Identification of Songbird Species in Field Recordings

Hsiao-Yu Tung, De-An Huang, Xiao-Feng Xie, Yurui Zhou

Introduction

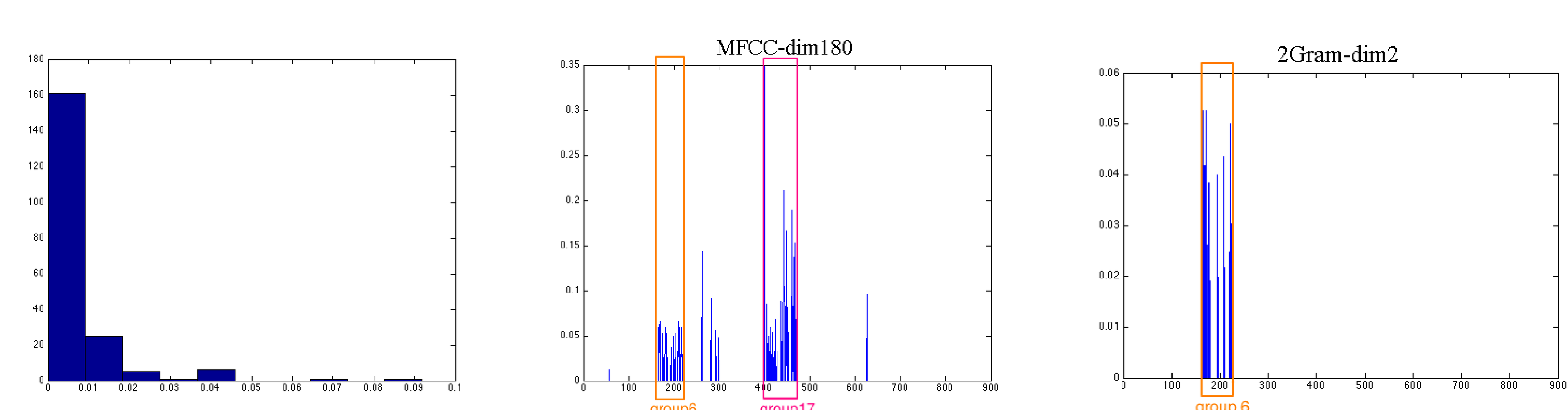
It is important to gain a better understanding about the climate and ecological changes in the world. One way to address this is to study seasonal migration patterns in songbird populations, since birds respond quickly to environmental changes During migratory periods, many species of songbirds use flight 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. Although there are some recent advances 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.

Features

- Spectrogram based[a] (MLSP 2013 Bird Classification Challenge Baseline):
  - Mask descriptors:
    - min-f, max-f, bandwidth (min-max), duration (T)
    - area, perimeter, non-compactness, rectangularity
  - Profile statistics:
    - gini, mean, variance, skewness, kurtosis,
    - area, perimeter, non-compactness, rectangularity
  - Histogram of gradients (HOG)
- Mel-Frequency Cepstrum Coefficients (MFCC) based:
  - Has been successful in speech recognition.
  - 39 dimensional vector. First dimension is energy.
  - $T \times 39$ matrix M for each audio file (T is not fixed.)
  - Continuous features: $\frac{1}{T}\sum_t M_t$, $M_{t_{max}}$, and first PC of M
  - Discretized features: Quantize MFCC by k-means. ($K = 200$)
    - Bag-of-words: 200-D histogram
    - N-gram ($N = 2, 3$): 200^N-D histogram. Select occurrence ≥ 3.
  - Denoising: Only use t with energy above threshold.

[a] F. Briggs *et al.*, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *Journal of the Acoustical Society of America*, vol. 131, pp. 4640–4650, 2012

Experiment: Classifiers

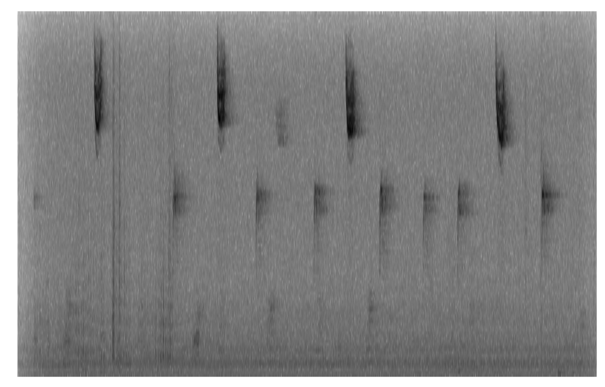


Acuuracy of different classifiers:
The accuarcy for classifier (using Random Forest) on mask descriptors by Briggs *et al.* is 0.25.

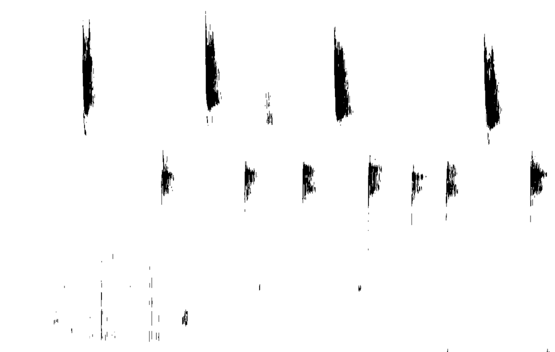
Classifier	Accuracy	Features	Settings
linear SVM	67.7273/ 70.4545	BoW/ denoised	
poly SVM	69.0909/ 70.4545	BoW/ denoised	degree: 1
	70	denoised BoW	degree: 2
	70.4545	denoised BoW	degree: 3
rbf SVM	70/ 70.4545	BoW/ denoised	
	68	BoW (log)	
	76.8182	denoised BoW	$\gamma = 7.9433$
	78.1818	denoised BoW + 2gram	
sigmoid SVM	70.9091/70.4545	BoW/ denoised	
random forest	57/63	BoW/ denoised	100 trees; 5 splits
	63	denoised BoW	100 trees; 2 splits
NN-euclidean	54.09/62.73	BoW/ denoised	
NN-chisquare	62.27/71.82	BoW/ denoised	

Preprocessing/Segmentation

We first convert audio files into spectrogram images, and for each segments we use Hanning windows with %75 overlap. 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 median clipping algorithm works best because it not only removes most background noise, but also capture the sound feature clearly and precisely.

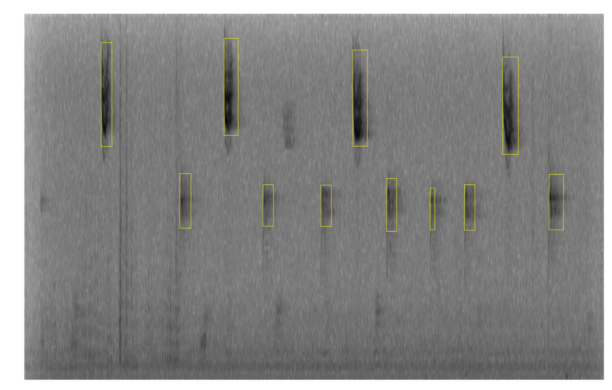


a. Original Spectrogram



b. Median Clipped

c.Eroded and Propagated



d. Labeled

Classifiers/Ensembles

The classifiers we used are

Nearest Neighbor (NN). We used euclidean and χ^2 distance.

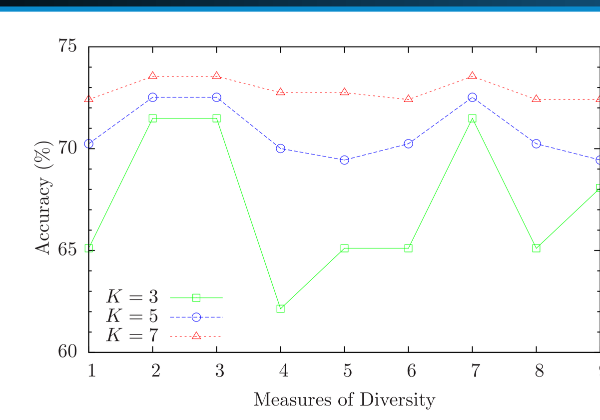
Support Vector Machine (SVM). This is the most common approach for multi-label classification. We tried linear SVM, sigmoid SVM and SVM with polynomial and rbf kernel.

Random Forest. Random Forest is operated by constructing decision tree structure by the training examples.

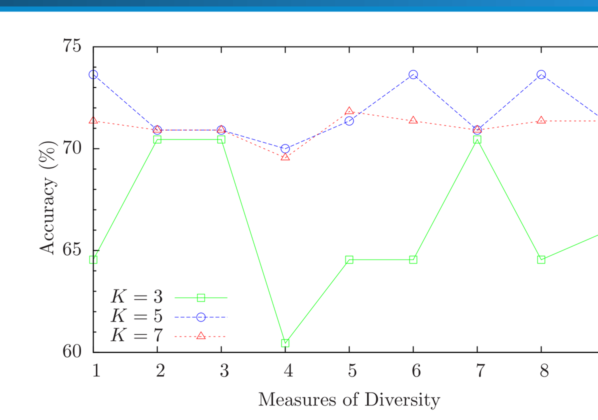
Ensemble Learning=$< F$-strategy$(\tilde{K}), C$-strategy$, B$-strategy$, \bar{V} >$.

- F-strategy: return a binary array F of chosen classifiers, using nine measures of diversity, i.e., *disagreement*, *correlation*, *Q-test*, *double-fault*, *coincident failure*, *entropy*, *interrater agreement*, *Kohavi-Wolpert*, and *generalized diversity*.
- C-strategy: assign weights C, using {*uniform*, *performance*, *optimization*}.
- B-strategy: obtain the belief matrix B, using *basic* and *new* forms.
- $\bar{V}$: enhance accuracy on most sure instances, leave others "UNKNOWN".

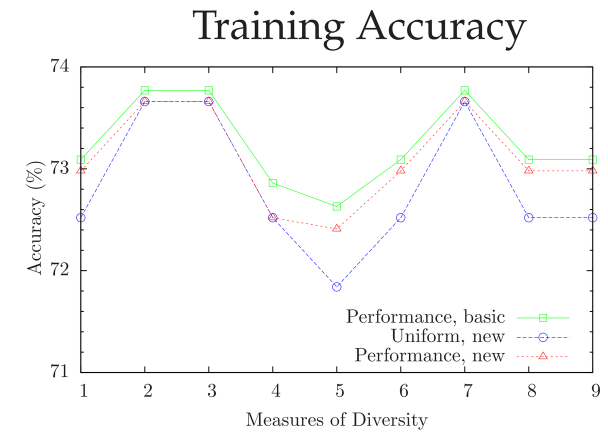
Experiment: Ensembles



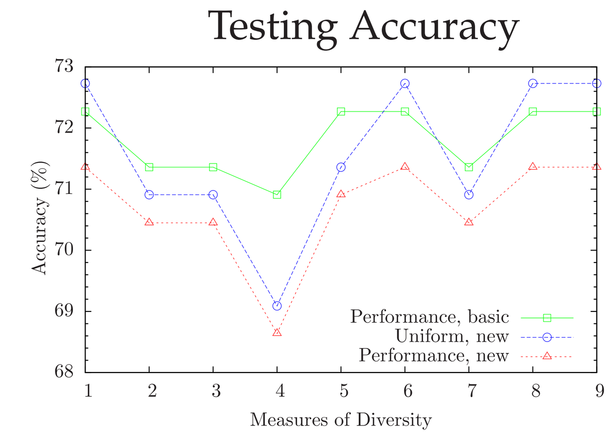
Training Accuracy



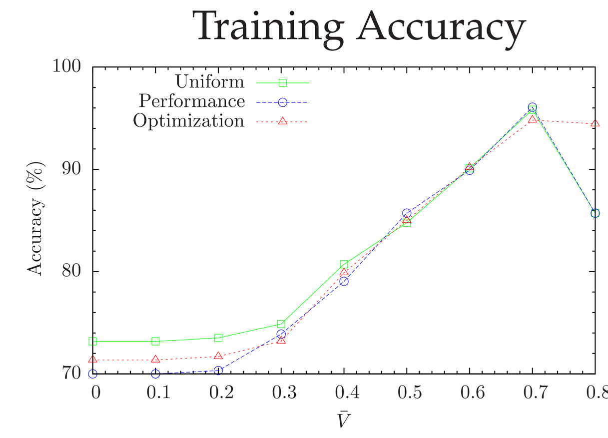
Testing Accuracy



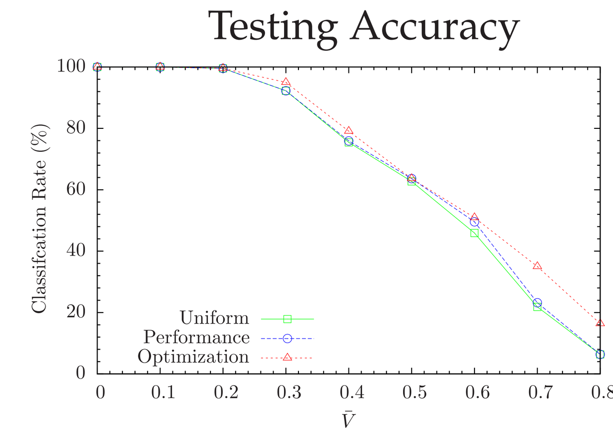
Training Accuracy



Testing Accuracy



Classification Accuracy



Classification Rate