

[61] Jain, A. K.; Duin, Robert P. W.; Mao, Jianchang: Statistical pattern recognition: a review. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* . 22 (2000), Nr. 1, S. 4–37

[62] Jančovič, Peter; Kküer, Münevver: Automatic detection and recognition of tonal bird sounds in noisy environments. In: *Eurasip Journal on Advances in Signal Processing* . 2011 (2011), S. 1–10

[63] Kasten, Eric P.; Gage, Stuart H.; Fox, Jordan; Joo, Wooyeong: The remote environmental assessment laboratory's acoustic library: An archive for studying soundscape ecology. In: *Ecological Informatics* . 12 (2012), S. 50 – 67

[64] Kasten, Eric P.; McKinley, Philip K.; Gage, Stuart H.: Ensemble extraction for classification and detection of bird species. In: *Ecological Informatics* . 5 (2010), Nr. 3, S. 153 – 166. – ISSN 1574–9541

[65] Keen, Sara; Ross, Jesse C.; Griffiths, Emily T.; Lanzone, Michael; Farnsworth, Andrew: A comparison of similarity-based approaches in the classification of flight calls of four species of North American wood-warblers (Parulidae). In: *Ecological Informatics* . 21 (2014), S. 25–33. – Ecological Acoustics. – ISSN 1574–9541

[66] Kim, Sang-woon; Duin, Robert P. W.: Dissimilarity-Based Classifications in Eigenspaces. In: San Martin, César (Hrsg.); Kim, Sang-Woon (Hrsg.): *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications SE - 50* Bd. 7042. Springer Berlin Heidelberg, 2011. – ISBN 978–3–642–25084–2, S. 425–432

[67] Kirschel, Alexander; Cody, Martin; Harlow, Zachary; Promponas, Vasilis; Vallejo, Edgar; Taylor, Charles: Territorial dynamics of Mexican Ant-thrushes Formicarius moniliger revealed by individual recognition of their songs. In: *Ibis* . 153 (2011), Nr. 2, S. 255–268

[68] Lakshminarayanan, Balaji; Raich, Raviv; Fern, Xiaoli: A Syllable-Level Probabilistic Framework for Bird Species Identification. In: *Proceedings of the Fourth International Conference on Machine Learning and Applications.* Los Alamitos, CA,

USA : IEEE Computer Society, 2009, S. 53–59

[69] Lee, Chang-Hsing; Han, Chin-Chuan; Chuang, Ching-Chien: Automatic Classification of Bird Species From Their Sounds Using Two-Dimensional Cepstral Coefficients. In: *IEEE Transactions on Audio, Speech, and Language Processing* . 16 (2008), November, Nr. 8, S. 1541 –1550

[70] Lee, Daniel D.; Seung, H. S.: Algorithms for Non-negative Matrix Factorization. In: Leen, T. K. (Hrsg.); Dietterich, T. G. (Hrsg.); Tresp, V. (Hrsg.): *Advances in Neural Information Processing Systems 13*. MIT Press, 2001, S. 556–562

[71] Lee, Honglak; Battle, Alexis; Raina, Rajat; Ng, Andrew Y.: Efficient sparse coding algorithms. In: *Advances in neural information processing systems*, 2006, S. 801–808

[72] Lee, Wan-Jui; Cheplygina, Veronika; Tax, David M. J.; Loog, Marco; Duin, Robert P. W.: Bridging Structure and Feature Representations in Graph Matching. In: *International Journal of Pattern Recognition and Artificial Intelligence* . 26 (2012), Nr. 5

[73] Li, Yan; Tax, David M. J.; Duin, Robert P. W.; Loog, Marco: Multiple-instance Learning as a Classifier Combining Problem. In: *Pattern Recognition* . 46 (2013), Nr. 3, S. 865–874. – ISSN 0031–3203

[74] Liu, Qingju; Wang, Wenwu; Jackson, Philip J B.; Barnard, Mark; Kittler, Josef; Chambers, Jonathon: Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking. In: *IEEE Transactions on Signal Processing* . 61 (2013), Nr. 22, S. 5520–5535

[75] Mairal, Julien; Bach, Francis; Ponce, Jean; Sapiro, Guillermo: Online Learning for Matrix Factorization and Sparse Coding. In: *Journal of Machine Learning Research* . 11 (2010), S. 19–60

[76] Maron, Oded; Lozano-Pérez, Tomás: A Framework for Multiple-Instance Learning. In: *Advances in neural information processing systems* . (1998), S. 570–576

[77] Marsh, David M.; Trenham, Peter C.: Current Trends in Plant and Animal Population Monitoring. In: *Conservation Biology* . 22 (2008), Nr. 3, S. 647–655

[78] Mason, R.; Roe, P.; Towsey, M.; Zhang, Jinglan; Gibson, J.; Gage, S.: Towards an Acoustic Environmental Observatory. In: *IEEE Fourth International Conference*

on eScience, 2008, S. 135–142

[79] Mennill, Daniel; Vehrencamp, Sandra: Context-dependent functions of avian duets revealed through microphone array recordings and multi-speaker playback. In: *Current Biology* . 18 (2008), S. 1314–1319

[80] Neal, L.; Briggs, F.; Raich, R.; Fern, X. Z.: Time-frequency segmentation of bird song in noisy acoustic environments. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011*, 2011, S. 2012 –2015

[81] Odom, Karan J.; Mennill, Daniel J.: A quantitative description of the vocalizations and vocal activity of the Barred Owl. In: *The Condor* . 112 (2010), Nr. 3, S. 549–560

[82] O'Grady, Paul D.; Pearlmutter, Barak A.: Convolutive Non-Negative Matrix Factorisation with a Sparseness Constraint. In: *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2006).* Maynooth, Ireland, September 2006, S. 427–432

[83] O'Grady, Paul D.; Pearlmutter, Barak A.: Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint. In: *Neurocomputing* . 72 (2008), Nr. 1-3, S. 88–101

[84] Otsu, Nobuyuki: A Threshold Selection Method from Gray-Level Histograms. In: *IEEE Transactions on Systems, Man and Cybernetics* . 9 (1979), Nr. 1, S. 62–66

[85] Owren, M. J.; Bernacki, R. H.: Animal Acoustic Communication: Sound Analysis and Research Methods, Springer Berlin Heidelberg, 1998, S. 129–162

[86] Ozerov, Alexey; Févotte, Cédric: Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. In: *IEEE Transactions on Audio, Speech, and Language Processing* . 18 (2010), Nr. 3, S. 550–563

[87] Pekalska, Elżbieta; Duin, Robert P. W.: Dissimilarity representations allow for building good classifiers. In: *Pattern Recognition Letters* . 23 (2002), Nr. 8, S. 943–956. – ISSN 01678655

[88] Plasencia-Calaña, Yenisel; Cheplygina, Veronika; Duin, Robert P. W.; García-Reyes, Edel B.; Orozco-Alzate, Mauricio; Tax, David M. J.; Loog, Marco: On the Informativeness of Asymmetric Dissimilarities. In: Hancock, Edwin (Hrsg.); Pelillo, Marcello (Hrsg.): *Similarity-Based Pattern and Recognition: Second International Workshop, SIMBAD 2013*, Springer, 2013, S. 75–89

[89] Potamitis, Ilyas: Automatic classification of a taxon-rich community recorded in the wild. In: *PLoS ONE* . 9 (2014), Nr. 5, S. 1–11

[90] Potamitis, Ilyas; Ntalampiras, Stavros; Jahn, Olaf; Riede, Klaus: Automatic bird sound detection in long real-field recordings: Applications and tools. In: *Applied Acoustics* . 80 (2014), S. 1 – 9

[91] Pourhomayoun, Mohammad; Dugan, Peter J.; Popescu, Marian; Clark, Christopher W.: Bioacoustic Signal Classification Based on Continuous Region Processing, Grid Masking and Artificial Neural Network. In: *CoRR* . abs/1305.3635 (2013)

[92] Rempel, Robert S.; Francis, Charles M.; Robinson, Jeffrey N.; Campbell, Margaret: Comparison of audio recording system performance for detecting and monitoring songbirds. In: *Journal of Field Ornithology* . 84 (2013), Nr. 1, S. 86–97

[93] Romberg, Justin: Imaging via compressive sampling [introduction to compressive sampling and recovery via convex programming]. In: *IEEE Signal Processing Magazine* . 25 (2008), Nr. 2, S. 14–20

[94] Ross, Jesse C.; Allen, Paul E.: Random Forest for improved analysis efficiency in passive acoustic monitoring. In: *Ecological Informatics* . 21 (2014), S. 34 – 39. – Ecological Acoustics. – ISSN 1574–9541

[95] Ruiz-Muñoz, José F.; Castellanos-Dominguez, German; Orozco-Alzate, Mauricio: Enhancing the dissimilarity-based classification of birdsong recordings. In: *Ecological Informatics* . 33 (2016), S. 75 – 84. – ISSN 1574–9541

[96] Ruiz-Muñoz, J. F.; You, Zeyu; Raich, Raviv; Fern, Xiaoli Z.: Dictionary Learning for Bioacoustics Monitoring with Applications to Species Classification. In: *Journal of Signal Processing Systems* . (2016), S. 1–15. – ISSN 1939–8115

[97] Ruiz-Munoz, José F.; You, Zeyu; Raich, Raviv; Fern, Xiaoli Z.: Dictionary extraction from a collection of spectrograms for bioacoustics monitoring. In: *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on.* IEEE, 2015, S. 1–6

[98] Sawada, Hideyuki; Kameoka, Hirokazu; Araki, Shunsuke; Ueda, Naonori: Multichannel extensions of non-negative matrix factorization with complex-valued data. In: *IEEE Transactions on Audio, Speech, and Language Processing* . 21 (2013), Nr. 5, S. 971–982

[99] Scott, Stephen; Zhang, J U N.; Brown, Joshua: On generalized multiple-instance learning. In: *International Journal of Computational Intelligence and Applications* . 5 (2005), Nr. 1, S. 21–35

[100] Sebastián-González, Esther; Pang-Ching, Joshua; Barbosa, Jomar M.; Hart, Patrick: Bioacoustics for species management: two case studies with a Hawaiian forest bird. In: *Ecology and Evolution* . 5 (2015), Nr. 20, S. 4696–4705

[101] Smaragdis, Paris: Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In: *Independent Component Analysis and Blind Signal Separation.* Springer, 2004, S. 494–499

[102] Smaragdis, Paris: Convolutive speech bases and their application to supervised speech separation. In: *IEEE Transactions on Audio, Speech, and Language Processing* . 15 (2007), Nr. 1, S. 1–12

[103] Smaragdis, Paris; Fevotte, Cedric; Mysore, Gautham J.; Mohammadiha, Nasser; Hoffman, Matthew: Static and Dynamic Source Separation Using Non-negative Factorizations: A unified view. In: *IEEE Signal Processing Magazine* . 31 (2014), Nr. 3, S. 66–75

[104] Somervuo, P.; Härmä, Aki; Fagerlund, Seppo: Parametric Representations of Bird Sounds for Automatic Species Recognition. In: *IEEE Transactions on Audio, Speech, and Language Processing* . 14 (2006), November, Nr. 6, S. 2252 –2263

[105] Stowell, Dan; Plumbley, Mark D.: Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. In: *PeerJ* . 2 (2014), Nr. e488, S. 1–31. – ISSN 2167–8359

[106] Strang, Gilbert: A proposal for Toeplitz matrix calculations. In: *Studies in Applied Mathematics* . 74 (1986), Nr. 2, S. 171–176

[107] Tax, David M. J.; Loog, Marco; Duin, Robert P. W.; Cheplygina, Veronika; Lee, Wan-Jui: Bag Dissimilarities for Multiple Instance Learning. In: Pelillo, Marcello (Hrsg.); Hancock, Edwin R. (Hrsg.): *Similarity-Based Pattern Recognition* Bd. 7005. Springer, 2011, S. 222–234

[108] Tax, David M.: MIL, A Matlab Toolbox for Multiple Instance Learning, 2013. – version 0.8.1

[109] Ten Cate, Carel: Birdsong and Evolution. In: *Nature's music: the science of*

birdsong. 2004, Kapitel 10, S. 296–317

[110] Trifa, Vlad M.; Kirschel, Alexander N. G.; Taylor, Charles E.; Vallejo, Edgar E.: Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models. In: *The Journal of the Acoustical Society of America* . 123 (2008), Nr. 4, S. 2424–2431

[111] Venier, Lisa A.; Holmes, Stephen B.; Holborn, George W.; Mcilwrick, Kenneth A.; Brown, Glen: Evaluation of an automated recording device for monitoring forest birds. In: *Wildlife Society Bulletin* . 36 (2012), Nr. 1, S. 30–39

[112] Ventura, Thiago M.; de Oliveira, Allan G.; Ganchev, Todor D.; de Figueiredo, Josiel M.; Jahn, Olaf; Marques, Marinez I.; Schuchmann, Karl-L.: Audio parameterization with robust frame selection for improved bird identification. In: *Expert Systems with Applications* . 42 (2015), Nr. 22, S. 8463 – 8471. – ISSN 0957–4174

[113] Wang, Dong; Vipperla, Ravichander; Evans, Nicholas; Zheng, Thomas F.: Online non-negative convolutive pattern learning for speech signals. In: *IEEE Transactions on Signal Processing* . 61 (2013), Nr. 1, S. 44–56

[114] Wang, Dong; Vipperla, Ravichander; Evans, Nicholas W D: Online pattern learning for non-negative convolutive sparse coding. In: *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication, August 28-31, Florence, Italy*, 2011

[115] Wang, Fei; Li, Ping: Efficient Nonnegative Matrix Factorization with Random Projections. In: *Proceedings of the 2010 SIAM International Conference on Data Mining*, 2010, S. 281–292

[116] Wang, Jigang; Neskovic, Predrag; Cooper, Leon N.: Improving nearest neighbor rule with a simple adaptive distance measure. In: *Pattern Recognition Letters* . 28 (2007), Nr. 2, S. 207–213. – ISSN 01678655

[117] Wang, Wenwu: Convolutive non-negative sparse coding. In: *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, S. 3681–3684

[118] Wang, Yu-Xiong; Zhang, Yu-Jin: Nonnegative matrix factorization: A comprehensive review. In: *Knowledge and Data Engineering, IEEE Transactions on* . 25 (2013), Nr. 6, S. 1336–1353

[119] Wimmer, Jason; Towsey, Michael; Planitz, Birgit; Roe, Paul; Williamson, Ian: Scaling acoustic data analysis through collaboration and automation. In: *Sixth IEEE International Conference on e-Science*. IEEE, 2010, S. 308–315

[120] Yaghoobi, M.; Blumensath, T.; Davies, M.E.: Dictionary Learning for Sparse Approximations With the Majorization Method. In: *Signal Processing, IEEE Transactions on* . 57 (2009), Nr. 6, S. 2178–2191. – ISSN 1053–587X

[121] Yeh, Chin-Chia M.; Yang, Yi-Hsuan: Supervised Dictionary Learning for Music Genre Classification. In: *Proceedings of the 2Nd ACM International Conference on Multimedia Retrieval*. New York, NY, USA : ACM, 2012. (ICMR '12). – ISBN 978–1–4503–1329–2, S. 55:1–55:8

[122] Zakaria, Jesin; Rotschafer, Sarah; Mueen, Abdullah; Razak, Khaleel; Keogh, Eamonn: Mining Massive Archives of Mice Sounds with Symbolized Representations. In: *SIAM International Conference on Data Mining (SDM)*, 2012, S. 588–599

[123] Zhang, Min-Ling; Zhou, Zhi-Hua: A review on multi-label learning algorithms. In: *IEEE Transactions on Knowledge and Data Engineering* . 26 (2014), Nr. 8, S. 1819–1837

[124] Zhou, Guoxu; Cichocki, Andrzej; Xie, Shengli: Fast nonnegative matrix/tensor factorization based on low-rank approximation. In: *IEEE Transactions on Signal Processing* . 60 (2012), Nr. 6, S. 2928–2940