MLSP 2013 BIRD CLASSIFICATION CHALLENGE DOCUMENTATION FOR PARTICIPANTS

Forrest Briggs, Raviv Raich

Oregon State Unversity, Dept. of EECS
briggsf@onid.orst.edu, raich@eecs.oregonstate.edu

Yonghong Huang

Intel Labs
catherine.huang@intel.com

1. OVERVIEW

It is important to gain a better understanding of bird behavior and population trends. Birds respond quickly to environmental change, and may also tell us about other organisms (e.g., insects they feed on), while being easier to detect. Traditional methods for collecting data about birds involves costly human effort. A promising alternative is acoustic monitoring. There are many advantages to recording audio of birds compared to human surveys, including increased temporal and spatial resolution and extent, applicability in remote sites, reduced observer bias, and potentially lower cost. However, it is an open problem for signal processing and machine learning to reliably identify bird sounds in real-world audio data collected in an acoustic monitoring scenario. Some of the major challenges include multiple simultaneously vocalizing birds, other sources of non-bird sound (e.g., buzzing insects), and background noise like wind, rain, and motor vehicles.

2. BACKGROUND

The audio dataset for this challenge was collected in the H. J. Andrews (HJA) Long-Term Experimental Research Forest, in the Cascade mountain range of Oregon. Since 2009, members of the OSU Bioacoustics group have collected over 10TB of audio data in HJA using Songmeter audio recording devices. A Songmeter has two omnidirectional microphones, and records audio in WAV format to flash memory. A Songmeter can be left in the field for several weeks at a time before either its batteries run out, or its memory is full.

HJA has been the site of decades of experiments and data collection in ecology, geology and meteorology. This means, for example, that given an audio recording from a particular day and location in HJA, it is possible to look up the weather, vegetative composition, elevation, and much more. Such data enables unique discoveries through cross-examination, and long-term analysis.

Previous experiments on supervised classification using multi-instance and/or multi-label formulations have used audio data collected with song meters in HJA [4, 2, 3, 6, 7]. The dataset for this competition is similar to, but perhaps more difficult than that dataset used in these prior works; in earlier work care was taken to avoid recordings with rain and loud wind, or no birds at all, and all of the recordings came from a single day. In this competition, you will consider a new dataset which includes rain and wind, and represents a sample from two years of audio recording at 13 different locations.

3. OBJECTIVE

The goal in this challenge is to predict the set of bird species that are present given a ten-second audio clip. This is a multi-label supervised classification problem. The training data consists of audio recordings paired with the set of species that are present.

4. DATA SOURCES AND DESCRIPTION

All files for the competition are available to download on the Kaggle.com page for the contest.

The dataset for this challenge is a representative sample of HJA in 2009 and 2010 during summer, and presents a real-world acoustic monitoring scenario.

The full dataset consists of 645 ten-second audio recordings in uncompressed WAV format (16kHz sampling frequency, 16 bits per sample, mono). Participants may start from the original WAV audio files, or use pre-computed features that we provide.

There are 19 species of bird in the dataset (Table 1). Each ten-second audio recording is paired with a set of species that are present. These label sets were obtained by listening to and looking at spectrograms. Several experts inspected each recording, and each provided their own label set, along with estimates of their confidence. The final label set was formed by confidence-weighted majority voting.

These 645 ten-second recordings are split into train and test sets. You are given labels for the training set, and are asked to make predictions about the test set.

There is some relevant information in the WAV filenames about location and time. For example, one file is named `PC1_20090606_050012_0040.wav`.

978-1-4673-1026-0/12/$31.00 ©2012 IEEE

Table 1. The 19 bird species in the dataset.

Code	Name
BRCR	Brown Creeper
PAWR	Pacific Wren
PSFL	Pacific-slope Flycatcher
RBNU	Red-breasted Nuthatch
DEJU	Dark-eyed Junco
OSFL	Olive-sided Flycatcher
HETH	Hermit Thrush
CBCH	Chestnut-backed Chickadee
VATH	Varied Thrush
HEWA	Hermit Warbler
SWTH	Swainson's Thrush
HAFL	Hammond's Flycatcher
WETA	Western Tanager
BHGB	Black-headed Grosbeak
GCKI	Golden Crowned Kinglet
WAVI	Warbling Vireo
MGWA	MacGillivray's Warbler
STJA	Stellar's Jay
CONI	Common Nighthawk

- The first part, `PC1` indicates the location ("penology control 1"). There are 13 distinct location codes in the form `PC#`. Figure 1 shows a map of these locations.

- The second part, `20090606` is the date in `YYYYMMDD` format.

- The third part, `050012`, is the time of day in `HHMMSS` format (24hr, pacific standard).

- The four part, `0040`, is an offset in seconds (this 10-second clip of a recording recording starts at the time indicated by the 3rd part + the fourth part offset in seconds).

4.1. Essential Files

We provide some files which are essential for participants, and some which are extra which may or may not be used. All of these files are described in greater detail in the README file, so we describe them only briefly here.

- `src_wavs` – a folder of WAV audio files. This is the source data for the challenge (both training and test sets)

- `CVfolds_2.txt` – specifies which fold (train/test) each recording is in. 0 = train, 1 = test.

- `rec_id2filename.txt` – Gives a unique `rec_id` ("recording id") to each audio file, in the range 0 – 644.

- `species_list.txt` – Gives a unique number to each species name, in the range 0 – 18.

- `rec_labels_test_hidden.txt` – These are the species label sets for the training data. A `?` is used for recordings in the test data. Your task is to make predictions for the recordings marked with a `?`.

- `sample_submission.csv` – You are required to submit your predictions as a text file. This is an example of the format that the predictions must be in.

5. SUBMISSION ON TEST DATA

Your task is to make predictions about the recordings in the test set. Specifically, for each recording in the test set, you are to give a probability that each species is present. These probabilities will be evaluated by AUC. Formally, let the set of classes/species be $\mathcal{Y} = \{1, \ldots, c\}$, where c is the number of classes. For each recording in the test set R_i, let its unknown species/label set be Y_i. Your task is to predict $P(j \in Y_i | R_i)$ for $j = 1, \ldots, c$.

Your submission must exactly match the format in `sample_submission.csv`. The first line of the submission should be the header:

```
rec_id,species,probability
```

Each of the following lines should list one estimate of $P(j \in Y_i | R_i)$ for a particular recording i in the test set. There should be 3 numbers on each line, separated by 2 commas –

1. The `rec_id` of a recording in the test set.

2. The class index j, for $j = 0, \ldots, 18$.

3. Your classifier's prediction of $P(j \in Y_i | R_i)$.

Make sure your predictions meet these requirements:

- Do not make predictions for recordings in the training set. Only include predictions for recordings in the test set. For example, in `sample_submission.csv`, the sequence of `rec_id`'s listed is 0,1,4,6,7,10,. . ., because these are the first recordings in the test set.

- There is no blank line at the end of your predictions.

- There should be exactly 6138 lines in your prediction file.

- Predictions are in the range [0,1].

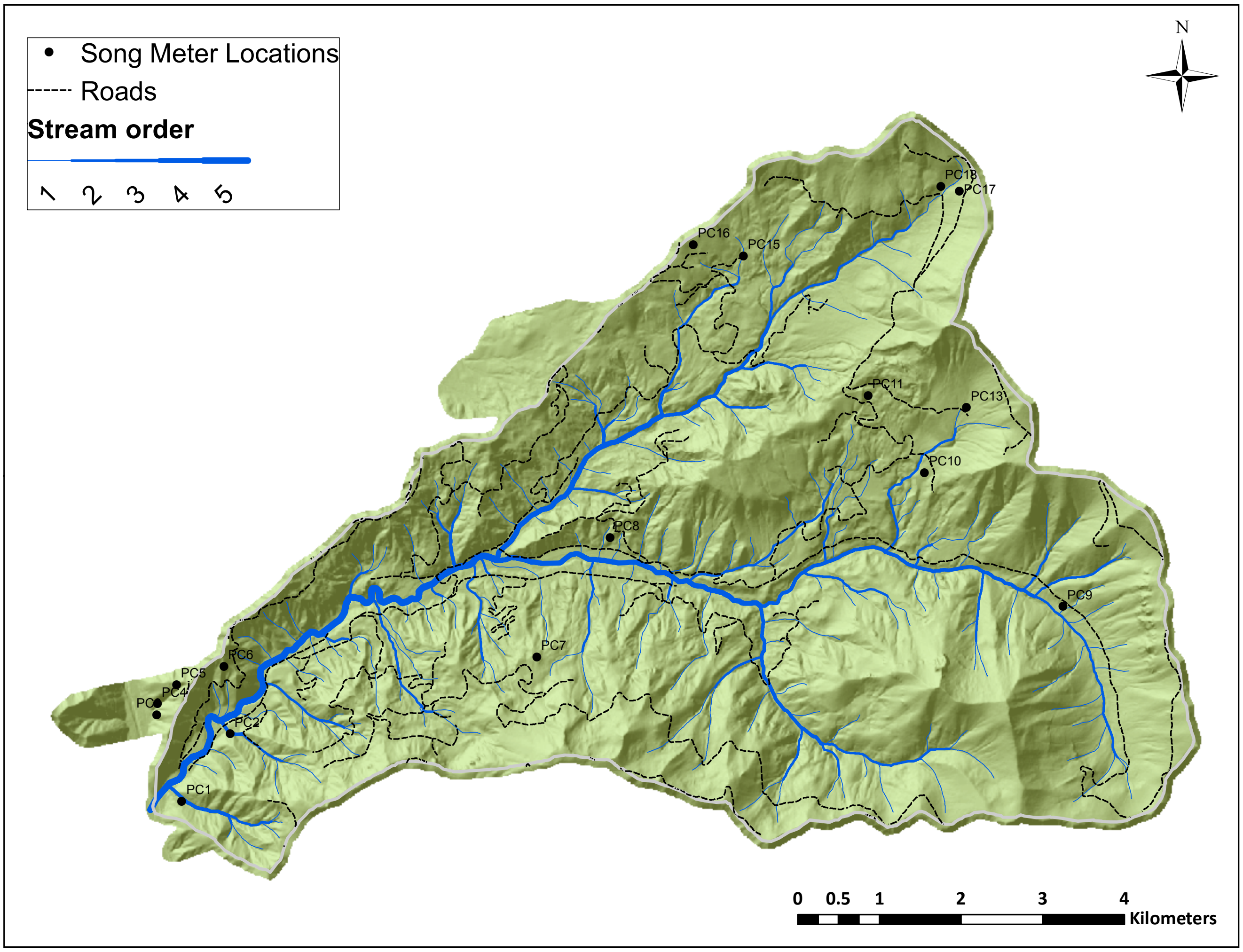

Fig. 1. Song meter data collection locations in the H. J. Andrews Experimental Forest.

6. BASELINE METHOD & SUPPLEMENTARY DATA

There several steps to go from a WAV audio file to a predicted probability. Participants may prefer to use some precomputed data that we provide, or they may instead prefer to do it all from scratch (or some combination). Many bird species classifiers involve some of these steps:

- Spectrogram – the raw audio signal is converted into a spectrogram (an image representing the sound), by dividing it into frames, and applying the FFT to each frame.
- Enhancement/noise reduction – filters are applied to the signal before or after creating the spectrogram to reduce noise and boost the sounds of interest.
- Segmentation – an algorithm identifies intervals of time or regions / rectangles in the spectrogram which correspond to distinct utterances or segments ("syllables") of bird sound.
- Features – At some level of structure, a part or the whole audio recording is summarized by feature vectors. Some features such as MFCCs describe individual frames of audio. Others describe syllables or segments. Still others describe a clip of audio as a hole. Some features are derived from the spectrogram, and others are not (e.g., they might be derived directly from the signal).
- Classifier - a supervised classifier is trained using the features and some labeled examples.

We provide one implementation of the some of these components, based on prior work [4]. Specifically, we provide the following data (described in more detail in the README), which may optionally be used:

- Spectrograms computed from the source WAV files, and stored as BMP images.
- Filtered spectrograms, with some wind/stream noise suppressed.
- A subset of 20 spectrograms in the training set have been annotated at the pixel level with course examples of correct segmentation. In this annotation, red denotes bird sound, and blue denotes rain or loud wind.
- Using the 20 spectrograms annotated with correct segmentation, we train a Random Forest on to classify each pixel of a spectrogram as bird sound or not. This classifier is applied to obtain an automatic segmentation of each spectrogram into distinct utterances / segments. The algorithm is similar to [4], with minor differences (see Appendix 1).
- We provide a bounding rectangle for each of the segments obtained by this supervised segmentation process. This may be useful for participants who want to compute their own features, but not run segmentation from scratch.
- For each segment, we provide a 38-dimensional feature vector, as in [4]. These segment features constitute a "multi-instance" representation of the dataset. In a multi-instance representation, the dataset is a collection of "bags-of-instances." In this formulation, a recording is a bag, and the segment/38-d features are instances. This kind of representation was previously used in [4, 2, 3, 6].
- The segments are clustered using k-means++ to form a codebook, then each recording is represented using a histogram-of-segments feature, as in [5]. This means each recording is represented by a fixed length-feature vector. This representation of the data may be useful for participants who want to focus on a standard "multi-label classification" formulation. This is the most processed version of the data we are providing; participants need to do the least to go from here to the predictions, but it may be difficult to win just starting from here.
- We provide some visualizations of the dataset to illustrate the diversity of sounds that appear.

7. OTHER RULES

With one exception, you are not allowed to use any other data outside of what is provided. If there is any doubt, don't use it.

- You may use the extra files in `/segmentation_examples` that we provide for training in additional to the species label sets.
- You may use additional data that you provide to annotate the training set with examples of correct segmentation into syllables in any form appropriate (e.g., bounding boxes). We are allowing this one kind of extra data because many systems require some kind of examples or templates for segmentation. The spirit of the rule here is that we want methods which require only a little bit of low-expertise human effort in the training set, and no human effort in the test set.
- You may not annotate data in the training set in any way that provides additional species labels. This means you are not allowed to label specific syllables or regions of a spectrogram with a single species label, or manually extract templates of syllables of certain species (your extra annotation is limited to bird/non-bird).

- You may not use any data outside of what is provided, with the exception of your own additional annotations in the training set. This includes, but is not limited to, other training examples from the dataset in [4], any data available about other experiments in HJA which have taken place at the same sites, and weather data.

- You may not do anything that requires manual/human effort in the test set.

- You may use the location code (`PC#`), or the time of day encoded in the filenames.

8. DEADLINE

Refer to the Kaggle.com competition page for announcements about dates.

9. PUBLICATION

The MLSP 2013 proceedings will include a publication, written by the competition chairs, which describes the competition and shows the comparison of the submitted methods. The committee chairs will invite up to three teams to submit a two-page summary that discusses the method they used in the competition (the mathematical notation must be consistent with the paper authored by the competition chairs). Pending approval, the invited summaries will be published in the conference proceedings.

10. AWARDS

The 2013 MLSP Organizing Committee is providing the prizes. The 2013 MLSP Competition Committee will distribute one award to each team or teams that they select, where the selection is based on: (1) the performance of the submitted methods and (2) the requirement that at least one member of each selected team attend the 2013 MLSP Conference. Members of the 2013 MLSP Competition Committee (and everyone belonging to any of the labs of the 2013 MLSP Competition Committee) are not eligible for this award.

11. APPENDIX 1: BASELINE METHOD SEGMENTATION

This appendix gives more information about the baseline segmentation method in the pre-processed data we provide. The approach is very similar to what was used in [4] and [5], with some differences in parameters and features. The main idea is classify each pixel in the spectrogram as bird sound or not using a Random Forest [1]. Each pixel in each spectrogram is described by a feature vector. For a collection of training images, each pixel is given a positive or negative label, as specified by manual annotation of the pixels (for example, Fig. 2).

Each pixel of a spectrogram is described by a feature vector with the following elements:

- The raw pixel intensity of all pixels in a 17x17 box around the pixel (this gives a 17^2 = 289-d feature).

- The average intensity of all pixels in that box (1-d)

- The y-coordinate of the pixel, which corresponds to frequency (1-d)

- The raw pixel intensity of all pixels in the same column as the pixel (256-d) (this feature is intended to help with segmentation in rain).

Based on the collection of manually annotated spectrograms, we form a training dataset pairing pixel features with the labels 0 = background noise, and 1 = bird sound. There are 318976 pixels in each spectrogram, so it is somewhat impractical to use all of them. Instead, we randomly sample 30% of red pixels as label 1, 30% of blue pixels as label 0, and 4% of uncolored pixels as label 0. From the 20 annotated spectrograms, this sampling process yields 467958 pixels which are used to train a Random Forest with 100 trees. The trees are grown to a maximum depth of 10, and histograms are stored in the leaves.

The trained RF classifier is applied to each pixel in every spectrogram, which gives a probability for the pixel to be bird sound. The probabilities may be noisy when viewing individual pixels in isolation, so they are averaged over a neighborhood by applying a Gaussian blur to an image of the probabilities, with a kernel parameter $\sigma = 3$. The blurred probabilities are then compared to a threshold of 0.4 (blurring reduces energy so a threshold of 0.5 is slightly too high). Pixels with probabilities above the threshold are considered to be bird sound and pixels with probabilities below the threshold are considered background. The files in `/supervised_segmentation` are a visualization of the results of this segmentation process (see Fig. 3 for an example).

These segments are then used to compute the same 38-d feature vector described in [4].

12. ACKNOWLEDGEMENT

Collection and preparation of this dataset was partially funded by NSF grant DGE 0333257, NSF-CDI grant 0941748, NSF grant 1055113, and the College of Engineering, Oregon State University. We would also like to thank Sarah Hadley for her role in data collection.

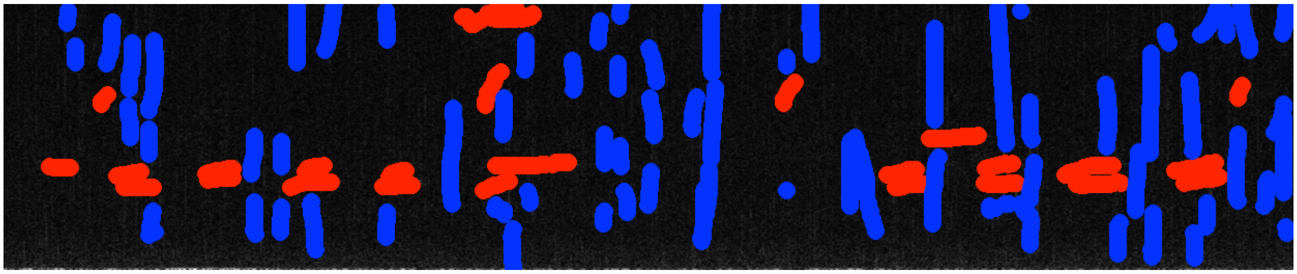

Fig. 2. Manually labeled spectrogram with example of correct segmentation. Red = bird, blue = rain. This spectrogram corresponds to 10 seconds of audio.

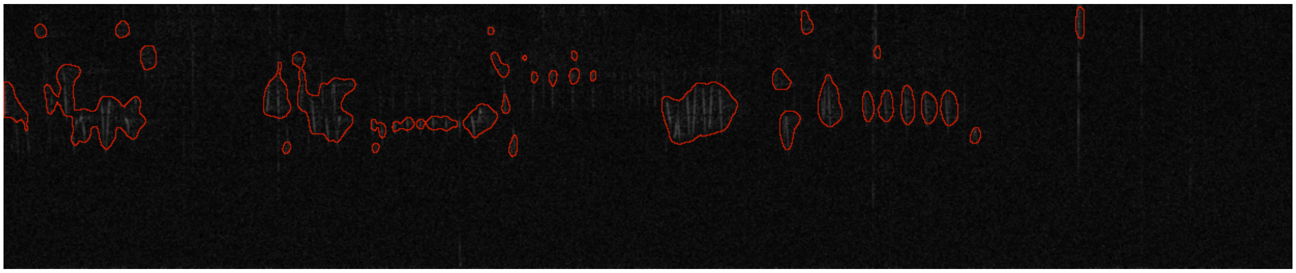

Fig. 3. Automatic segmentation of a spectrogram corresponding to 10-seconds of audio. This recording contains both rain and bird sound. In some cases the segmentation successfully ignored the rain and isolated bird sound, but in others in mistakenly labeled some rain as bird sound.

References

[1] L. Breiman. Random forests. *Machine learning*, 45(1): 5–32, 2001.

[2] F. Briggs, X.Z. Fern, and R. Raich. Rank-loss support instance machines for miml instance annotation. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 534–542. ACM, 2012.

[3] F. Briggs, X.Z. Fern, R. Raich, and Q. Lou. Instance annotation for multi-instance multi-label learning. *Transactions on Knowledge Discovery from Data (TKDD), 2012*, 2012.

[4] F. Briggs, B. Lakshminarayanan, L. Neal, X.Z. Fern, R. Raich, S.J.K. Hadley, A.S. Hadley, and M.G. Betts. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *The Journal of the Acoustical Society of America*, 131:4640, 2012.

[5] F. Briggs, X. Z. Fern, and J. Irvine. Multi-Label Classifier Chains for Bird Sound. *ArXiv e-prints*, April 2013.

[6] Liping Liu and Thomas Dietterich. A conditional multinomial mixture model for superset label learning. In *Advances in Neural Information Processing Systems 25*, pages 557–565, 2012.

[7] L. Neal, F. Briggs, R. Raich, and X. Fern. Time-frequency segmentation of bird song in noisy acoustic environments. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2011.