information

Article

Visual Active Learning for Labeling: A Case for Soundscape Ecology Data

Liz Huancapaza Hilasaca [1], Milton Cezar Ribeiro [2] and Rosane Minghim [3,*]

[1] Department of Computer Science, University of São Paulo, Sao Carlos, SP 13566590, Brazil; lizhh@usp.br
[2] Biodiversity Department, São Paulo State University—UNESP, Rio Claro, SP 13506900, Brazil; milton.c.ribeiro@unesp.br
[3] School of Computer Science and Information Technology, University College Cork, T12 XF62 Cork, Ireland
* Correspondence: r.minghim@cs.ucc.ie

Abstract: Labeling of samples is a recurrent and time-consuming task in data analysis and machine learning and yet generally overlooked in terms of visual analytics approaches to improve the process. As the number of tailored applications of learning models increases, it is crucial that more effective approaches to labeling are developed. In this paper, we report the development of a methodology and a framework to support labeling, with an application case as background. The methodology performs visual active learning and label propagation with 2D embeddings as layouts to achieve faster and interactive labeling of samples. The framework is realized through SoundscapeX, a tool to support labeling in soundscape ecology data. We have applied the framework to a set of audio recordings collected for a Long Term Ecological Research Project in the Cantareira-Mantiqueira Corridor (LTER CCM), localized in the transition between northeastern São Paulo state and southern Minas Gerais state in Brazil. We employed a pre-label data set of groups of animals to test the efficacy of the approach. The results showed the best accuracy at 94.58% in the prediction of labeling for birds and insects; and 91.09% for the prediction of the sound event as frogs and insects.

Keywords: active learning; sampling; clustering; soundscape ecology; visualization; labeling

Citation: Hilasaca, L.H.; Ribeiro, M.C.; Minghim, R. Visual Active Learning for Labeling: A Case for Soundscape Ecology Data. *Information* **2021**, *12*, 265. https://doi.org/10.3390/info12070265

Academic Editor: Fernando V. Paulovich

Received: 1 June 2021
Accepted: 22 June 2021
Published: 29 June 2021

1. Introduction

Active Learning (AL) is considered as a special case of machine learning [1]. AL is also called "query learning" because it actively requests information of selected data from the set of unlabeled data, from which the model will learn. It is widely used in hard cases for big data learning, an example of which is evidenced by [2] when data have only 2000 labeled instances and 250,000 unlabeled instances. There are different strategies used for each stage AL, such as those described and categorized by [1,3,4]. The effectiveness of AL was demonstrated in various tasks, such as: (a) automatic speech recognition [5]; (b) classification of voicemail messages [6]; (c) malicious code detection [7]; (d) text classification [8]; (e) speech emotion classification [9]; and (f) audio retrieval [10]. When the user takes part in the process, there is a cooperation between computer and analyst, and it becomes part of the Human In The Loop (HITL) machine learning paradigm [11], targeted at improving learning processes by employing the user's expertise as well as computing strategies.

The main goal of AL is that the algorithm learns using the smallest possible number of instances during training, but generating the best predictions of unlabeled data [1]. There are several sampling strategies, but at least three categories can be generalized according to [1]: (1) certainty-based sampling, (2) query-by-committee, and (3) expected error reduction. In the first sampling strategy, a small set of selected samples is annotated at the beginning, and then manually annotated labels are used to train a classifier, which classifies unlabeled samples. The second type of sampling strategy involves two or more classifiers, who may disagree with respect to some instances; if they agree, those instances are delivered for human annotation or validation. The third type of strategy aims to

estimate and select the instances that can have a high impact on the expected model error for human annotation; however, this last strategy may be the most computationally expensive.

The final goal of Active Learning is to learn from a small set of labeled samples. However, a large number of current applications require a large subset of labelled data from the start. Examples are the building of certain machine learning models, such as those applied in deep learning, and applications targeted at selecting and reducing the set of attributes to represent a phenomenon (examples are certain biological data sets, medical records, fraud detection, etc.). Our aim in this work is to find a stable balance of the active learning and label propagation strategies to support the actual labeling process, so that the stage of the data science pipeline can be made more effective.

In order to allow user participation in the labeling processes, visualization tools are necessary, since they can help the user navigate in large data with multiple attributes, by facilitating data interpretation through graphical and interactive representations. Multidimensional projections or embeddings allow the reduction of a multidimensional space of m dimensions to another reduced space with p dimensions $(X^m \Rightarrow X^p, p < m)$, while trying to preserve as much information as possible. The AL approach together with visualization techniques can facilitate the acquisition of knowledge from the data and support interpretation as well as sample labeling by the user [12,13].

In this study, we propose a method of user centered active learning that aims to optimize and give more dynamism to the entire process of labeling, by including visual strategies in the several stages of AL. In support of the strategy, we have compared different types of sampling strategies and estimated the appropriate number of samples that must be labeled by the annotator. Thus, we also performed an exhaustive assessment of learning power and performance of the models, in relation to the different sampling strategies and the defined number of clusters. We have also reflected on how the model learns based on a dataset with numerous features against and the same dataset taking as a starting point known discriminative features (as illustrated in Figure 4). After data samples are annotated by expert users, a classification model is trained with the recorded data. Finally, the predictions of the labels for the data set are visualized and evaluated (see Table 1). The projections afford visual analysis and interaction in each stage of the active learning process to give the user a better understanding of the data and of the process of labeling. (see Figure 5). Users can confirm or correct predicted labels. Therefore, this is a visual analytics approach that is meant to blend learning and user supervision in the process of labeling samples in general and in soundscape ecology data in particular (see Figures 3 and 5).

We realize the framework and its methods in the context of environment monitoring using sound recordings. Recordings are becoming central in understanding and describing the condition of natural environments. They are involved in a large number of studies, such as monitoring environmental noise [14], measuring biodiversity integrity and environmental health [15–17], freshwater lakes [18,19], and species identification [20]. Soundscape ecology studies [21] have increased in the recent years [22] and, between other things, this research field aims to monitor and understand how different environments respond to changes induced by human activities [23–25] and to assess the impact through altered soundscapes.

Extracting information from such data is both very challenging and expensive. Therefore, it is necessary build a bridge between ecoacoustics, machine learning, and visualization, which requires a multidisciplinary approach [26]. One important phase of soundscape ecology studies is the annotation or tagging of events of interest [27]. Therefore, developing new methods that improve the quality of labeling is essential for the success of many soundscape ecology studies worldwide. Here, we test our framework on the task of labeling soundscape ecology data and employing real data on birds, frogs, and insects to evaluate the performance of the method.

The main contributions of this work are summarized as follows:

- Proposal, implementation, and testing of a Visual Active Learning strategy for labeling and application of such strategy to soundscape ecology data;
- Anchoring the process of user centered labeling through visualization by multidimensional projections. Anchoring data detailing in visualizations of summarized data; in the case of our sample application that is done by the proposal of "Time Line Spectrogram" (TLS) visualizations (Figure 3).
- New sampling strategies incorporated in the process of Active Learning for labeling, and their evaluation (Figure 4) using this strategy, we demonstrate reduced annotation costs (Table 1);
- Presentation of extensive experimental results that demonstrate the relevance of the proposed methods (Figures 3–5 and Table 1).

Table 1. Accuracy results (in percent 100%) of the labeling for the DS1 set with the best 30 features: each line represents a different number of clusters (k). In the columns from left to right, three groups of columns are indicated (pk), which refers to the sample numbers per clustering (5, 10 e $\sqrt{|.|}$). In each block of columns from left to right are indicated: (1) the total number of samples (p); (2) the percentage of samples (%) in relation to the total number of instances in the data set; and (3) the accuracy for each of the sampling strategies (*random* (r), *medoid* (m), *contour* (c)) and their combinations rm, rc, mc e rmc. In the heatmap, the yellow color represents the maximum accuracy values achieved (labels prediction), the blue color represents the minimum values, the intermediate accuracy values are visualized with a color gradient between the yellow, orange, and blue colors.

k	$pk=5$									$pk=10$									$pk=\sqrt{\lvert.\rvert}$							
	(p)	(%)	(r)	(m)	(c)	(rm)	(rc)	(mc)	(rmc)	(p)	(%)	(r)	(m)	(c)	(rm)	(rc)	(mc)	(rmc)	(p)	(%)	(r)	(m)	(c)	(rm)	(rc)	(mc)	(rmc)
5	25	1.10	62.79	58.30	32.50	59.64	47.38	58.53	62.79	50	2.20	64.66	64.12	31.88	60.66	58.69	53.84	58.15	103	4.52	67.25	62.79	34.64	66.65	62.93	55.24	57.41
6	30	1.32	65.24	62.44	31.51	60.21	43.21	59.28	60.17	60	2.64	64.10	63.78	33.60	62.07	58.05	52.64	58.37	112	4.92	64.43	63.74	40.05	67.48	69.24	55.66	57.14
7	35	1.54	71.19	61.46	38.85	67.53	59.28	62.36	62.62	70	3.07	71.95	60.67	37.20	66.38	65.43	57.59	63.25	119	5.23	72.85	63.30	37.26	68.40	67.98	55.28	61.03
8	40	1.76	65.62	61.42	37.82	65.53	59.10	64.68	64.33	80	3.51	67.96	62.90	32.13	64.54	67.05	60.31	61.72	126	5.53	67.55	63.69	39.01	68.62	68.06	61.13	64.53
9	45	1.98	60.26	64.61	40.14	62.46	64.25	66.71	66.58	90	3.95	67.08	64.61	33.74	68.54	68.22	65.11	65.57	136	5.97	69.87	64.27	41.71	71.18	63.80	64.83	66.14
10	50	2.20	63.36	64.89	43.65	62.73	66.91	67.85	67.49	100	4.39	70.56	64.17	40.19	66.65	68.08	67.98	65.55	142	6.24	71.52	65.43	48.01	73.40	71.05	68.99	68.29
11	55	2.42	64.13	62.15	46.98	65.84	67.96	66.61	65.98	110	4.83	68.76	63.54	43.79	68.94	70.14	67.74	65.39	149	6.54	73.50	64.29	43.70	71.95	69.41	69.08	66.17
12	60	2.64	70.55	60.62	45.20	67.34	68.20	63.78	65.31	120	5.27	75.15	62.87	46.13	64.72	71.12	67.45	64.86	155	6.81	76.48	64.89	49.62	69.32	72.20	67.86	66.21
13	65	2.85	67.18	61.35	44.71	66.59	60.67	64.87	67.09	130	5.71	68.70	62.55	50.30	67.54	68.37	68.70	65.16	160	7.03	70.00	64.38	60.23	68.02	70.15	69.25	66.98
14	70	3.07	69.87	60.17	58.13	67.20	66.38	65.84	65.29	140	6.15	69.91	62.56	60.60	68.79	68.69	68.65	64.25	166	7.29	73.14	62.72	59.02	66.51	76.69	70.82	67.69
15	75	3.29	68.94	61.22	54.13	66.17	65.67	63.40	63.90	150	6.59	71.98	59.71	58.44	68.97	71.37	69.44	65.26	171	7.51	72.98	62.68	58.97	70.85	73.12	69.94	67.57
16	80	3.51	66.64	61.36	56.94	66.18	67.59	64.54	63.31	160	7.03	70.81	62.21	57.96	67.45	73.50	69.49	67.60	177	7.77	70.48	64.38	60.86	71.10	75.24	69.95	69.14
17	85	3.73	67.38	62.55	61.82	69.21	68.48	67.47	65.88	170	7.47	71.95	62.70	59.80	68.15	71.86	71.29	67.68	180	7.91	73.72	63.61	57.51	71.01	73.06	71.44	69.19
18	90	3.95	66.94	62.87	59.35	68.27	67.86	68.77	64.29	180	7.91	72.96	63.09	60.75	71.63	72.10	71.39	68.43	187	8.21	73.25	64.31	62.54	67.13	72.87	70.29	68.04
19	95	4.17	69.43	62.74	61.18	67.51	68.65	68.70	66.68	190	8.34	73.84	63.44	61.04	70.96	73.74	71.01	68.28	190	8.34	74.89	65.26	62.48	71.35	72.40	70.29	69.38
20	100	4.39	70.56	63.30	57.83	67.98	63.99	69.13	68.26	200	8.78	72.85	64.52	63.31	72.22	73.57	73.18	72.32	196	8.61	72.80	64.44	62.09	70.98	74.20	72.03	70.40
21	105	4.61	67.50	63.49	61.60	66.44	69.11	69.20	66.85	210	9.22	72.57	64.25	61.44	71.02	71.46	73.39	72.18	202	8.87	70.80	64.00	58.31	69.54	73.78	72.96	71.28
22	110	4.83	69.59	62.67	64.28	67.10	68.90	68.44	67.01	220	9.66	76.52	63.93	62.08	71.66	75.79	71.85	69.42	205	9.00	73.94	64.43	57.77	72.39	75.87	71.43	69.88
23	115	5.05	69.29	61.89	62.86	65.54	70.40	68.13	65.82	230	10.10	72.74	62.77	61.65	71.18	76.80	72.79	71.52	208	9.13	72.21	64.52	58.05	71.53	73.66	71.58	71.00
24	120	5.27	70.47	62.31	63.65	67.22	72.60	68.47	66.34	240	10.54	72.46	63.48	65.44	71.18	74.37	73.10	71.28	213	9.35	74.13	64.58	63.81	71.32	75.53	72.29	70.69
25	125	5.49	71.28	63.48	64.92	64.78	69.56	68.77	67.80	250	10.98	74.49	65.02	66.01	70.35	76.76	72.72	71.88	216	9.49	73.70	64.68	60.89	69.00	76.71	72.68	72.97
26	130	5.71	71.50	64.28	62.23	69.21	71.63	70.42	67.26	260	11.42	73.13	66.83	64.95	70.70	75.31	74.62	73.33	220	9.66	73.94	66.80	61.84	69.52	75.55	73.65	71.46
27	135	5.93	71.99	64.43	61.06	67.04	69.33	69.09	66.29	270	11.86	73.39	66.72	65.62	70.30	76.23	73.59	71.10	227	9.97	72.59	65.66	61.41	72.73	74.54	73.07	71.41
28	140	6.15	71.31	64.16	60.93	69.35	69.07	70.47	67.01	280	12.30	73.46	66.80	64.15	73.56	74.71	73.86	71.71	230	10.10	72.50	66.05	62.29	72.69	74.50	73.82	70.35
29	145	6.37	73.26	64.59	58.72	65.95	71.62	70.73	67.68	290	12.74	75.74	66.53	64.47	73.28	76.60	73.73	71.97	235	10.32	75.37	64.84	62.10	69.59	75.81	72.97	70.23
30	150	6.59	73.25	64.27	64.17	67.47	72.21	69.25	68.36	300	13.18	75.37	66.26	66.87	70.97	77.14	74.00	73.88	237	10.41	73.48	66.96	63.43	71.27	74.85	71.76	69.56
31	155	6.81	72.01	63.81	64.14	69.93	74.46	70.45	68.94	310	13.61	72.70	67.16	66.50	71.73	75.19	73.97	73.84	241	10.58	73.04	66.70	65.28	72.59	73.04	73.43	71.41
32	160	7.03	73.36	64.67	66.23	69.44	71.99	70.29	68.63	320	14.05	77.11	67.25	68.73	72.15	77.72	73.89	72.84	245	10.76	75.84	67.47	64.22	73.13	76.28	73.38	72.59
33	165	7.25	72.02	65.34	65.53	66.76	74.76	70.74	69.74	330	14.49	74.47	67.28	67.95	72.37	77.40	75.30	74.65	249	10.94	74.51	66.96	67.50	71.65	72.68	73.82	73.67
34	170	7.47	72.95	65.54	67.44	70.72	76.70	70.00	70.53	340	14.93	77.59	66.91	68.97	71.66	75.32	75.12	74.57	254	11.16	75.43	66.58	69.25	72.96	78.25	72.66	72.86
35	175	7.69	74.41	65.08	67.98	72.36	72.84	69.36	69.89	350	15.37	76.49	68.19	71.56	74.26	78.31	75.30	74.13	257	11.29	74.60	67.67	70.25	74.01	75.54	74.50	71.63
36	180	7.91	72.53	64.43	68.14	72.25	71.20	70.15	69.19	360	15.81	76.21	66.82	71.36	72.98	77.78	75.59	75.14	260	11.42	73.67	67.48	71.00	70.90	77.94	73.28	70.80
37	185	8.12	75.05	66.68	66.30	72.28	75.00	71.94	70.98	370	16.25	77.24	69.01	70.79	74.04	78.66	75.62	74.99	262	11.51	75.88	68.44	73.15	73.10	75.93	73.40	72.16
38	190	8.34	75.13	64.64	66.94	71.63	73.26	71.73	70.92	380	16.69	76.70	68.90	72.32	75.28	76.07	75.12	75.18	264	11.59	74.42	68.21	71.98	75.56	74.96	72.78	72.48
39	195	8.56	73.15	64.94	66.71	74.02	75.22	71.85	71.66	390	17.13	77.16	68.73	70.48	74.03	76.31	74.46	74.31	268	11.77	75.96	68.19	70.68	72.62	77.20	72.32	71.38
40	200	8.78	73.95	64.28	67.60	75.16	75.16	71.98	71.27	399	17.52	79.07	68.00	70.02	73.48	76.89	74.81	73.54	271	11.90	77.52	66.15	71.44	75.62	74.63	72.78	72.38
41	205	9.00	71.72	64.29	67.37	71.09	76.11	72.10	70.24	409	17.96	77.46	67.93	71.15	75.54	76.28	74.20	74.57	274	12.03	75.09	67.30	70.00	72.74	74.84	73.34	71.24
42	210	9.22	71.60	64.97	65.99	70.39	76.15	72.09	70.16	419	18.40	76.64	69.05	69.54	71.91	78.53	74.87	74.49	277	12.17	74.35	67.95	68.75	72.60	77.60	75.30	71.85
43	215	9.44	75.46	66.25	66.88	72.41	75.56	71.92	69.95	429	18.84	78.35	68.72	69.05	72.73	78.90	74.57	73.76	280	12.30	74.81	67.80	70.36	74.01	77.52	73.86	70.56
44	220	9.66	75.21	66.65	67.87	72.78	73.36	71.90	70.12	439	19.28	79.87	68.99	69.91	75.57	77.04	74.92	75.20	283	12.43	77.03	67.30	69.41	72.92	75.63	74.02	72.77
45	225	9.88	74.22	65.98	69.01	71.30	73.68	72.95	70.04	449	19.72	77.24	68.54	70.46	72.92	76.15	75.44	75.00	286	12.56	73.93	67.40	71.52	73.73	76.34	74.23	71.52

Table 1. *Cont.*

k	$pk = 5$									$pk = 10$									$pk = \sqrt{\lvert . \rvert}$							
	(*p*)	(%)	(*r*)	(*m*)	(*c*)	(*rm*)	(*rc*)	(*mc*)	(*rmc*)	(*p*)	(%)	(*r*)	(*m*)	(*c*)	(*rm*)	(*rc*)	(*mc*)	(*rmc*)	(*p*)	(%)	(*r*)	(*m*)	(*c*)	(*rm*)	(*rc*)	(*mc*)	(*rmc*)
46	230	10.10	75.38	65.51	69.47	74.26	76.50	71.81	70.21	459	20.16	78.49	68.04	70.52	77.06	77.67	74.70	74.51	288	12.65	76.62	67.67	70.29	73.81	74.56	74.76	72.40
47	235	10.32	73.90	65.38	69.83	70.96	75.17	71.35	69.60	469	20.60	76.55	68.36	70.58	75.28	77.43	75.77	74.51	290	12.74	74.94	67.04	71.36	74.13	75.84	72.97	71.62
48	240	10.54	75.75	65.93	68.53	71.87	75.11	72.12	70.36	479	21.04	77.25	68.91	70.69	74.53	76.25	75.36	74.10	293	12.87	76.41	66.99	71.07	73.74	75.35	74.90	72.18
49	245	10.76	75.69	66.19	69.14	72.15	77.46	72.44	70.39	489	21.48	76.68	68.40	69.30	74.55	79.53	77.46	75.86	296	13.00	75.42	66.33	72.34	75.42	76.73	75.27	72.69
50	250	10.98	73.06	67.00	68.18	74.10	74.99	72.72	71.15	499	21.91	79.53	70.13	70.36	75.70	78.40	76.43	76.11	300	13.18	76.53	67.17	70.21	73.60	77.79	74.51	72.23

2. Background

2.1. Visualization in Active Learning and Labeling

Visualization in support of active learning strategies has been observed for some time in a variety of applications such as image processing [28,29]. Years later, the work [30] presented *Visalix* as a good alternative that also combines AL and visualization, where the user can make the annotation of classes by managing the attributes in a 3D space, but with certain limitations. Reference [31] also contributed in the field by presenting the Case Base Topology Viewer for Active Learning (CBTV-AL), where they considered density, uncertainty, and diversity in the sampling strategies. In the case of diversity and uncertainty, the CBTV-AL needs to be recalculated until the last instance can be labeled. CBTV-AL allows for visualizing the entire labeling process through layouts based on force-directed graph drawing algorithms. The authors emphasize the importance of visualization techniques for user participation in AL. Reference [12] enhances sample selection in AL based on visualization using scatter plots and iso-contours, employing the semi-supervised metric learning method to train with data annotated by the user. Reference [32] presents a pilot study using the t-Distributed Stochastic Neighbor Embedding (t-SNE) [33], force-directed graph layout, and chord diagrams to visualize data and facilitate labeling. According to the authors, this improves the text document labeling process. In these last two approaches, the determination of samples is carried out manually on the visualization (scatter plot) by the user, and this leads to the conclusion that there is no sample selection and AL suggestion strategy directly involved as described by [13]. In addition, there are other strategies as incrementally involved within active learning models as can be seen in the approaches of [34–36].

Here, we present an approach to predict labels using label propagation by sample selection over a clustering process. The visualizations in our case are based on multidimensional projections, which aim to map data in 2D based on their similarity, giving an extra layer of effectiveness of the current stage of the labeling. In comparison to previous methods, we contribute both in terms of the interaction method and on the sample selection process.

To test the method, we built a framework for labeling of acoustic landscapes in soundcape ecology. We have employed a data set for animal group identification whose discriminant features we studied before [37] to evaluate the impact on results. Our results prove efficiency in the prediction task of labels and show reduced manual annotation effort with the methodology proposed for soundscape data.

2.2. Labeling of Sound Data

Sound data labeling is a key task that, in general, precedes most of the remaining data analysis tasks or the development of new approaches to automatic interpretation. For labeling sound data, there are some computer-based solutions, such as the ones presented recently by [27]. Manual labeling is very costly in time and degree of complexity, and studies in AL for labeling are meant to minimize manual annotation effort of samples by the users. In line with this objective, the study conducted by [38] addresses a methodology based on the combination of AL and self-training by considering the level of confidence of instances. Thus, instances with low confidence scores are delivered to expert users to be labeled and instances with high confidence scores are used in the prediction automatically. Because it is based on defining the scores, the determination of the confidence threshold is

crucial. Reference [38] used the FindSounds (FindSounds: https://www.findsounds.com/, accessed on 22 June 2021) database with duration ranging from 1 to 10 s for each instance of audio. Reference [39] proposed a new medoids-based active learning (MAL) method for generating clusters by K-Medoids; afterwards, the medoids from each cluster are presented to expert users and labeled by them. After obtaining the labels, instances are fed to a classifie. Reference [39] used the UrbanSound8k [40] database with maximum duration of four seconds for each instance of audio. Another study conducted by [41] proposed an AL strategy that combines two stages: (i) The first stage implements the same strategy as [39]; (ii) In the second stage, they proposed a selection of samples based on the prediction mismatch, looking mainly for segments with incorrect labels; labels are then corrected and predictions are updated in the groupings using a nearest-neighbor approach. Reference [42] presents an AL-based framework for classifying soundscape recordings. According to the methodology, the authors first generate 60 clusters, then randomly selected 10 instances of each cluster to be manually labeled. It allowed to define the most appropriate class names for each cluster. The reduced manual annotation effort with the active learning methodology in the paper was demonstrated empirically. Reference [42] used a database with maximum duration of 1 min for each instance of audio. As presented above, several venues can be pursued by researchers when advancing the labeling task, with a varying degree of automation and confidence. On the other handl, Reference [43] proposed an active learning system for detecting sound events where the main objective is to improve the process of selection of samples based on the identification of audio segments with the presence of sound activity for annotation; and, in the same line of work, reference [44] proposed a new strategy to determine samples from the audio database, thus asserting the high influence of this sample selection task in reducing the manual labeling effort by the user.

Our approach focuses on displaying the data as submitted to several strategies or techniques for each stage of active learning process, with the main goal of labeling data with improved accuracy, but with less effort by users than with manual labeling. Here, we propose a visual analytics approach that is meant to blend learning and user supervision in the process of labeling samples in general and in soundscape ecology in particular. The goals are: (i) include the learning and supervision of the user in the model for labeling; (ii) know what the appropriate strategies or techniques are in this labeling process; and (iii) support evaluating the quality of labeling to soundscape data.

3. The Labeling Method

In Figure 1, we illustrate the main steps of our proposed method for labeling strategy. Given a data set that has no labels, the Clustering stage is initially undertaken. Then, samples are extracted from each cluster at the stage named Sampling. Expert users interact with the samples by listening and labeling the audios in the Annotation step. Considering the samples labeled by the user, a learning model (classifier) is trained. Then, the model is then used to predict the rest of the labels of the instances of data set (excluding the samples), and these predictions are performed in the step learning–prediction. The visualization task is intrinsically present in all stages. Finally, the results obtained with the proposed method are evaluated. The steps of each stage of our method are given in the Algorithm 1, and each step is described in the next subsections. We implemented our method as a framework with application in soundscape ecology. A description of the interface of the system is shown in Appendix A.

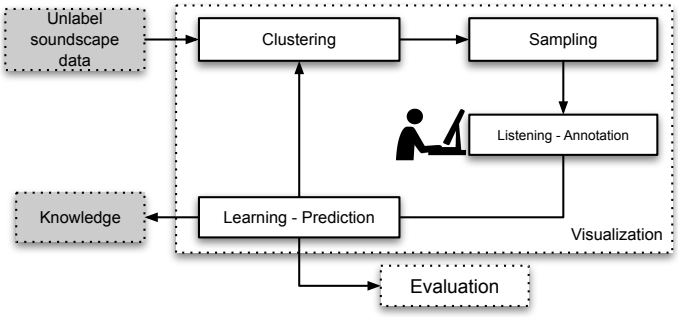


Figure 1. Framework of the proposed method: visual active learning for labeling data.

Algorithm 1: Graph building.

Input : *data*: unlabeled set of sound files from soundscape recording; *model_clust*: clustering method ex. *KM* ou *HAC*; *k*: number of clustering; *model_smp*: sampling type ex. *r*, *m*, *c*, *rm*, *rc*, *mc* ou *rmc*; *size_smp*: number of samples per cluster; *model_learn*: learning model ex. *RFC*, *SVC*, *KNNC* ou *XBGC*;

Output: $\widehat{data}$: labeled data, where the labels represent the segregated event categories.

$repeat \leftarrow True$;
while $repeat == True$ **do**
  $X \leftarrow$ Features($data$);
  $Y \leftarrow$ Labels($data$);
  `/* stage (1)` `*/`
  $clusters \leftarrow$ Clustering($model\_clust, X, k$);
  `/* stage (2)` `*/`
  $samples \leftarrow$ Sampling($clusters, model\_smp, size\_smp$);
  $unsamples \leftarrow (X - samples)$;
  `/* stage (3)` `*/`
  $view\_p \leftarrow$ Projection($X, samples$);
  $view\_s \leftarrow$ TimeLineSpectrogram($X, samples$);
  $Y_{samples} \leftarrow$ Annotation($view\_p, view\_s,$);
  `/* stage (4)` `*/`
  Learing($model\_learn, X_{samples}, Y_{samples}$);
  $y_{unsamples} \leftarrow$ Prediction($model\_learn, X_{unsamples}$);
  UpdadeLabels($view, Y_{unsamples}$);
  $repeat \leftarrow$ UserVisualEvaluation($view$);

3.1. Clustering

To start this stage, a data set of unlabeled instances for which features have been extracted is required. In the case of soundscape data, numerical features are computed from the audio itself and from the image of the audio spectrogram. In general, the input to clustering is only the features extracted for all instances. Firstly, k clusters are extracted from the data employing the Euclidean distance. Hierarchical Agglomerative Clustering and K-Means were used in our framework, both using the Scikit-learn package (available in scikit-learn: https://scikit-learn.org/, accessed on 22 June 2021). We consider that, through clustering, it will be possible to identify patterns between the instances of audio data, and, consequently, these patterns will allow the segregation of the event categories from the soundscape and produce good samples for labeling.

3.2. Sampling

The main goal of the sampling step is to extract additional representative samples from the data set, to be used later in learning tasks. Thus, p samples are extracted from each of the clusters, employing for this the following sample extraction methods: random (r), medoid (m) and contour (c); as well as their combinations: (random-medoid (rm), random-contour (rc), medoid-contour (mc) and random-medid-contour (rmc)). In the first

method, samples are taken randomly. In the case of medoid, samples are the instances closest to the cluster centroid. The method contour takes samples furthest from the centroid of the clusters. Figure 2 illustrates the three types of sampling methods.

3.3. Annotation

In this step, the induction of learning from the AL paradigm is initiated, through the interaction of expert users. The goal of this step is to abstract information from users who are experts in recognizing and differentiating sound categories. Thus, this stage deals with the tasks of listening and labeling audio files corresponding to the most representative samples. To perform these tasks, multidimensional projections are employed to allow visualization and interaction with the samples. In a parallel view, the same projection facilitates the visualization of previously generated clustering between instances.

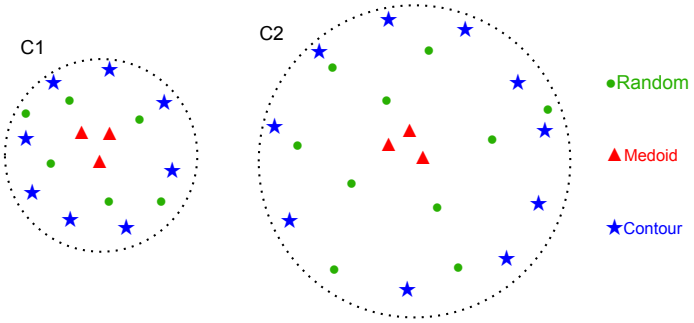


Figure 2. Types of sampling (random, medoid and contour). C1 (cluster 1) and C2 (cluster 2).

In order to assist interaction with projections, visualizations from data summarization should be presented on the top banner of the interface. In our implementation as a new proposal in visual AL, we deal with combined spectrograms over time. They are denominated Time-Line-Spectrogram (TLS)—see the top of Figure 3. The goal of the TLS visualization is to provide more visual information about samples while the user interacts with the projections, in the tasks of labeling by listening. TLS is a supportive visual representation that summarizes the spectrograms of each audio recording (instance) according to the time of recording in the landscape. Other types of data, such as documents, image collections, and videos, can also be summarized by tag clouds or representative pictures.

3.4. Learning-Prediction

This step aims to train a learning model from features and labels of samples, and this learning is used to predict the labels of the other instances to the data set. To accomplish this step, the user needs to perform the following tasks:

(i) *Learning*: Model training. In this case, the model to be used is Random Forest Classifier (RFC).

(ii) *Prediction*: After the learning, labels of the instances other than the samples are predicted. Then, by examining the results using the same visualizations, and the criteria of the application, the steps of the proposed method can be repeated starting from the Clustering step (Section 3.1).

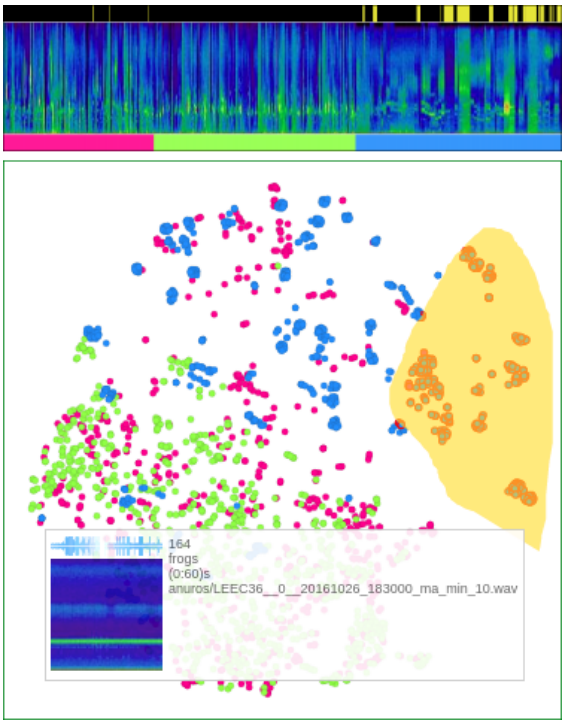


Figure 3. Example of visual iteration with Time-Line-Spectrogram (TLS): the points selected in the yellow color area in the projection are also visualized with yellow color in the spectrogram timeline (top of figure). In this view, projection multidimensional t-SNE is used for visualizing each data point.

3.5. Validation

To validate the proposed method, it is necessary to use a data set that has true labels, that is to say, that all instances of the audios had been previously labeled by expert users. Therefore, the validation in this stage of the method consists of comparing the real labels with predicted labels. In order to validate the prediction, we use the classification Accuracy (AC), which is defined by Equation (1):

$$AC = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

where TP are the true positives, TN are the true negatives, FP are the false positives, and FN are the false negatives.

3.6. Visualization

Visualization techniques—particularly multidimensional projections—are used in all stages of the proposed method (Clustering, Sampling, Annotation, and Learning), to support verification of results and user interaction where feasible.

By employing projections, users have the same mental model of all steps, and, by interpreting similarity between samples and their neighbors on the projections, the user is equipped with a powerful tool for browsing through data. This should increase the performance of expert users tasks.

4. Data Description and Case Study

To validate our proposed method, we used a data set provided in partnership with the Spatial Ecology and Conservation Lab (LEEC) of São Paulo State University (UNESP—Rio Claro). The soundscape recordings are part of Long-Term Ecological Research within the Ecological Corridor of Cantareira-Mantiqueira (LTER CCM or PELD CCM in Portuguese). The audio data were recorded within 22 landscapes distributed in the LTER CCM region, where the following types of environments were sampled: forest, swamps, and open area (mainly pasture). Originally, the region had been covered by forest, but, due to the

expansion of agriculture, pasture, and urbanization, the region shows varying forest cover from 16% to 85% in different portions of study areas. The recordings occurred between October 2016 and January 2017, and each landscape was surveyed during three consecutive months (30 days for forest, 30 days for swamps, and 30 days for open areas). Half of the recordings were collected in the morning (from sunrise to 8:30 a.m.) and half in the evening (6:30 p.m. to 10:00 p.m.). For the purpose of the current study, the raw data were re-sampled, in order to represent the soundscape heterogeneity of all the sampled environments. Therefore, the re-sampled data set has more than 40,000 sound files of one minute each. A total of 2277 sound files were labeled by experts. To assess our method, we chose to work with the data set containing those 2277 instances of audio. The sound files were labeled according to the most dominant sound in each minute, which were divided into three labels: 615 for frogs, 822 for birds, and 840 for the insects. Hereafter, we will refer to this data set as DS1. In the same region of LTER CCM, two other soundscape ecology studies were conducted: (1) [22], which aimed to assess how spatial scale (i.e., extents) influences acoustic indices responses and how these indices behave according to natural vegetation cover (%); and the authors in (2) [37] developed a method to identify the most discriminant features for categorizing sound events in soundscapes. This present study and its realization in the field of soundscape ecology represents an important tool for further ecological studies in this and other natural areas.

For data analysis, we follow the methodology described in article [37], which first suggests a preprocessing of the audios with a set of parameters to generate the Spectrum creation. *soundfile* (Available in: https://pypi.org/project/SoundFile, accessed on 22 June 2021) and *librosa* (Available in: https://librosa.github.io/librosa, accessed on 22 June 2021) in Python were used for this. The real part of the spectrum was used for obtaining spectrograms. The study also suggests that feature extraction from three different sources (descriptors based on acoustic indices, descriptors based on cepstral information, and descriptors based on the image of spectrogram) can be used together to the benefit of the analysis. For feature extraction, we employed *Essentia* (Available in: https://essentia.upf.edu, accessed on 22 June 2021), Python, Cython, and C.

In our experiments, for each audio minute, a total of 238 features were extracted. For the experiments, we set up four data sets (DS): (DS1) frogs, birds, and insects, with 2277 instances; (DS2) frogs and birds, with 1437 instances, (DS3) frogs and insects, with 1455 instances; and (DS4) birds and insects, with 1662 instances. Each data set has two feature settings: the first configuration has 102 features in total for each set (original features). For the second configuration, we select best features using the feature selection method based on important features known as Extra-Trees-Classifier (see [37]). After this step, we remained with 30 features for DS1, 30 for DS2, 46 for DS3, and 31 for DS4.

Data Availability and Bioethics

All the raw data used in this study are available on the following platform: https://github.com/LEEClab/soundscape_CCM1_exp01, accessed on 22 June 2021. The Spatial Ecology and Conservation lab of Biodiversity Department of UNESP is in charge of keeping this repository available and updated over time. By the nature of data (audio data only), the UNESP bioethical committee does not require any specific authorization, as no live animals were handled during the sampling period.

5. Results and Discussion

The predictions of data set labels were computed by performing all of the steps described in Section 3, which are: *Clustering*, *Sampling*, *Annotation*, *Learning-Prediction*, and *Visualization*. In this scenario, it is important to highlight that, for the purpose of our analysis, user interaction in the step *Annotation* is recreated or simulated using exceptionally the true labels only to assign the labels to the initial samples. This was so to guarantee the observation of learning effectiveness in the context of the experiments.

For the *Learning-Prediction* stage, the classifier used was *Random Forest* (RFC), for showing robust classification in soundscape data in comparison with other classifiers that had been tested such as: Support Vector Classifier (SVC), K-nearest Neighbor Classifier (KNNC), and X-Gradient Boosting classifier (XGBoost). Details of the experiments carried out to analyze the steps of the methodology are described in the next sections.

5.1. Clustering and Sampling Analysis

In order to define the best parameters for the method, the goal of this experiment was to evaluate the first steps of *Clustering* and *Sampling*. The accuracy of these steps contributed to increasing the accuracy in the prediction of labels. Therefore, an accurate prediction of the labels can, at the same time, mean a certain segregation of categories of events in data of soundscapes.

In this experiment, we evaluated the following parameters: the number of clusters (k), the total number of samples (p), the number of samples per cluster (pk), and the types of strategies for extracting samples (r, m, c, rm, rc, mc; and rmc—see Section 3.2). The setup of the experiment was defined as follows: (1) The four data sets DS1, DS2, DS3, and DS4 were used; (2) The number of clusters was analyzed in the range of $k = \{5, 6, ..., 49, 50\}$; (3) The number of samples per cluster was established considering the smallest possible number of instances per cluster, so the number of samples per cluster was fixed at$pk = \{5, 10, \sqrt{|.|}\}$, where $|.|$ represents the total number of instances per grouping and is usually used to define samples.

The visual active learning method was executed 7728 times for the four data sets, and eight sets of results were obtained in the form of tables; four for data sets with all 102 features, and four for data sets using the best features. The results for the DS1 data set are displayed in Table 1, where each row presents the seven types of samples (r, m ,c, rm, rc, mc, rmc) in the form of *heatmaps*. The maximum accuracy values are displayed in yellow, the minimum accuracy values are displayed in blue; and the intermediate accuracy values are displayed in a color gradient between yellow and blue.

Initially, the clusters were computed using the *K-Means* (KM) and *Hierarchical Agglomerative Clustering* (HAC) algorithms, but the best results were obtained with HAC for clusters larger than 20. As expected, we noticed that the larger the number of samples, the higher the accuracy. However, ideally, one wishes to use as few samples as possible to later predict most of the other instances. Therefore, the number of samples is limited by a threshold for the smallest possible number of samples. In this scenario, from the results, we can infer that the proposal to set the number of samples per clustering in $\sqrt{|.|}$ is the best option; thus, this parameter can be calculated automatically. Regarding the method for determining initial samples, the analysis focuses on the predominance of maximum accuracy values. After doing a visual analysis of the information in the table using *heatmaps*, the most suitable strategies for extracting samples in order (from best to worst) are: r, rc, rm, mc, rmc, m, and c.

In order to make a more clear and specific analysis, Figure 4 illustrates the comparison of results between the best features and all 102 features, specifically considering: the two best sample strategies (r and rc) and the worst sampling method (c) with 10 samples per cluster. In the results presented in Figure 4, we can observe the superiority of the samples r and rc for the data sets under analysis. Thus, through these experiments, the discriminatory capacity of the selected features was verified, achieving high accuracy in prediction of sound categories.

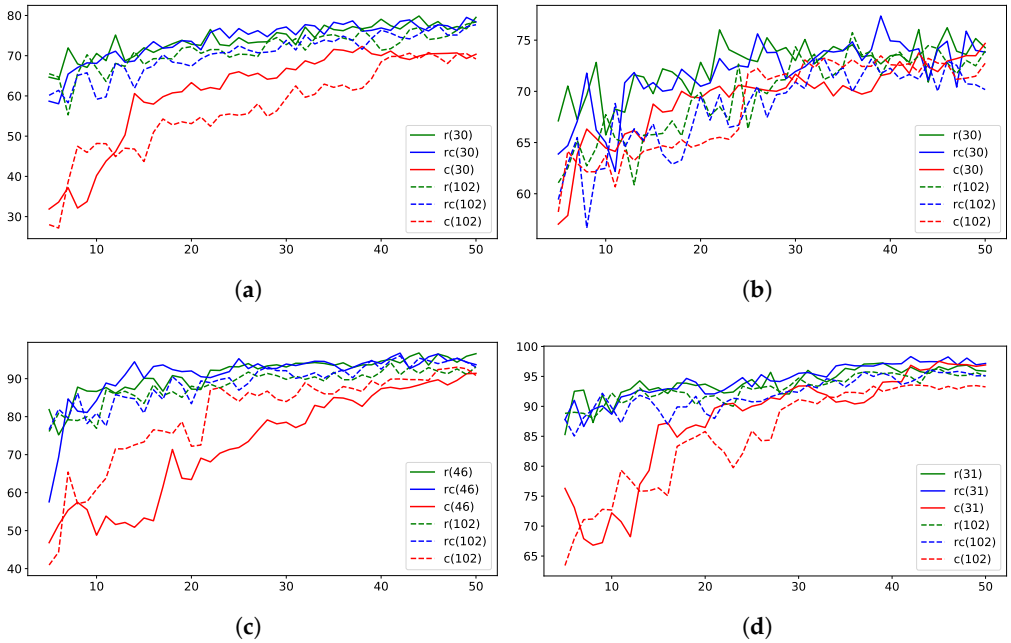
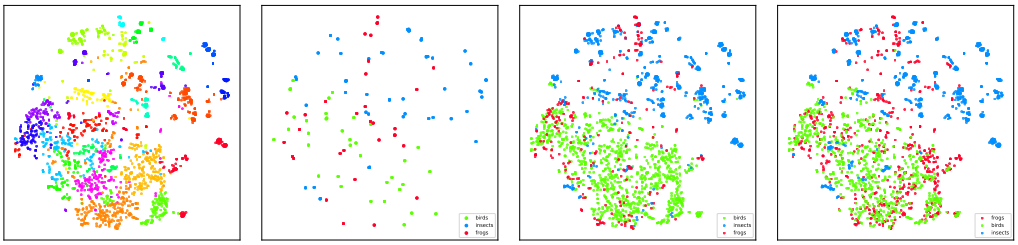


Figure 4. Comparison of accuracy results (in percentage 100%) according to the types of features and the sampling strategies: the graphs show results when the best features are used (continuous lines) and when all 102 features are used (dashed lines). This comparison is made considering the types of samples *r*, *rc*, and *c*, with 10 samples per cluster for data sets DS1 (**a**), DS2 (**b**), DS3 (**c**), and DS4 (**d**). In the graphs, the *x*-axis corresponds to the cluster values and the *y*-axis corresponds to the accuracy values.

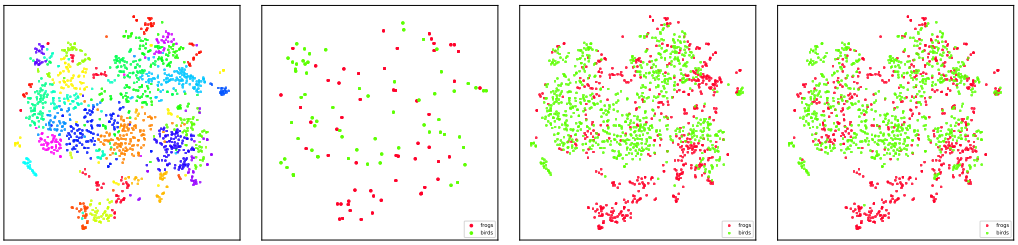
5.2. Visual Analysis via Projections

Some of the results presented in Table 1 can be visualized in Figure 5. A set of the resulting visualizations from the framework is presented: (1) $k = 24$ as the number of clusters; (2) $pk = 5$ as the number of samples per cluster; and (3) *rc* was selected as the method or strategies for extracting samples. The visualizations were generated using t-SNE projections. Figure 5 illustrates each step of the proposed method: Clustering, Sampling, Annotation, and Learning-Prediction. The ground truth is added for comparison.

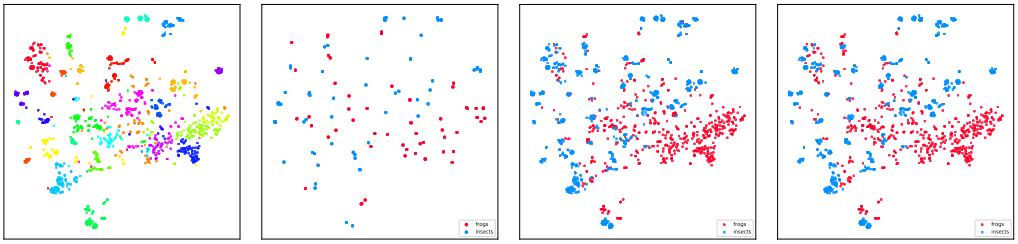
For each data set, the points of varying colors in Figure 5a,e,i,m, represent the 24 clusters generated in the Clustering stage. The colored dots with up to three colors of Figure 5b,f,j,n, represent the samples from the Sampling step, specifically the colors represent the user's labeling interaction in the step *Annotation*. The colored dots in Figure 5c,g,k,o represent the instances with labels that were determined by the prediction in the step *Learning-prediction*. It is important to report that—for the training of learning—the labels of the samples were considered, and, in the prediction of the rest of the instances, unlabeled of data sets were also used. Finally, the colors of the points in Figure 5d,h,i,p represent the actual true labels of the data set instances. Visually, we can observe the great similarity between the pairs of instances labeled in Figure 5c,d,g,h,k,l,o,p, which, in reality, translates as the visual degree of similarity of the labels in the prediction in relation to the true labels respectively for the four data sets. The accuracy achieved in the labeling task for each of the four data sets varied from high to pretty high: in DS1 equal to 72.6% (Figure 5c), DS2 equal to 72.36% (Figure 5g), DS3 equal to 91.09% (Figure 5k), and DS4 equal to 94.58% (Figure 5o). Based on these results—at least for our experiments—we can say that, in order to achieve high accuracy in automatic labeling, the experts mostly provide manual annotation for the following percentage of the data: 5.3% for DS1, 8.35% for DS2, 8.2% for DS3, and 7.2% for DS4.



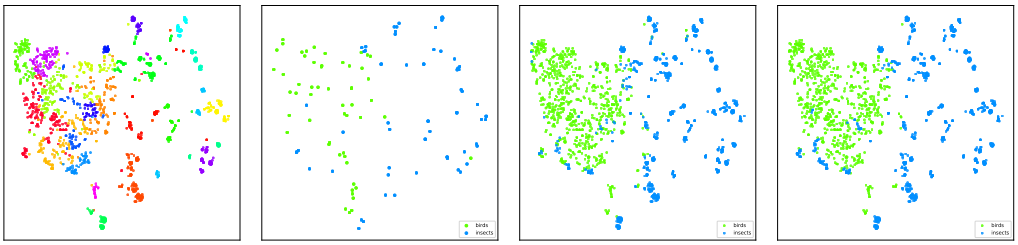
(**a**) $k = 24$ (**b**) $p = 120$ (5.27%) (**c**) prediction ($AC = 72.60$) (**d**) *ground truth*



(**e**) $k = 24$ (**f**) $p = 120$ (8.35%) (**g**) prediction ($AC = 72.36$) (**h**) *ground truth*



(**i**) $k = 24$ (**j**) $p = 120$ (8.25%) (**k**) prediction ($AC = 91.09$) (**l**) *ground truth*



(**m**) $k = 24$ (**n**) $p = 120$ (7.22%) (**o**) prediction ($AC = 94.58$) (**p**) *ground truth*

Figure 5. Sample visualizations of label predictions: the lines from top to bottom show the t-SNE projections of datasets DS1, DS2, DS3, and DS4 with their best features. Along the columns from left to right, colors represent each of the steps of the method: *Clustering*; *Sampling and Annotation*; and *Learning-Prediction*, as well as the *Ground-truth*.

6. Conclusions, Future Work, and Opportunities

In this study, we presented a method for support sample labeling by employing visual active learning as a label prediction strategy. A framework was tested in the context of a framework for labeling soundscape ecology data. Experiments evaluated each step of the process by the employment of a pre-labeled data set provided by our application partners. We have evaluated the number of clusters, the number of samples per cluster, and the sample strategy within each cluster. Results of the experiments were evaluated

according to classification accuracy, which determines the level of prediction of the labels. For clustering unlabeled data, Hierarchical Agglomerative Clustering (HAC) was adapted to the application since it allows for starting the process from user labeled samples, building groups incrementally. According to the results, the best sample strategies were *rc* and *r* because these strategies reached the most representative and informative samples from each cluster. The tables in the form of *heatmaps* with all cluster results identify that trend.

Therefore, we identify an effective parameter configuration for the method and for the current experiment. The best parameters were: (1) which clustering strategy to employ, (2) the number of clusters to generate, (3) the method of strategy to extract samples, (4) the number of samples per clustering, and (5) the technique of visualization where the user will interact in the labeling. Thus, the optimal configuration was: (1) Agglomerative Hierarchical Clustering (AHC) as a strategy to generate clusters; (2) clusters greater than 20; (3) samples of the method random or random combined with contours; and (4) number of samples $\sqrt{|.|}$, per cluster, where $|.|$ is the number of instances per cluster.

The main visual tool employed for user tracking of the process and interpreting the results of each step was multidimensional projections, which can reflect the clustering, sample selection, result of label prediction, and comparison with the ground truth.

With the results achieved in our study, we demonstrate that AL and multidimensional visualization can play an important role in achieving favorable performance with significantly less manual annotation effort. In addition, the accuracy of automatic labeling for our case study varied from high (72.36%) to very high (94.58%) for the four data sets, also reflecting a very good advance for the field of soundscape interpretation in the task of discriminating categories of sounds (in our case, groups of animals).

While our approach can be applied to any vector representation of the data set undergoing labeling, the best success in terms of accuracy is achieved when the data set is described by a set of features that has a good potential to discriminate target labels. Our experiments presented here show that aspect of the problem. It is our plan to tackle this problem in the next progression of our efforts in visual analytics for labeling.

Another contribution of our study is the framework itself (see Figure A1) that encapsulates all steps of our proposed method and its calculations. The framework is openly available. Although the framework is dedicated to soundscape data, it should be applicable to implementing the strategies for other data sets where a set of attributes can be extracted which consistently discriminates labels of interest.

In the particular case of soundscape ecology, recordings are being collected constantly, and the amount of data to be labeled increases exponentially in biodiversity monitoring worldwide [22]. Thus, developing solid and easy-to-use methods are of utmost importance for project managers that want to speedily extract information from their data. After data recording in the field, labeling is one of the most time-consuming tasks that precedes extracting knowledge. Therefore, contributions such as automatic labeling combined with a good visualization tool can be crucial for the success of conservation projects. The application of this approach was limited to experimenting with categorical sound events. While this could be construed as one of its limitations, the applicability of the results of the study can bring benefits to science such as understanding the behavior of certain environments, which can lead to the development of environmental monitoring strategies and policies of conservation.

In the next steps of the application studies, we will try to focus the framework in other categories, such as primates, bats, rain, dogs, human conversation, cars, airplane, guns, stream, wind, and background noises, and also verify applicability to another level of audio data resolution, such as identifying species.

More crucial to the development of visual analytics strategies in the future, it essential that the problem of labeling, which drives data analysis and understanding, machine learning, and crucial feature selection in many different applications, is dedicated more effort in order to find a balance and also to generalize the approach so that it can be applicable to a large variety of domains. Examples of areas where accelerating labeling can

have a real impact are document analysis, biological data interpretation, and image and video interpretation, to name a few.

Author Contributions: Conceptualization: L.H.H. and R.M.; all authors participated in regular meetings during the project and contributed to ideas to shape the concepts and complement their the execution. Implementation: L.H.H.; Contextualization and verification: L.H.H. and R.M.; Testing and Use Case Scenarios: All authors.; Writing: all authors.; Revision: M.C.R. and R.M. All authors have read and agreed to the published version of the manuscript.

Funding: L.H.H. was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001; MCR was partly funded by the São Paulo Research Foundation—FAPESP (processes 2013/50421-2; 2020/01779-5) and CNPq (processes 312045/2013-1; 312292/2016-3; 442147/2020-1); RM was partly funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) Grant No. 307411/2016-8.

Data Availability Statement: The data presented in this study are available on the following platform: https://github.com/LEEClab/soundscape_CCM1_exp01 (accessed on 28 June 2021).

Acknowledgments: The authors acknowledge the work of students in LEEC lab, in particular Lucas Gaspar, for the labeling of the data set.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Overview of the Visual Active Learning Framework for Soundscape Ecology

Our methodology was realized by a framework that we call SoundscapeX (code available at https://github.com/hhliz/vizactivelearning, accessed on 22 June 2021). The software implementation is based on a Client/Server architecture developed in Python that mainly employs the following libraries for the back-end: Tornado, Scikit-learn, Librosa, Pandas, and MongoDB. For visualizations, we employ the D3 Javascript library.

On the server, algorithms written in Python are used to perform clustering, sampling, learning, prediction, and projection tasks. On the client-side, algorithms written in D3/js are used to generate the layout of the views, mainly the projections, and, with these projections, the user interacts to perform the manual labeling. The application window is a view created with D3/JS in the web browser (client), but the visualization geometry is created by complex calculations from algorithms written in Python (server). Python and D3 communicate via JSON message passing.

Here, we provide an overview of the data exploration and labeling functionalities. Figure A1 presents an overview of SoundscapeX and the interface functions for exploring and labeling the data. The main interface components are shown in the regions labeled: (A) The process of labeling starts in the configuration panel, where the user issues a query of a data set to be used, the set of features to represent the data, type of the normalization, clustering technique to be used (k-means, HCA), number of clusters, and visualization (t-SNE or Uniform Manifold Approximation and Projection UMAP [45]). (B) At the top, the first seven buttons are used to interact and explore the data within the projection in region D. Then, the remaining buttons are a mini configuration panel to determine the type of sampling, number of samples, and launch of a small interface where the user can listen to the audio and label selected data samples. (C) The region presents the spectrogram of the whole set of audios in the form of a timeline; it is named the "Time-Line-Spectrogram" (TLS) and offers an overview of the audios under analysis. In addition, TLS is coordinated with the projections data allowing exploration. (D) In this region, the visualizations of the active learning process can be observed, by displaying the projection with colored clustering, the selection of samples to be labeled by the user, and finally the prediction of the labels. An example of the alternative visualizations in this region is seen in Figure 5. (E) The region presents the ground truth of the data when one is available, or the final labeling in normal circumstances. (F) Polar histogram of features. (G) Current set of features used in the data set; colors indicate the source (acoustic indices, spectrograms, or sound signal). (H) Legend of labels in the data set.

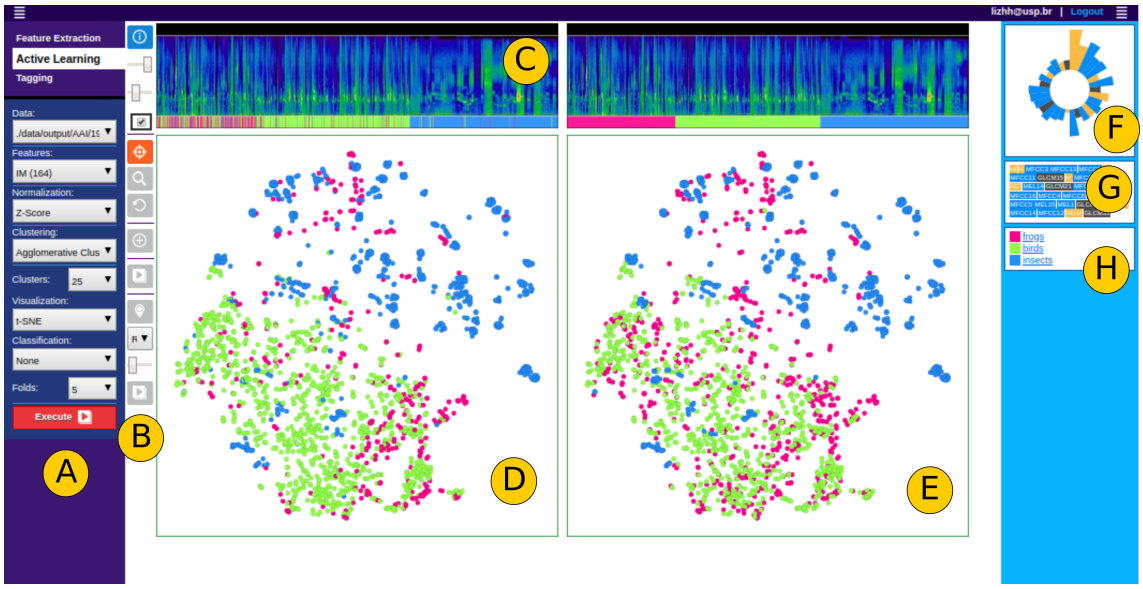


Figure A1. Screenshot of the main interface of the *Framework*. (A) Initially, a configuration parameter panel (B) Allows interact and explore the data within the projection in region D. (C) TLS is coordinated with the projection's data, allowing exploration interactive. (D) Multiple visualizations of the active learning process. (E) The ground truth. (F) Polar histogram of features. (G) Features used in the experiment. (H) Legend of labels in the data set.

References

1. Settles, B. Active Learning Literature Survey. In *Computer Sciences Technical Report 1648*; University of Wisconsin–Madison: Madison, WI, USA, 2009.
2. Piczak, K.J. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd ACM International Conference on Multimedia (MM '15)*; Association for Computing Machinery: New York, NY, USA, 2015; pp. 1015–1018. [CrossRef]
3. Tuia, D.; Volpi, M.; Copa, L.; Kanevski, M.; Munoz-Mari, J. A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification. *IEEE J. Sel. Top. Signal Process.* **2011**, *5*, 606–617. [CrossRef]
4. Pereira-Santos, D.; Prudêncio, R.B.C.; de Carvalho, A.C. Empirical investigation of active learning strategies. *Neurocomputing* **2019**, *326–327*, 15–27. [CrossRef]
5. Riccardi, G.; Hakkani-Tur, D. Active learning: Theory and applications to automatic speech recognition. *IEEE Trans. Speech Audio Process.* **2005**, *13*, 504–511. [CrossRef]
6. Kapoor, A.; Horvitz, E.; Basu, S. Selective Supervision: Guiding Supervised Learning with Decision-Theoretic Active Learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07)*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2007; pp. 877–882.
7. Moskovitch, R.; Nissim, N.; Stopel, D.; Feher, C.; Englert, R.; Elovici, Y. Improving the Detection of Unknown Computer Worms Activity Using Active Learning. In *KI 2007: Advances in Artificial Intelligence*; Hertzberg, J., Beetz, M., Englert, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 489–493.
8. Hu, R. Active Learning for Text Classification. Ph.D. Thesis, Technological University Dublin, Dublin, Ireland, 2011. [CrossRef]
9. Abdelwahab, M.; Busso, C. Active Learning for Speech Emotion Recognition Using Deep Neural Network. In Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), Cambridge, UK, 3–6 September 2019; pp. 1–7.
10. Mandel, M.I.; Poliner, G.E.; Ellis, D.P. Support Vector Machine Active Learning for Music Retrieval. *Multimed. Syst.* **2006**, *12*, 3–13. [CrossRef]
11. Xin, D.; Ma, L.; Liu, J.; Macke, S.; Song, S.; Parameswaran, A. Accelerating Human-in-the-loop Machine Learning: Challenges and opportunities. In *Conjunction with the 2018 ACM SIGMOD/PODS Conference (DEEM 2018), 15 June 2018, Proceedings of the 2nd Workshop on Data Management for End-To-End Machine Learning*; Association for Computing Machinery, Inc.: New York, NY, USA, 2018; pp. 1–4. [CrossRef]
12. Liao, H.; Chen, L.; Song, Y.; Ming, H. Visualization-Based Active Learning for Video Annotation. *IEEE Trans. Multimed.* **2016**, *18*, 2196–2205. [CrossRef]
13. Limberg, C.; Krieger, K.; Wersing, H.; Ritter, H. Active Learning for Image Recognition Using a Visualization-Based User Interface. In *Artificial Neural Networks and Machine Learning—ICANN 2019: Deep Learning*; Tetko, I.V., Kůrková, V., Karpov, P., Theis, F., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 495–506.

14. Maijala, P.; Shuyang, Z.; Heittola, T.; Virtanen, T. Environmental noise monitoring using source classification in sensors. *Appl. Acoust.* **2018**, *129*, 258–267. [CrossRef]
15. Servick, K. Eavesdropping on Ecosystems. *Science* **2014**, *343*, 834–837. [CrossRef] [PubMed]
16. Farina, A.; Pieretti, N.; Piccioli, L. The soundscape methodology for long-term bird monitoring: A Mediterranean Europe case-study. *Ecol. Inform.* **2011**, *6*, 354–363. [CrossRef]
17. Farina, A.; Pieretti, N. Sonic environment and vegetation structure: A methodological approach for a soundscape analysis of a Mediterranean maqui. *Ecol. Inform.* **2014**, *21*, 120–132. [CrossRef]
18. Putland, R.; Mensinger, A. Exploring the soundscape of small freshwater lakes. *Ecol. Inform.* **2020**, *55*, 101018. [CrossRef]
19. Kasten, E.P.; Gage, S.H.; Fox, J.; Joo, W. The remote environmental assessment laboratory's acoustic library: An archive for studying soundscape ecology. *Ecol. Inform.* **2012**, *12*, 50–67. [CrossRef]
20. LeBien, J.G.; Zhong, M.; Campos-Cerqueira, M.; Velev, J.; Dodhia, R.; Ferres, J.; Aide, T. A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. *Ecol. Inform.* **2020**, *59*, 101113. [CrossRef]
21. Pijanowski, B.C.; Farina, A.; Gage, S.H.; Dumyahn, S.L.; Krause, B.L. What is soundscape ecology? An introduction and overview of an emerging new science. *Landsc. Ecol.* **2011**, *26*, 1213–1232. [CrossRef]
22. Scarpelli, M.D.; Ribeiro, M.C.; Teixeira, F.Z.; Young, R.J.; Teixeira, C.P. Gaps in terrestrial soundscape research: It's time to focus on tropical wildlife. *Sci. Total. Environ.* **2020**, *707*, 135403. [CrossRef]
23. Hu, W.; Bulusu, N.; Chou, C.T.; Jha, S.; Taylor, A.; Tran, V.N. Design and Evaluation of a Hybrid Sensor Network for Cane Toad Monitoring. *ACM Trans. Sen. Netw.* **2009**, *5*, 4:1–4:28. [CrossRef]
24. Joo, W.; Gage, S.H.; Kasten, E.P. Analysis and interpretation of variability in soundscapes along an urban-rural gradient. *Landsc. Urban Plan.* **2011**, *103*, 259–276. [CrossRef]
25. Parks, S.E.; Miksis-Olds, J.L.; Denes, S.L. Assessing marine ecosystem acoustic diversity across ocean basins. *Ecol. Inform.* **2014**, *21*, 81–88. [CrossRef]
26. Sueur, J.; Farina, A. Ecoacoustics: the Ecological Investigation and Interpretation of Environmental Sound. *Biosemiotics* **2015**, *8*, 493–502. [CrossRef]
27. Bellisario, K.; Broadhead, T.; Savage, D.; Zhao, Z.; Omrani, H.; Zhang, S.; Springer, J.; Pijanowski, B. Contributions of MIR to soundscape ecology. Part 3: Tagging and classifying audio features using a multi-labeling k-nearest neighbor approach. *Ecol. Inform.* **2019**, *51*, 103–111. [CrossRef]
28. Abramson, Y.; Freund, Y. SEmi-automatic VIsuaL LEarning (SEVILLE): A tutorial on active learning for visual object recognition. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20–26 June 2005, San Diego, CA, USA; p. 11.
29. Turtinen, M.; Pietikänien, M. Labeling of Textured Data with Co-Training and Active Learning. Proc. Workshop on Texture Analysis and Synthesis, 2005; pp. 137–142. Available online: https://core.ac.uk/display/20963071 (accessed on 22 June 2021).
30. Lecerf, L.; Chidlovskii, B. Visalix: A web application for visual data analysis and clustering. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, France, 29 June–1 July 2009.
31. Namee, B.M.; Hu, R.; Delany, S.J. Inside the Selection Box: Visualising active learning selection strategies. nips2010, 2010; p. 15. Available online: https://cseweb.ucsd.edu/~lvdmaaten/workshops/nips2010/papers/namee.pdf (accessed on 22 June 2021).
32. Huang, L.; Matwin, S.; de Carvalho, E.J.; Minghim, R. Active Learning with Visualization for Text Data. In *Proceedings of the 2017 ACM Workshop on Exploratory Search and Interactive Data Analytics (ESIDA '17)*; Association for Computing Machinery: New York, NY, USA, 2017; pp. 69–74. [CrossRef]
33. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
34. Ristin, M.; Guillaumin, M.; Gall, J.; Van Gool, L. Incremental Learning of Random Forests for Large-Scale Image Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 490–503. [CrossRef] [PubMed]
35. Tasar, O.; Tarabalka, Y.; Alliez, P. Incremental Learning for Semantic Segmentation of Large-Scale Remote Sensing Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2019**, *12*, 3524–3537. [CrossRef]
36. Shin, G.; Yooun, H.; Shin, D.; Shin, D. Incremental Learning Method for Cyber Intelligence, Surveillance, and Reconnaissance in Closed Military Network Using Converged IT Techniques. *Soft Comput.* **2018**, *22*, 6835–6844. [CrossRef]
37. Hilasaca, L.M.H.; Gaspar, L.P.; Ribeiro, M.C.; Minghim, R. Visualization and categorization of ecological acoustic events based on discriminant features. *Ecol. Indic.* **2021**, 107316. [CrossRef]
38. Han, W.; Coutinho, E.; Ruan, H.; Li, H.; Schuller, B.; Yu, X.; Zhu, X. Semi-Supervised Active Learning for Sound Classification in Hybrid Learning Environments. *PLoS ONE* **2016**, *11*, 1–23. [CrossRef] [PubMed]
39. Shuyang, Z.; Heittola, T.; Virtanen, T. Active learning for sound event classification by clustering unlabeled data. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 751–755. [CrossRef]
40. Salamon, J.; Jacoby, C.; Bello, J.P. A Dataset and Taxonomy for Urban Sound Research. In *Proceedings of the 22nd ACM International Conference on Multimedia (MM '14)*; Association for Computing Machinery: New York, NY, USA, 2014; pp. 1041–1044. [CrossRef]
41. Shuyang, Z.; Heittola, T.; Virtanen, T. An Active Learning Method Using Clustering and Committee-Based Sample Selection for Sound Event Classification. In Proceedings of the 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo, Japan, 17–20 September 2018; pp. 116–120. [CrossRef]

42. Kholghi, M.; Phillips, Y.; Towsey, M.; Sitbon, L.; Roe, P. Active learning for classifying long-duration audio recordings of the environment. *Methods Ecol. Evol.* **2018**, *9*, 1948–1958. [CrossRef]
43. Shuyang, Z.; Heittola, T.; Virtanen, T. Active Learning for Sound Event Detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2895–2905. [CrossRef]
44. Wang, Y.; Mendez Mendez, A.E.; Cartwright, M.; Bello, J.P. Active Learning for Efficient Audio Annotation and Classification with a Large Amount of Unlabeled Data. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 880–884. [CrossRef]
45. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2020**, arXiv:1802.03426.