Identification of Songbird Species in Field Recordings

Hsiao-Yu Tung
htung@andrew.cmu.edu
htung

De-An Huang
deanh@andrew.cmu.edu
deanh

Xiao-Feng Xie
xfxie@cs.cmu.edu
xfxie

Yurui Zhou
yuruiz@andrew.cmu.edu
yuruiz

Joseph Russino
jrussino@rec.ri.cmu.edu
jrussino

1 Introduction

It is important to gain a better understanding about the climate and ecological changes in the world. One way to address this is to study seasonal migration patterns in songbird populations, since birds respond quickly to environmental changes [26]. During migratory periods, many species of songbirds use flight calls, which are species-specific and are distinct from other vocalizations. Therefore, flight calls information can be used to determine the relative abundance of species and is important to understand long-term population trends. Due to costly human effort to collect data about birds in traditional methods, using machine learing (ML) methods to identify bird species from continuous audio recordings has been a hot topic in in recent conference competitions.[1]. Although there are some recent advances [4, 19, 20, 24], it is still an open ML problem to reliably identify bird sounds in field recordings data due to simultaneously vocalizing birds and various background noise [6].

In this project, we will focus on critical aspects of this problem. We start from some state-of-the-art algorithms, e.g., [3, 4, 7, 19, 20, 24], adapt components to the data we have, and finally develop a scalable, integrated software tool. The total process is divided into four steps. First, Audio data are first preprocessed into spectrograms. The spectrograms are further cleaned by applying background noise reduction and image processing techniques, and connected pixels (acoustic patterns) in the spectrograms are labeled into rectangle segments. Second, features are then extracted and selected from different sources, e.g., file statistics, segment statistics and probabilistics, and mel-frequency cepstral coefficients (MFCC). Third, the classification is then done by using multiple algorithms, e.g., naive Bayes, decision trees, k-nearest neighbors (k-NN) [9], support vector machines (SVM) [8], etc. Finally, we will explore some ensemble methods [1, 2, 14, 17, 23] for furthering improve the overall performance by combining the predictions of models, as well as facilitating scalability in real-world usages. The software developed for this project will be used by the Carnegie Museum of Natural History, and possibly shared with other land managers, researchers, and educators to enhance the use of flight calls as a method to study the populations of migratory songbirds.

2 Related work

As mentioned in the Introduction, the toal process is divided into standard steps of ML components.

The first part is about *preprocessing and segmentating* audio data of songbirds. Normally audio files are first processed into grayscale image by applying the Fourier transform using a Hanning or Hamming window of samples with some overlap [5, 19]. Only relevent frequency range of the scope of domain interests are kept. The narrowed spectrograms are then treated as graysale images. To reduce the background noise, a *median clipping* method can be applied on each frequency band and time

[1] ICML 2013: The Bird Challenge; NIPS 2013: Multi-label Bird Species Classification; MLSP 2013: Bird Classification Challenge

frame to not only remove most background noise, but also capture the sound feature clearly and precisely [19]. An alternative is to apply iterations of a whitening filter [5]. A spectrogram can also be processed using a series of subprocesses including Gaussing filtering, local gradient, thresholding, and morphological nosing removal [13]. The resulting images can be further handled using standard image processing techniques such as diluation and median filter (e.g. using scikit-image). In [19], neighboring pixels exceeding certain spatial threshold (acoustic patterns) in the spectrograms are labeled into rectangle segments. In [13], small segements are discarded and remaing holes are filled. In [5], a supervised single-instance single-label classifier is used to label the probability of each pixel as bird sound or noise, and then obtain predicted segementations by applying a thershold.

The second part is about *feature extraction and selection*, which can have a large impact on later classification results. In [5], segement features are divided into two different categories - "mask descriptors", which describe the general shape of a segment, and "profile statistics", which describe the frequency and time profiles within segments. Histogram of gradients (HOG) have also been used [5]. In [13], features are obtained by applying the template matching function of scikit-image to compute the similarity at the maximum value of the normalized cross-correlation map with templates. In [19], features come from three different sources, i.e., file statistics, segement statistics, and segment probabilities. The template matching in OpenCV library is applied only on absolute-intensive spectrograms. Many existing work [7, 10, 20, 25] considers mel-frequency cepstral Coefficients (MFCC), which have been proved useful for speech recognition, as features. Features can also be extracted using unsupervised deep learning [21]. Addtional methods, e.g., rescaling [5], concatenation [10], bag-of-words (BoW) model [12], and principal component analysis (PCA) [18], can also be used for feature engineering.

The third part is about *classification*. Typical methods include naive Bayes, neural networks, logistic regression, Gaussian mixture model, radial basis function (RBF), decision trees, k-NN [9], SVM [8], etc. Binary classifers can be turned into multi-class ones by using ome general strategies, e.g., one-versus-all and pairwise decomposition, to classify instances into multiple classes. Some existing methods have been used for songbird identification. In [10], a LibSVM is used in a one-versus-all fashion, and best scores have been obtained with C-SVC SVM type and linear kernel function. In [21], pairwise SVM, LibSVM, decision trees, and neural networks are used, and the merged SVM and RDT often leads to better results. Mutiple-instance multi-label (MIML) classsifers, e.g., MIML-SVM, MIML-RBF, MIML-kNN are considered in [5].

Finally, different *ensemble learing* methods have also been used for combining the predictions of several models. In theory, ensembles can be more flexibility in the functions they can represent. Typical methods including gradient boosting (GB) [15], random forest (RF) [2], extremely randomized trees (ERT) [16], bootstrap aggregating (or bagging) [1], bayesian model averaging (BMA) [17], and ensemble of classifier chains (ECC) [23]. Some of them, e.g., GB [7], RF [7, 13, 25], ERT [19], have been directly used for songbird identification. There are also some hybrid methods. In [7], a simple linear blending is ued to combine GB, RF, and a Lasso and elastic-net regularizion of generalized linear model. In [20], an ensemble of logistic regression and GB classfiers are considered. In [3], an ensemble of classifier Chains with Random Forest is applied.

Many classification and ensemble methods can be obtained in scikit-learn library.

3 Methods

3.1 Preprocessing

[Yurui] In preprocessing we first re-sampled the audio files to 22050 Hz guarantee a uniform audio sample format for all training and test data. For each audio file we split it into 256 segments. Then the Power Spectral Density of each segments can be calculated. For each segments we use a Hanning window with %75 overlap. Finally, the spectrogram of each audio file is treated as gray-scale image for further noise reduction and segmentation.

Notice the case that in a processed grayscale image most area was occupied by the random noise. What we want is to get rid of the background noise completely and increase the contrast between real signal and the background. Given the several different algorithm tested, the best one is the so called Median clipping algorithm. The idea is simple, for each pixel we compute the median and variance

of its corresponding row and column. If the pixel is above the median plus three times variance we set the pixel value to 1, otherwise the value would be set to 0. The median clipping algorithm works best because it not only removes most background noise, but also capture the sound feature clearly and precisely.

Such a algorithm, though perform well in noise removing and feature capturing, requires large amount of computation because the average and variance of each column of the row need to be calculated. Given the huge size of the processed grayscale image, we need a more efficient algorithm to segment the image. Thus further experiments is needed to explore an balance point between noise reduction effectivity and efficiency.

Similar to the methods used in [19], we apply standard image processing techniques to further reduce residual noise dots. Finally we would use find connected pixels from the image and label the segments.

3.2 Features

3.2.1 Mel-Frequency Cepstral Coefficients

While flight calls recognition is a new application, audio recognition is a well developed area in signal processing. We will build features based on mel-frequency cepstral coefficients (MFCCs), which are commonly used as features in speech recognition systems. A temporal signal is first transformed into a series of frames where each frame consists of 13 MFCCs. Each frame represents a duration of 12ms. The step size is 4ms. We also include the fist and second derivatives of MFCCs, which results in a 39 (13×3) dimensional vector for each frame.

3.2.2 Bag-of-Words Model over MFCCs

In contrast to [11], which assumes that the number of frames is fixed for all audio segments to classify, we make no assumption on the length of the audio segment, since the length of flight calls might variate. In this case, we want a feature that is temporally scale invariant. Intuitively, even when a flight call is temporally scaled (extended or shorten), it should still be classified as the flight call of the same species. We leverage the progress in image classification and applied the bag-of-words (BoW) model [12] over our MFCCs. By treating audio features (MFCCS in our case) as 'words', each audio segment is represented by a sparse vector of occurrence counts of words in BoW model; that is, a sparse histogram over the vocabulary.

We learn the vocabulary, also called the codebook, by performing k-means clustering over sampled training MFCCs. Codewords are then defined as the centers of the learned clusters. In test time, we extract MFCCs from the test audio segment and quantize the 39 dimensional MFCCs to a learned codeword. The audio segment is represented by a K dimensional histogram over the codeword, where K is the size of the codebook.

One limitation of BoW model on audio segment is that the temporal information are lost in the model. However, as shown in the experiments, BoW is able to achieve satisfactory classification result on single-label single-instance case. We are considering to build higher order N-gram model on the learned vocabulary to capture the temporal information.

3.3 Classifiers

We consider several classifiers/regressors.

Nearest Neighbor (NN). For each testing data, the NN classifier finds the nearest training data point and transfer the corresponding label. The NN classifier directly reflect the effect of our features and is used as baseline. Depends on the features, different distance metric should be used. We will compare the performance of euclidean and χ^2 distance on the BoW feature in the experiments.

Support Vector Machine (SVM). The most common approach for multi-label classification is to use an ensemble of binary classifiers, where each classifier predicts if an instance belongs to one specific class or not. SVMs trained in a pairwise fashion has obtained state-of-the-art results on standard multi-label benchmark datasets [22]. Furthermore, we will be able to integrate different kernels to improve the performance using the SVM. For example, the χ^2 kernel will have a better

Table 1: Accuracy of different classifier

classifier	setting	accuracy	classifier	setting	accuracy
KNN (distance)	chi-square	62.27	SVM (kernel types)	chi-square kernel	72.73
	euclidean	54.09		RBF	70
				linear	67.7273
Naive Bayes	kernel density	15.91		polynomial	69.0909
RandomForest		7.27		Sigmoid	70.9091

performance for histogram than the linear kernel. We also consider support vector regression to produce soft output for ensemble learning.

Random Forest. Random forests has been widely used for multi-label classification [7, 19, 25, 27]. Random Forest is operated by constructing decision tree structure by the training examples. One of the popular algorithm is tree bagging, in which the training process includes repeatedly selecting a bootstrap sample of the training set and fitting the trees to them. After the training process, the label decision is made either on the majority of the votes or a weighted combination from all individual tree.

Naive Bayes.

Logistic Regression. Ensemble of logistic regression has been used for multi-label bird song classification [20] because of its simplicity. We also consider logistic regression as the baseline for soft output.

3.4 Ensemble Learning

4 Experiments

We first evaluate our system on single-label data to see if the flight calls are separable using our method.

Data Set Large amounts of audio data (about 20 terabytes) for bird calls have been collected. The data contain flight calls from approximately 20-30 species of songbirds, examples of other background noises, and long audio files that contain flight calls (some of them might be too faint to identify them to species) and other background sounds. The software developed can be tested using long audio files where all the flight calls of songbirds have been detected, and with manual identification by Amy Tegeler, an Avian Ecologist from the Carnegie Museum of Natural History.

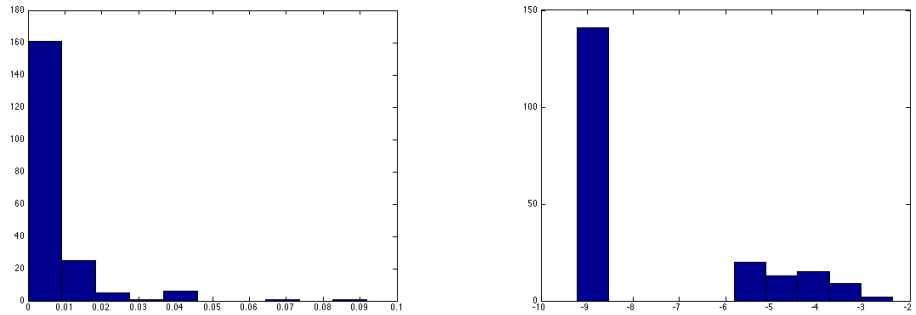
Data. We use the flight calls of songbirds manually segmented and labeled by Amy Tegeler, an Avian Ecologist from the Carnegie Museum of Natural History, to evaluate the performance of our system on indoor flight calls classification. We used the data from year 2008 to 2013. The total number of species is 32. But only 11 species has more than 100 data. Therefore, the experiment is performed on those 11 species, and 100 data points are randomly sampled from each species. The 100 data points are randomly partitioned into 20 testing data and 80 training data for each species.

Segmentation. Since the flight call segments are already manually identified, we do not perform segmentation in this experiment.

Features. We use the BoW over MFCCs in Section 3.2.2 as the features for this experiment. The size of the code book is 200, and the MFCCs consider the frequencies from 1500 Hz to 22000Hz.

Classifiers. Before beginning with the training process, we first examine the property of the data. We can see that most of the value in the original data is zero. After taking the log value of the data, we can find that a small group of the data is separated from the others. In some of the dimension, this small contains only a number of specific species of birds, which might be helpful in the classification.

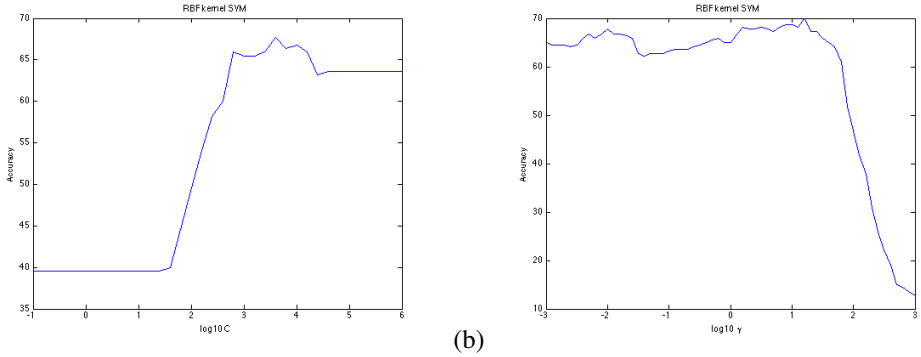
We compare the results using nearest neighbor, support vector machine, random forest and naive bayes classifier.



(a) (b)

Figure 1: (a) is the histogram of second feature. Most of the values are close to zero. (b) is the histogram of the log value of the original data.

We can see from our results that kernel SVM has better performance than the other method. First we show the parameter selection in using RBF kernel. Figure 2 shows the accuracy to the penalty term C and the parameter γ. The parameter we selected are $\gamma = 15$ and $C = 4000$.



(a) (b)

Figure 2: (a) Accuracy to penalty term c. (b) Accuracy to parameter γ.

In considering that original feature distribution is not a Gaussian distribution, for x_i, j representing the j^{th} feature in the object i, we first use a log transformation such that $x_{i,j} = log(x_{i,j} + 0.0001)$ to make the feature distribution more like a combination of Gaussian distribution (see 1(b)). The operation leads to a 1.27 improvement in the accuracy after parameter selection process.

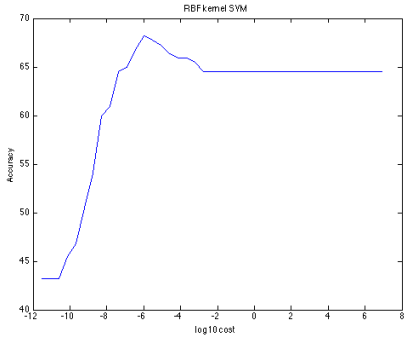
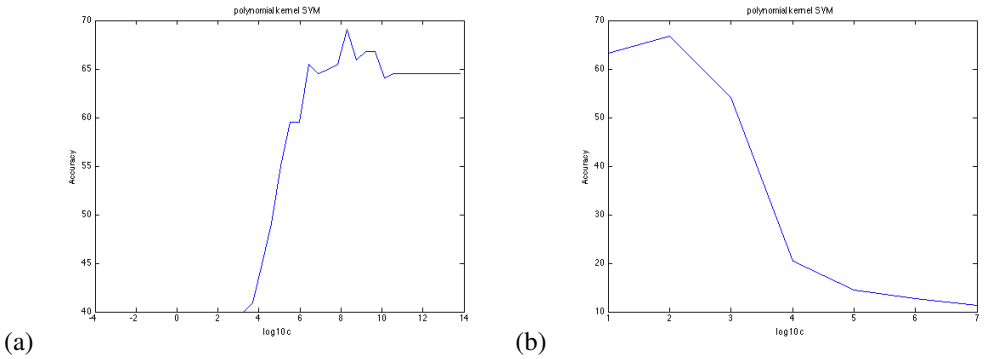


Figure 3: Accuracy to penalty C after log transformation on the original feature.

Next, we show the results for applying SVM with polynomial kernel. Figure 4 shows that the performance of accuracy to penalty C is similar to RBF kernel and the optimal value of this parameter

is 16000. Besides, according to the experiment results, the optimal degree for the data is around 2. The reason might due to small data size that we use recently so that increasing model complexity will lead to overfitting problem. So we focus on simple model for recent data.



(a) (b)

Figure 4: (a) Accuracy to penalty term c. (b) Accuracy to parameter degree d

The reason why random forest and naive Bayes failed is that the training example for each species is now limited to 80. From the predicted classification labels from the results of random forest, we found that the identities in some of the class are all mis-classified into another class, which means that the tree is ill-trained and biased such that many cases will be made to the same decision.

References

[1] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

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

[3] F. Briggs, X. Z. Fern, and J. Irvine. Multi-label classifier chains for bird sound. *arXiv:1304.5862*, (abs/1304.5862), 2013.

[4] F. Briggs, X. Z. Fern, R. Raich, and Q. Lou. Instance annotation for multi-instance multi-label learning. *ACM Transactions on Knowledge Discovery from Data*, 7(3):14, 2013.

[5] F. Briggs, B. Lakshminarayanan, L. Neal, X. 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. *Journal of the Acoustical Society of America*, 131:4640–4650, 2012.

[6] F. Briggs, R. Raich, K. Eftaxias, Z. Lei, and Y. Huang. The ninth annual mlsp competition: Overview. In *IEEE International Workshop on Machine Learning for Signal Processing*, 2013.

[7] W. Chen, G. Zhao, and X. Li. A novel approach based on ensemble learning to nips4b challenge. In *Workshop on Neural Information Processing Scaled for Bioacoustics*, pages 195–197, 2013.

[8] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[9] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[10] O. Dufour, T. Artieres, H. Glotin, and P. Giraudet. Clusterized mel filter cepstral coefficients and support vector machines for bird song idenfication. In *Workshop on Machine Learning for Bioacoustics*, pages 89–92, 2013.

[11] O. Dufour, H. Glotin, T. Artires, Y. Bas, and P. Giraudet. Multi-instance multi-label acoustic classification of plurality of animals : birds, insects & amphibian. In *Workshop on Neural Information Processing Scaled for Bioacoustics*, pages 164–174, 2013.

[12] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 524–531. IEEE, 2005.

[13] G. Fodor. The ninth annual mlsp competition: First place. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–2. IEEE, 2013.

[14] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[15] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[16] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[17] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical science*, pages 382–401, 1999.

[18] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.

[19] M. Lasseck. Bird song classification in field recordings: Winning solution for NIPS4B 2013 competition. In *Workshop on Neural Information Processing Scaled for Bioacoustics*, pages 176–181, 2013.

[20] L. Massaron. Ensemble logistic regression and gradient boosting classifiers for multilabel bird song classification in noise. In *Workshop on Neural Information Processing Scaled for Bioacoustics*, pages 190–194, 2013.

[21] E. L. Mencıa, J. Nam, and D.-H. Lee. Learning multi-labeled bioacoustic samples with an unsupervised feature learning approach. pages 184–189, 2013.

[22] E. L. Mencía, J. Nam, and D.-H. Lee. Learning multi-labeled bioacoustic samples with an unsupervised feature learning approach. In *Workshop on Neural Information Processing Scaled for Bioacoustics*, pages 184–189, 2013.

[23] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine learning*, 85(3):333–359, 2011.

[24] E. Stattner, W. Segretier, M. Collard, P. Hunel, and N. Vidot. Song-based classification techniques for endangered bird conservation. In *Workshop on Machine Learning for Bioacoustics*, pages 67–73, 2013.

[25] D. Stowell and M. D. Plumbley. Feature design for multilabel bird song classification in noise. In *Workshop on Neural Information Processing Scaled for Bioacoustics*, pages 182–183, 2013.

[26] G.-R. Walther, E. Post, P. Convey, A. Menzel, C. Parmesan, T. J. Beebee, J.-M. Fromentin, O. Hoegh-Guldberg, and F. Bairlein. Ecological responses to recent climate change. *Nature*, 416(6879):389–395, 2002.

[27] X. Zhang, Q. Yuan, S. Zhao, W. Fan, W. Zheng, and Z. Wang. Multi-label classification without the multi-label cost. In *SIAM International Conference on Data Mining*, pages 778–789. SIAM, 2010.