International Journal of application or innovation in engineering and management

ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

Using artificial intelligence to identify birdsong in natural environments

Vijaya Varma L, Jaipur, India

Vijaya Varma76@gmail.com

Article Info

Received: 17-09-2020 Revised: 21 -10-2020 Accepted: 06-11-2020 Published:15/12/2020

ABSTRACT:

In this study, we simulated a real-world scenario to see how well an automated system could identify bird sounds. We rigorously avoided any kind of human interaction during the training of our classification algorithms using a crowdsourced dataset consisting of audio recordings of birds. Therefore, the method may be used to analyze collections of different species with the use of crowdsourced collecting for species labeling. We used a realistic number of candidate classes, representative of the kinds of numbers found in the actual world, to test the bird sound identification system's performance. Approaches: Within the training dataset's crowdsourced recordings, we used a threshold selection approach to distinguish between clear bird sound and quiet. The test data was derived from carefully selected recordings and was selected to align with an application scenario where the user chooses a segment of pure bird sound and feeds it to the identification system without any further information. Because of their popularity and relative simplicity, we looked at two classic classification methods: a k-Nearest Neighbor (kNN) classifier that uses histogram-based features and a Support Vector Machine (SVM) that uses time-summarization features. In order to make the class judgments more reliable and easier to understand, we looked at using a certainty measure that was based on the classifiers' output probabilities.

• Outcomes: Even though we claim that the k Nearest Neighbor classifier provides somewhat more flexibility, our findings show that both identification approaches performed similarly. In addition, we demonstrate that using an The reliability of categorization findings may be valuablely and consistently gauged by using the outcome certainty measure.

• Broader implications: Our research into probabilistic classification methodologies and use of generic training data directly contribute to the creation of a practical system for identifying bird sounds, which could have global applications. This system can adapt to the field's variable number of candidate species and classes. We go even beyond by demonstrating how the certainty metrics linked to identification results may greatly enhance the system's overall practical utility.

Keywords: Automatic bird sound recognition, Crowd-sourced training data, Machine Learning, Real-world performance evaluation, Flexible geo-temporal species selection,Probabilistic uncertainty measure, Classification rejection option

International Journal of application or Innovation in engineering and management

ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

Introduction

Over the last decade rapid advances in the sensing capabilities, data storage, network connectivity and computation power of mobile devices have occurred. This has been recognized as a unique opportunity for the deployment of biological recording systems to detect and identify biodiversity using mobile device applications (August etal. 2015). The widespread user base of such devices, as well as the ability to geo-tag & time-stamp observations and to exchange data makes the intelligent gathering of biodiversity information at a massive scale a real possibility. The potential of such methods to underpin the design of data-driven species population models, ecosystem sustainability evaluation and biodiversity protection strategies has been realised in publicly available, either automatic or crowd-sourced, identification applications for

birds (ChirpOMatic, Warblr, BirdSongID, Merlin Bird Photo ID)[1], bats (BatMobile, Echo Meter Touch Bat Detector) [2] , cetaceans (Automatic Whale Detector) [3] , insects (Cicada Hunt)[4] and plants (Plantifier, NatureGate, Leafsnap, Plantsnap, Wild Flower Id)[5].

The use of such approaches to enable automatic bird sound detection and identification started receiving research attention some two decades ago (see e.g. Anderson et al. 1996). In recent years there have been a number of studies carried out to investigate the effectiveness of different audio feature extraction methods and classification algorithms for bird sound identification (see for example Somervuo et al. 2006; Brandes 2008a,b; Trifa et al. 2008; Acevedo et al. 2009; Kirschel et al. 2009; Lakshminarayanan et al. 2009; Farina et al. 2011; Towsey et al. 2013; Wimmer et al. 2013; Stowell & Plumbley 2014b). Identification methods have also been used in the context of audio identification of vocalisations from species, such as bats (see e.g. Walters et al. 2012; Zamora-Gutierrez et al. 2016). More recently, the development of new recognition methods is mainly carried

International Journal of application or innovation in engineering and management
ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

and presented as part of online competitions[6] (see Stowell et al. 2016; Xie et al. 2018 and references therein). The method of choice in these more recent developments is predominantly that of deep convolutional neural networks operating on 2D spectrograms treated as images. Our work does not aim to investigate'best classification' in the narrow machine learning sense. We consider instead the pipeline for creation of detection algorithms, focusing on the role of data for both training and testing as well as issues of labelling.

First, with the exception of Lopes et al. 2011, Chou & Ko 2011, Stowell & Plumbley 2014a,b and the most recent online BirdClef competitions, most of the research undertaken to date presents automated classification results associated with approximately 5 to 20 species of birds. This number of candidate classes/species is far below the number of vocalizing avian species that are likely to be encountered in the field. For example, data from the BirdTrack project[7] for the region of Oxford, UK (an area of nine 10-km squares around the city of Oxford) collected by the British Trust for Ornithology list a total of approximately 240 species in an annual cycle[8] with the number of occurring species per week ranging from approximately 80 to 140 (with a mean of 110). An automated classification system deployed in the field would therefore have to operate on a considerably larger set of species than typically investigated in most studies so far and probably around 50-100 classes/species in a typical deployment scenario.

Second, for a species identification method to be scalable to global application, the classifier's training method needs to be applicable to as many of the total number of vocalizing avian species -namely some 10000 species that are found worldwide.

assembling training data for such a number of classes and, more importantly, relying on a few individual experts for segmenting and annotating adequate volumes of training data, is thus a major undertaking and not practicable for most regions of the world. This fact renders inapplicable a large number of classification methods presented in the literature so far, especially those that rely on manual segmentation to the syllable or phrase level (see e.g. Anderson et al. 1996; Chen & Maher 2006; Vilches et al. 2006; Leeet al. 2008; Trifa et al. 2008; Acevedo et al. 2009; Vallejo & Taylor 2009; Wei & Blumstein 2011; Aide et al. 2013; Tsai et al. 2014; Tsai & Xue 2014). On the other hand, a publically available database of bird sounds exists, with the sounds both recorded and annotated ina crowd-sourced manner - albeit not at the syllable or phrase level- and provides nearly worldwide coverage of bird species data which is constantly enriched. This database isthe xeno-canto project (xeno-canto.org). The identification methods we consider in this paper rely on data obtained from the xeno-canto data source in a manner that is directly and automatically repeatable for any selection of bird species from any region of the world.

Third, we find that classification methodologies that allow the straightforward and flexible use of time-of-year and location information for the selection of candidate classes (bird species) on-the-spot, for example the *k* nearest neighbours (*k*NN), have received virtually no attention in the existing literature. Falling under the broader category of 'Instance-based learning' (IBL) classifiers (Mitchell 1997), *k*NN classifiers do not generalise to classification rules during training but rather store training instances and defer further action until a test instance appears for classification. Such information can significantly reduce the number of candidate classes from several thousands (worldwide) down to a few hundred (e.g. in the U.K.) and even fewer (e.g. 50-100 for a given week anda given geographic location in the U.K.). In turn, this can help avoid the discriminative

performance degradation due to exceedingly confusable classes (Gupta et al. 2014), make the per-class balancing of training data easier by considering underrepresented species only when necessary and also make trivially easy the incorporation of more training instances as they become available without the need to retrain or implement incremental learning methods.

The overall aims of this paper were therefore to examine the utility of using simple, but flexible, classification methodologies combined with publicly available crowd-sourced training data to aid automatic species identification of bird song excerpts and to evaluate the expected level of performance of such practical applications.

The contributions of this paper are summarised as follows:

• We describe the 'blueprint' of a practical automated bird sound identification system which is directly applicable worldwide.

• We present classification results that directly indicate the expected level of performance and usability of such a bird species identification system.

• We introduce a probabilistic measure of uncertainty associated with the classification output and discuss how this can be used to increase the reliability of the identification results.

Methods

Test and Training Audio Recording Data

We used recordings from the 'Reference Animal Vocalisations' section [9] of the Animal Sound Archive dataset as test data (hereafter denoted as RAV recordings). The recordings included in the RAV collection have been manually annotated by the curator

International Journal of application or innovation in engineering and management
ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

contained at least 10 open access recordings. This resulted in a collection of 4132 recordings from 99 species with duration ranging from 0.35sec to 36sec and median duration of 2.26sec (see Table 2 in the Appendix, included in the supplementary material, for details). In terms of a real-world application, this kind of test data correspond to the case where the user singles out a recorded excerpt of clean bird sound (e.g. recorded on their mobile device and 'chopped' from the beginning to the end of a bird vocalisation) and provides it to the recognition system in order to identify the species. Modern mobile devices with large touch display interfaces make such a process feasible on the field.

We used recordings from the xeno-canto online dataset[10] (hereafter denoted as XC recordings) to train the classifiers we investigated. This database currently contains more than 270000 recordings from around the world with tens of thousands of new recordings added every year. For approximately 9300 species there is at least one recording annotated as dominantly containing the corresponding species (species in the 'foreground'). The mean of 'foreground' recordings per species is approximately 25 and the median is 10; there are more than 4600 species represented by at least 10 'foreground' recordings. By selecting XC recordings given a 'Quality A' (highest) ratingand marked as having no other species in the background we obtained a collection of 6182 recordings of duration ranging between 0.73sec and 71min42sec and with median duration of 44.7sec (see supplementary material for details).

Even though the selection of species used for our experiments was largely dictated by the availability of publicly accessible, reliably annotated test instances (as was the case

with the RAV dataset), it offers a quite good indication of bird species prevalence inEurope. Taking as an indication of prevalence the number of recordings per species currently available on the xeno-canto database, and with the exception of the Canary Islands Chiffchaff (*Phylloscopus canariensis*) which is not a European species, the remaining 98 species used in our experiments contain the 14 most frequent and 18 of the 20 most frequent species in xeno-canto. Out of the total number of 104100 recordings of the 731 European species in xeno-canto, 50914 recordings, nearly the half, are from the 98 European species of our collection.

Pre-processing and feature extraction

The audio features we used were based on standard FFT-based spectrograms. We extracted spectral statistics from separate spectrogram frames (frame-level features) and consequently aggregated these over collections of spectrogram frames to create feature vectors. For the test (RAV) recordings, frame-level features were computed for the whole length of the recording while for the training recordings we applied a frame selection method. This method is a modification of that used in the works of Briggs et al. (2009) and Stowell & Plumbley (2014a) and is described in the appendix (see supplementary material). Four types of frame-level features were considered, all choices from the various types described in Briggs et al. (2009) and Stowell & Plumbley (2014a). These were the (i) mean (denoted here as f_{mean}); (ii) standard deviation (f_{std}); (iii) mode (f_{mode}); and (iv) difference between the mode of two consecutive frames (Δf_{mode}). The feature vectors that were used for classification were consequently determined by computing binned histograms (for the IBL classifier based on Briggs et al. 2009) and time- summarisation statistics (for the SVM classifier based on Stowell & Plumbley 2014a). For the former case we consider 100x50-bin two-dimensional histograms of the pairs (f_{mean} and f_{std}) and (f_{mode} and Δf_{mode}) frame-

International Journal of application or Innovation in engineering and management

ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

level features and 100-bin one-dimensional

histograms of the f_{mode} frame-level feature. For the latter case we use the 6-dimensional time-summarisation features described in Stowell & Plumbley (2014a) comprising the 5th, 50th and 95th percentiles of the f_{mode} and the 50th, 75th and 95th percentiles of the Δf_{mode} frame-based features (see the appendix for details on the spectrogram computation parameters).

In both cases (histogram binning and time summarisation) the feature vectors used for training were obtained from sequences of 100 frame-level features (from one or more XC recordings) selected by the power threshold method mentioned above. With the chosen spectrogram parameters (listed in the appendix), each such sequence corresponds to between 1sec and 2sec of clean bird sound. The feature vector for each test instance (RAV recordings) was computed by histogram binning or time summarisation over the whole length of the recording.

The total number of frames selected by use of the power threshold method from the XC recordings for each species is listed in Table 2 in the appendix. We balanced the training dataset by subsampling according to the class with the fewer selected frames. As can be seen in Table 2, this is *Emberiza pusilla* for which the selection method returned in 1830 training frames. Following the procedure described above, 18 '1sec' traininginstances of 100 frames each were randomly selected (without resubstitution) for each species. We also investigated a second selection of species, namely the 72 species for which the power threshold selection method returned at least 20000 frames (again listed in Table 2). In this case, 200 training instances were used, again comprising 100 frames each. The corresponding test dataset (RAV recordings) for the collection of 72 species comprised 3354 RAV recordings (see Table 2).

International Journal of application or innovation in engineering and management

ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

Classifier and performance evaluation methods

The IBL-type classifier that we investigated is a *k*NN classifier. It draws from the work presented in Briggs et al. (2009) where a nearest neighbour classifier with time-distribution histograms as feature vectors was tested successfully for classification of 6 species. In the present work we modified this method to a *k*NN voting scheme with a tie-breaking rule (rather than a single nearest neighbour methodology as was used in Briggs et al. (2009)). We compared the results of this modified classification scheme against a more recent work (Stowell & Plumbley 2014a) in which a Support Vector Machine (SVM) classifier was used in a setup that again adheres to the practical application requirements outlined in the introduction (i.e. training data obtained from field recordings only annotated at the recording level, with no manual segmentation to the phrase or syllable level).

It is important to note that in both the aforementioned works, the training and test data come from the same dataset, although some care was taken to avoid training and test data sharing an individual bird origin. In an effort to better investigate real-world conditions of application, in our study we used completely separated training and test data.

We used the L1 (Manhattan) distance between histogram feature vectors for the *k*NN classifier (we also investigated the use of the Kullback–Leibler divergence and an approximation to the Hellinger distance as described in Briggs et al. (2009) in small-scale tests and the performance was not influenced considerably). We took the voting score with an addition of a tie-breaking bias (see details in the supplementary material) as the posterior probability of class membership in the case of the *k*NN classifier (Bishop, C. M. 2006 pp. 124-126). For the SVM classifier we used the Matlab interface of the LIBSVM

International Journal of application or innovation in engineering and management
ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

software package [11] with its default settings for multiclass probabilistic classification (Chang & Lin 2011). These settings amount to a radial basis function kernel and a "one-against-one" method for recasting the problem of training a k-class multiclass classifier to the training of $k(k\text{-}1)/2$ binary SVM classifiers. We used the estimate provided by LIBSVM for the posterior class-membership probabilities which is based on the methods described in (Wu et al. 2004 and Lin et al. 2007).

A concept that can be employed for the improvement of the performance and the usability elicited from a probabilistic classification scheme, is the separation of the probabilistic inference and classification decision stages and the introduction of a classification rejection region (Bishop, C. M. 2006). It is worth noting that despite the fact that both the kNN and SVM methods considered in the present work provide a probabilistic output, this is effectively disregarded in the classification stage with the decision made solely on the ranked list of probabilities and not their actual values. Further to that we note that, to the best of our knowledge, none of the works presented so far on the subject of automated bird sound identification have investigated this possibility.

In a binary classifier (and assuming constant gain for correct classification and loss for misclassification), the expected loss due to a wrong classification decision is related to how close to unity is the assigned class-membership probability (Bishop, C. M. 2006). For the application of this notion to a multiclass setup, we use here the *entropy* (Bishop, C. M. 2006) of the class-membership probability vectors as the classification rejection criterion. Being a measure of the information content encoded in the probability distribution of a discrete random variable, the entropy value can be used as an indicator

of the certainty associated with the classification decision. In simple terms, low values of entropy correspond to class-membership probability distributions that tightly peak in one class or a small number of classes (and are hence associated with higher certainty) whereas high levels of entropy correspond to distributions that are closer to the uniform distribution.

We used the Receiver Operating Characteristic (ROC) in a one-class-versus-all setup for the evaluation of the classification performance. In a binary probabilistic classifier, the ROC curve traces the points with coordinates equal to the achieved true positive and false positive rates as the value of the probability threshold discriminating the positive from the negative class ranges from 0 to 1 (Fawcett, 2006). The constructed Area Under Curve (AUC) metric ranges from 1 (absolutely correct assignment of instances to classes) to 0 (assignment of all instances to the opposite class) with a value of AUC equal to 0.5 corresponding to a classification result equivalent to chance assignment of test instances to classes. In the multiclass setup considered here, the AUC was computed taking each class in turn as the positive with the remaining classes taken as negative. Using a binary classification example of a highly non-balanced dataset, Davis & Goadrich (2006) show that the area under the Precision-Recall curve can be a more informative metric than that of the ROC curve. We include results of that metric again in a one-class-versus-all setup.

The AUC-ROC metric offers a method for the comparative evaluation of different classifiers' performance that is robust in the case of non-balanced test datasets and which is arguably superior to accuracy-based evaluation methods even when a balanced test dataset is used (Huang and Ling, 2005). However, its intuitive interpretation (namely, the probability of a randomly chosen negative test instance being ranked by the classifier lower than a randomly chosen positive test instance) does not lend itself to a direct gauge of the

International Journal of application or Innovation in engineering and management
ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

practical effectiveness of an identification system such as the one investigated here.

International Journal of application or Innovation in engineering and management
ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

If we consider the use case where the classifier returns an ordered list (of a chosen length *N*) of most likely species, the most easily interpretable measure of its effectiveness would be how often the correct species is within the returned list as a function of its length; a measure that we call *accuracy*@*N* and which we evaluate in the results section on a balanced test dataset.

Finally, a measure that is more widely established in the topic of information retrieval, and which unlike raw accuracy takes into account not only whether the correct class is within the returned ordered list of length *N* but also how high it is on that list, is the *mean average precision at N* (MAP@*N*) (Manning et al. 2009). In the case where thereis only one relevant retrieval option (as our classification setup) this metric becomes equivalent with the *mean reciprocal rank at N* (MRR@*N*) which is used in the results section below. The MRR@*N* metric ranges from 0 (in the case where the correct class is not within the *N* returned classes for any of the test instances) to 1 (when all test instances return the correct class ranked as highest).

Results

Figure 1 plots the summary statistics of the classification results (median, interquartile range and whole range of the vector of AUC numbers for the ROC curve for each class against the rest, as well as the mean of the per-class AUC-ROC vector weighted by the number of test instances in each class). Figure 2 plots the same statistics but this time for the AUC of the Precision-Recall curve. The same results are given in tabular from in Table 1. Furthermore, while noting that this is not a consistent evaluation metric in a non-balanced test set (see Huang and Ling, 2005), in Table 1 we list for completeness the corresponding accuracy scores.

International Journal of application or innovation in engineering and management
ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

The weighted mean and the median AUC-ROC for the histogram-based *k*NN classifier are practically constant over different choices of audio features and number of voting neighbours and marginally higher for the 72 species than the 99 species case. The AUC-ROC results for the SVM classifier with time-summarised audio features are slightly lower in all cases. The accuracy results show an increase in performance with increasing number of voting neighbours. Again, in terms of accuracy, the (f_{mode}, Δf_{mode}) features perform better compared to (f_{mode}) and (f_{mode}) perform better than (f_{mean}, f_{std}) . The same characteristics are associated with the AUC of the Precision-Recall results.

Despite the fact that the AUC-ROC results were nearly constant in the *k*NN experiments, there was a clear differentiation in the per-species profile of the performance achieved for different choices of features. For all three features cases (f_{mean} and f_{std} ; f_{mode} ; f_{mode} and Δf_{mode}), the Pearson correlation between the one-class-versus-rest AUC-ROC vectors obtained with the same features but different numbers of voting neighbours (1, 5, 11, 17 for the 99 species dataset) ranged from a minimum of 0.978 to a maximum of 0.998. For the 72 species dataset and with number of voting neighbours taking the values 1, 5, 11, 21, 51, 101, 201, the Pearson correlation between AUC-ROC vectors obtained with the same features ranged from 0.944 to 0.999 indicating that different species were recognised consistently better with different types of audio features. Contrary to that, keeping the number of voting neighbours constant and changing the features used, resulted in a Pearson correlation between the obtained AUC- ROC vectors that ranged from 0.576 to 0.893 and from 0.586 to 0.863 for the 99 and 72 species datasets respectively.

As can be seen in Table 1, the highest AUC-ROC score is achieved in 10 out of the 12 *k*NN setups for the Common grasshopper warbler (*Locustella naevia*). Garden warbler (*Sylvia borin*) and Hooded crow (*Corvus cornix*) scored the highest AUC-ROC in the

remaining two *k*NN cases of Table 1 while Canada goose (*Branta canadensis*) and European nightjar (*Caprimulgus europaeus*) scored highest for the two SVM cases. This level of performance was consistent for these species on a total of 13 parameter setupsfor the 99 species dataset (three feature cases and 1, 5, 11, 17 voting neighbours, plus the SVM setup) and 22 parameter setups for the 72 species dataset (three feature cases and1, 5, 11, 21, 51, 101, 201 voting neighbours, plus the SVM setup). A score of more than 0.95 ROC-AUC was obtained 12 and 22 times respectively for the Common grasshopper warbler. The same performance (more than 0.95 AUC-ROC), was obtained 9 times in the 99 species dataset for the Canada goose (that species was not in the 72 species dataset),4 and 19 times respectively for the Hooded crow, 4 and 14 times respectively for the Garden warbler and 1 and 8 for the European nightjar. Other than the species appearingin Table 1, the same level of performances was also obtained for the Eurasian wren (*Troglodytes troglodytes*) 13 and 18 times respectively and for the Common firecrest(*Regulus ignicapilla*) 12 times in the 99 species dataset (that species was not in the 72 species dataset). We did not find any systematic misclassifications between species pairs in our experiments.

In order to get a more direct measure of the performance, we also obtained the accuracy@*N* metric on a balanced test dataset. We used the same 99 and 72 species collections and the same training procedure with xeno-canto data that was described above. We selected 10 recordings for each species from the Animal Sound Archive dataset as the test set. Test recordings were chosen by minimizing the number of test instances labelled as coming from the same individual (the identifiers of the chosen recordings are provided in the supplementary material). In Figure 3 we plot the percentage of classification results in which the correct species was within the first *N* returned classification outputs, as a function of *N*. In these results we compared the SVM classifier

discussed above with one case of the *k*NN classifier, namely the one where one-dimensional histograms of the f_{mode} frame-level feature are used and the number of voting neighbours is set to 17 and 101 for the 99 and 72 species datasets respectively. The other parameter settings of the *k*NN classifier returned very similar results. In the case where only one class was returned by the classifier, the accuracy can be seen in Figure 3 to be equal to 9% and 6% for the *k*NN and SVM classifiers for classification among 99 species and 16% and 10% for classification among 72 species. The corresponding accuracy scores for a returned list of 10 species by the *k*NN classifier was approximately 40% (50% for classification among 72 species) and approximately 30% and 40% respectively for the SVM classifier. Interpreted with the expected user engagement in mind, this identification performance is still in need of improvement. When the number of candidate species is limited to 72 (a number which is at the lower end of the species expected to be encountered in the field) the correct species is expected to be the top result of the best performing *k*NN classification method approximately one in six times. Even when the identification scheme is allowed to provide the 10 most probable results (a list length that is already rather cumbersome to display on a mobile device) the correct species is expected to be within the returned list of the *k*NN classifier only half of the time.

In Figure 4 we plot the MRR@10 obtained for the same training and test datasets and the same classifier configurations as in Figure 3 as a function of the proportion of test instances that are classified using the entropy-based rejection criterion described in the 'Methods' section. The application of the rejection criterion is effected by taking a uniform grid of possible entropy levels ranging from 0 to the maximum level of entropy (equal to ln(99) = 4.59 and ln(72) = 4.28 for the 99 and 72 species cases) with points spaced at a distance of 0.1. For each of these entropy levels, instances associated with higher entropy

International Journal of application or Innovation in engineering and management
ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

are rejected prior to the determination of the MRR@10 performance.

International Journal of application or Innovation in engineering and management
ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

The MRR@10 metric when all instances are classified is equal to 0.18 and 0.12 for the *k*NN and SVM classifiers respectively in the 99 species dataset and 0.28 and 0.18 in the 72 species dataset. The performance is consistently rising as the rejection threshold becomes more stringent (i.e. the maximum allowed level of entropy for trusting a classification is reduced). It reaches the maximum MRR score of 1 albeit at very high classification rejection levels (approximately 99%) with the exception of the *k*NN results of the 72 species dataset that show a drop in performance at the 98% classificationrejection level. Starting at lower performance levels when all tests samples are accepted for classification, the SVM classifier shows better performance at higher classification rejection rates. It thus appears that the per-class assignment of posterior probabilities by the SVM scheme is more successful than the nearest neighbour scheme. All the aforementioned characteristics appear consistently in the other cases of frame-level features and number of voting neighbours settings of Table 1 (not considering the single nearest neighbour setting for which the entropy-based classification rejection criterion is clearly not applicable). The consistent behaviour displayed in these results suggests that entropy can indeed provide a quite useful classification reliability measure, but evidently, very high performance is only obtained when a very large percentage of testsis not classified

An evaluation of the introduced classification reliability measure which is more directly related to practical application is presented in Figure 5. In that figure we plot again the MRR@10 metric but this time as a function of the entropy threshold level (i.e. the maximum value of entropy over which the classification result for a test sample is not considered reliable and is not included in the evaluation). To account for the fact that the entropy scaling is different for different numbers of candidate classes, we normalise the entropy value by dividing it by the maximum possible value of entropy in each of the 99

International Journal of application or innovation in engineering and management
ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

and 72 classes cases. Such a normalised value of the reliability measure can be provided to the user of an application together with the list of most likely species returned by the probabilistic classifier. As can be seen in Figure 5, with the exception of the SVM results in the 99-classes experiment, all other cases show a consistent behaviour whereby a marked increase in reliability occurs for values of normalised entropy below 0.6 with values below 0.4 being associated with MRR@10 equal or exceeding a value of approximately 0.7. While there is clearly need for a more detailed investigation, these results suggest that a calibration of the introduced reliability measure which is consistent over different sets of candidate species may well be possible a least for the *k*NN method.

Discussion – Conclusions

The results presented in this paper are focused on classification of bird audio recordings among a large number of species (as is realistically required) making use of training data that are presently available for nearly global scalability. The rationale for undertaking this work was driven by a perceived knowledge gap in the expected level of performance of such a practical bird audio identification system. For example, among previous related investigations, the work presented in Lopes et al. (2011) covers audio recordings from 73 species from the Southern Atlantic Brazilian Coast but presents classification results for up to only 20 species. The performance of the methods they investigate is consistently reduced as the number of classes is increased (from 3 up to 20). Chou & Ko (2011) present classification results among 420 species of Japan birds. However, in this study little information is available about the characteristics of their training audio dataset and about the actual degree of train and test data separation in the experiments.

In addition to the SVM-based classification method presented in Stowell & Plumbley (2014a) (and which we partly replicated showing similar performance with the *k*NN method investigated here) the same authors have investigated (Stowell & Plumbley, 2014b) the use in bird audio recognition using the unsupervised feature learning method of Coates, A., & Ng, A. Y. (2012). This approach is applied to various classification experiments of recordings containing vocalisations from approximately 80 to 500 species with very positive results. On the other hand, their feature learning method seems to be closely tied to the particular instance of training data and selection of candidate classes.It is thus questionable if it can be easily applied to an ad-hoc selection of candidate specieswithout significant retraining requirements. The authors of that work also find that their method is possibly demanding in terms of the amount of training data required in orderto achieve its best performance.

The IBL classification scheme we investigated in this paper mainly draws from the method proposed in Briggs et al. (2009). The use of higher dimensional histograms with codebook clustering during the training process was also investigated in that work. The performance improvement that was quoted with that approach was rather moderate (increase in accuracy from 88% to 92% for the leading choice of parameters in their experimental setup while at the same time tying the training process to a particulartraining dataset and selection of candidate classes). On the other hand, we found that the introduction of the f_{mode} frame-level feature (position of maximum frequency) instead of the f_{mean} and f_{std} spectral statistics achieved the same performance in our experimental setup by using only one-dimensional histograms (and hence having reduced requirements in storage size and computation time). We also found that, with the use ofa tie-breaking bias addition, the performance of the histogram-based *k*NN classifier is practically constant over

International Journal of application or Innovation in engineering and management

ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

the number of voting neighbours. This seems to be in agreement

with the finding in Briggs et al. (2009) that taking into account the distance of more distant neighbours in a Bayes risk minimizing classifier formulation gives practically identical results with the single nearest neighbour approach.

Rather than offering gains in performance compared to the single nearest neighbour case, the extension to a *k*NN method (with a probabilistically interpretable output) allows the determination of an uncertainty measure of the classification result as an additional output. It is evident that among classification systems, it is the moderately performing ones (such as many current-day bird audio recognition systems) that can significantly benefit in terms of their practical usability from a consistent uncertainty measure of the output. At the expense of not having a result in many cases, a large number of very likely wrong classification results can be discarded and a smaller number of classification results can be relied upon. In our experimental setup, when all test instances are classified (no rejection option introduced), the combination of histogram- based features with a *k*NN classifier performs better than the SVM method using time- summarised features but this performance comparison is reversed at high classification rejection rates. We cannot yet conclusively determine whether this result is due to the different classifiers or the different types of features and whether it is systematic indifferent sets of species/classes.

As discussed, the flexibility of an instance-based classifier comes at the cost of increased data storage requirements and computation power in the classification stage. Taking as an example the xeno-canto training data used for the set of 99 bird species and the balancing subsampling of the training dataset in the experiments presented here, the required data storage (making use of efficient storing of the histograms' sparse arrays) ranges from less than 1MByte to approximately 35Mbytes for the different types of audio features used here. The processing time needed for feature computation and *k*NN

classification of the whole set of 4132 test instances (total time duration of 198 minutes) in the Matlab programming environment running on a PC with Intel Core i3-4160 CPU@3.6GHz, ranged from approximately 250sec to 500sec (for the different histogram feature and classification parameters considered here). This corresponds to a maximum of 40ms of computation for each second of recorded sound. Evaluating how these requirements would translate in a practical deployment for mobile devices is another important challenge.

The process that we described for the creation of the IBL classifier training instances from xeno-canto recordings is completely free of manual intervention. When combined with (i) the fact that such crowd-sourced data sources offer practically global coverage of bird species recordings and (ii) with the ability to choose the candidate species at run-time with no need for further classifier training, the IBL methodology that we present in this paper offers a blueprint for the development of a globally applicable bird sound identification system on mobile devices which readily provide geo-location information. Our further work plans include larger-scale experiments using more recently established standard datasets (such as the BirdCLEF dataset [12]) for the determination of the bearing of different feature sets and classifier parameters on the system's performance as well as the direct comparison with other classification methods. Our plans also include the use of existing detailed global coverage species distribution information (provided for research purposes by the 'Bird species distribution maps of the world' project [13]) for the scaling of those evaluation experiments to scenarios ofapplication to different geographical regions..

[12] http://www.imageclef.org/lifeclef/2017/bird
[13] BirdLife International and NatureServe (2015) Bird species distribution maps of the

International Journal of application or Innovation in engineering and management

ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

world. BirdLife International, Cambridge, UK and NatureServe, Arlington, USA.

In conclusion, in this paper we present evaluation results that directly quantify the expected performance of a practical automated bird sound identification system. We focus on the use case scenario where a mobile device user provides an excerpt of recorded bird sound along with geographical and temporal information provided by the device. The latter is used to select a sufficiently (but not unnecessarily) extensive list of candidate species. In our work, publicly available, crowd-sourced, training data was used in a fashion free from manual preprocessing. We couple this with an appropriately flexible classification scheme to provide a list of most likely species identification results. The current, constantly increasing, collection of bird audio recordings from the xeno- canto dataset allows the described identification scheme to have global application. The results we present compare favourably with previous work adhering to similar application requirements. We evaluate the application of a method for the improved use of the classifier's probabilistic output in refining the classification. There is, however, significant room for improvement in terms of the accuracy of the results presented to user in order for a system such as the one described here to be positively appealing and engaging. To that end, our current work is focused in tuning and optimizing the parameters of the training data selection and audio feature extraction methods in an effort to further improve the identification performance.

Acknowledgements

Dr. Timos Papadopoulos is supported by a James Martin Fellowship from the Oxford Martin School, University of Oxford. We are thankful to Dr. Andy Musgrove of the British Trust for Ornithology, Dr. Paul Jepson of the School of Geography and the Environment, University of Oxford and Dr. Karl-Heinz Frommolt of Museum für Naturkunde, Berlin for sharing data and for providing valuable insights and discussions.

International Journal of application or innovation in engineering and management

ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

References

In this work, Acevedo, Corrada-Bravo, Villanueva-Rivera, and Aide (2019) present

In 2009, T. M. Methods for automated machine learning-based amphibian and bird call classification: A comparative analysis. This sentence is a citation for an article written by Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G., and Alvarez, R. in 2013 that appears in Ecological Informatics, volume 4, issue 4, pages 206-214. Real-time bioacoustics monitoring and automatic species identification. Doi:10.7717/peerj.103 PeerJ, 1(1), e103.

- In 1996, Anderson, Dave, and Margoliash published a paper. Syllable identification for birdsong automatically using templates from continuous recordings. Publication date: 100(2), pages 1209-1219; DOI: 10.1121/1.415968. Published by the Acoustical Society of America.

This sentence is a citation for a paper written by August et al. (2015). Emerging technology for biological recording. [Online]. Biological Journal of the Linnean Society, Volume 115, Pages 731–749, doi:10.1111/bij.12534.

"Bishop, C. M." since 2006. Machine learning and pattern recognition. London: Springer.

(a) Brandes, T. S. Instruments for the automated recording and analysis of bird calls for use in conservation efforts. Vol. 18, Issue S163–S173, Bird Conservation International, doi:10.1017/s0959270908000415 (1995).

Brazilians, T. S. (2008b). Feature Vector Selection and Hidden Markov Models for Detecting Bioacoustic Signals with Frequency Modulation in Noisy Environments. Reference: Audio,

Speech, and Language Processing, IEEE Transactions on, vol. 16, no. 6, pp. 1173–1180, doi:10.1109/tasl.2008.925872.

An article published in 2009 by Briggs, Raich, and Fern (X. Z.). A Statistical Manifold Approach to Audio-Based Species Classification in Birds. In the field of data mining, about 2009. The 2009 IEEE International Conference on Data Mining (ICDM 2009) (pp. 51-60). Journal Article: 10.1109/icdm.2009.65

In 2011, Chang and Lin published two articles. LIBSVM: A Support Vector Machine Library. 2-27. Published in the ACM Transactions on Intelligent Systems and Technology. The DOI for this article is 10.1145/1961189.1961199.

Publication date: 2006 by Chen and Maher. Semi-automatic spectral peak track-based vocalization categorization in birds. Article number: 120(5), pages 2974–2984, published in the Journal of the Acoustical Society of America. the scientific reference is "10.1121/1.2345831"

In 2011, Chou and Ko published a paper. The extraction of syllable features from bird songs automatically using MFCC. With C.-H. Hsu, L. T. Yang, J. Ma, and C. Zhu at the helm, this volume contains the proceedings of the 8th annual international conference on ubiquitous intelligence and computing, UIC 2011. It spans pages 185 to 196. Authors: Coates and Ng (2012) Published by Springer. Learning Feature Representations using K-Means. Chapters 561–580 of Neural Networks: Tricks of the Trade, Second Edition, edited by G. Montavon, G. B. Orr, and K.-R. Müller. Berlin, Heidelberg: Springer. Here is the link to the article: http://doi.org/10.1007/978-3-642- 35289-8_30.

(2006) in Davis and Goadrich. On the Connection Between ROC Curves and Precision-Recall Measures. Articles 233–240 from the ICML'06 proceedings, the 23rd annual international

conference on machine learning. Link: http://doi.org/10.1145/1143844.1143874 In 2011, Farina, Pieretti, and Piccioli published a study. Methods for long-term bird monitoring using soundscapes: a case study in Mediterranean Europe. Environmental Informatics, vol. 6, no. 6, pp. 354-563. Published online: July 4, 2011 the work of Fawcett, T. (2006). What is ROC analysis? A primer. Journal of Pattern Recognition, 27(8), 861-874. Citation: doi:10.1016/j.patrec.2005.10.010. Abstract:

The authors of this work are Gupta, Bengio, and Weston (2014). Train Multiclass Classifiers to Perform Well.

Volume 15, Issue 15, Pages 1461–1492, Journal of Machine Learning Research.

Based on research by Huang and Ling (2005). Assessing learning algorithms using area under the curve and precision. Knowledge and Data Engineering: An IEEE Transactions Journal, 17(3), 299-310. Article DOI: 10.1109/TKDE.2005.50.
- Kirschel et al. (2019) have published a study in which they detail their findings.

In 2009, E. One Mexican antthrush, Formicarius moniliger, may be identified from its song by comparing four different categorization schemes. Volume 19, Issue 2, Pages 1–20, Bioacoustics: The International Journal of Animal Sound and Its Recording.
- AIn 2009, Lakshminarayanan, Raich, and Fern published a paper. Identifying Bird Species using a Probabilistic Framework at the Syllable Level. Pages 53–59 are from the Proceedings of the Eighth International Conference on Machine Learning and Applications, edited by M. A. Wani, M. Kantardzic, V. Palade, L. Kurgan, and Y. Qi. Computer Society of Los

Alamitos, 2009, p. 79. doi:10.1109/icmla.2009.79.

In 2008, Lee, Han, and Chuang published a paper. Two-Dimensional Cepstral Coefficients for Automatic Species Classification in Bird Songs. this article was published in the IEEE Transactions on Audio Speech and Language Processing, volume 16, issue 8, pages 1541–1550, with the DOI number 10.1109/tasl.2008.2005345.

- Weng, R. C., Lin, C.-J., and Lin, H.-T. (2007). Consideration of Platt's Probabilistic Results for SVMs with a Remark. Article 68(3) of Machine Learning, pages 267-276, with the DOI 10.1007/s10994-007-5018-6.

This information is sourced from a 2011 publication by Lopes, Gioppo, Higushi, Kaestner, Silla, and Koerich [1]. A System for the Automated Species Identification of Birds. The 2011 IEEE International Symposium on Multimedia (ISM) (pp. 117-122). Article DOI: 10.1109/ism.2011.27

The authors of this work are Manning, C. D., Raghavan, P., and Schütze, H. Introduction to Information Retrieval. Published by Cambridge University Press, with the DOI of 10.1109/LPT.2009.2020494.

Mitchell, T. M. (1997). Artificial Intelligence. Doi:10.1145/242224.242229 McGraw-Hill, 2009.

In 2006, Somervuo, Harma, and Fagerlund published a paper. Bird Song Parametric Representations for Automated Species Identification. Journal of Audio, Speech, and Language Processing, 14(6), 2252-2263, published by IEEE. The doi for this article is 10.1109/tasl.2006.872624.

It was written by Stowell and Plumbley (2014a). An extensive study of frequency

modulation in bird song datasets, published in Methods in Ecology and Evolution, volume 5, pages 901–912, 2011. this article's DOI is 10.1111/2041-210X.12223/. In a 2014b publication, Stowell and Plumbley provide... Unsupervised feature learning significantly enhances automated large-scale bird sound categorization. (doi:10.7717/peerj.488.) PeerJ, 2(2), e488.

In 2016, Stowell, Wood, Stylianou, and Glotin published a document. Identifying birds in recorded sound: A survey and a test. Machine Learning for Signal Processing: 26th IEEE International Workshop

- In 2013, Towsey, M., Wimmer, J., Williamson, I., and Roe, P. published that. Assessing the abundance of bird species in field recordings using auditory indexes. Journal of Ecological Informatics, 100. The accepted citation is "doi:10.1016/j.ecoinf.2013.11.007." The authors of this work are Trifa (2008), Kirschel (2008), Taylor (2008), and Vallejo (2008). Antbird species identification using hidden Markov models automated in a Mexican jungle. Volume 123, Issue 4, Pages 2424–2431, Journal of the Acoustical Society of America. 10.1121/1.2839017 is the DOI for reference.

- In 2014, Tsai, Xu, and Lin published a paper. Tibre and pitch characteristics for the purpose of bird species identification. Volume 30, Issue 4, Pages 1927–1944, Journal of Computer Science and Engineering

In 2014, Tsai and Xue published a paper. Regarding the Use of Voice Recognition Methods for Bird Species Identification. Press, 19(1), 55-68, in the field of computational linguistics and Chinese language processing.

- Vallejo and Taylor (2009). Sound monitoring of bird activity and variety using adaptive sensor arrays: first findings on source identification with support vector machines. Journal

of Artificial Life and Robotics, 14(4), pages 485-488. article DOI: 10.1007/s10015-009-0705In 2006, the authors yVilches, Escobar, Vallejo, and Taylor published a paper. Acoustic Bird Species Recognition By Data Mining. Included in the proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006), Volume 3, pages 400–403. this article's DOI: 10.1109/icpr.2006.426'

The authors of the following works are: Walters (C. L.), Freeman (R.), Collen (A.), Dietz (C.), Brock Fenton (M.), Jones (G.), and Jones (K.).

Eliza (2012). A system for the acoustic detection of bats throughout Europe. Pages 1064–1074, in the Journal of Applied Ecology, volume 49, issue 5, published in 2015. This is the online version of the article: http://doi.org/10.1111/j.1365-2664.2012.02182.x]. It was published in 2011 by Wei and Blumstein. Using hidden Markov models based on syllable patterns, we can robustly recognize bird songs as noise. This paper was presented at the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), with the abstracts included on pages 345–348. the working citation is "10.1109/icassp.2011.5946411"

Based on research by Wimmer, Towsey, Roe, and Williamson (2013). I am collecting ambient audio recordings to count the species of birds. Article published in Ecological Applications, volume 23, issue 6, pages 1419–1428. The DOI for this article is 10.1890/12-12088.1.

It was written in 2004 by Wu, T.-F., Lin, C.-J., and Weng, R. C. Estimates of Probabilities for Pairwise Coupling Multi-Class Classification. Press, 5(975), 975–1005. Journal of Machine Learning Research. the research paper's DOI is 10.1016/j.visres.2004.04.006. In 2018, Xie, J.-J., Ding, C.-Q., Li, W.-B., and Cai, C.-H. published a paper. Audio-only Bird

Species Automated Identification Method with Limited Training Data Based on Multi-Channel Deep Convolutional Neural Networks. E-prints on ArXiv, March 3, 2018 This sentence is a citation for a paper published in 2016 by Zamora-Gutierrez et al., with authors Macswiney Gonzalez, Fenton, Jones, Kalko, and Jones.

Tables

Classifier	Features	Voting neighbours	Per-species AUC (ROC) in one class v. all				Species of maximum and minimum AUC (ROC)		Per-species AUC (Precision-Recall) in one class v. all		Accuracy (%)
			Min	Max	Weig. Mean	Median	Min	Max	Weig. Mean	Median	
Results for 99 species dataset											
*k*NN (histograms)	f_{mean}	1	0.44	0.98	0.77	0.80	lox_cur	loc_nae	0.086	0.024	4.53
	f_{std}	17	0.42	0.98	0.77	0.79	lox_cur	loc_nae	0.110	0.035	6.17
	f_{mode}	1	0.50	1.00	0.78	0.81	emb_aur	loc_nae	0.084	0.031	6.58
		17	0.49	1.00	0.78	0.81	chl_chl	loc_nae	0.102	0.033	7.41
	f_{mode} Δf_{mode}	1	0.46	1.00	0.79	0.81	tur_pil	loc_nae	0.089	0.034	7.74
		17	0.46	1.00	0.79	0.80	tur_pil	loc_nae	0.112	0.037	8.66
SVM (summ. stats)	f_{mode} Δf_{mode}	-	0.31	0.97	0.73	0.76	lox_cur	bra_can	0.091	0.024	6.15
Results for 72 species dataset											
	f_{mean}	1	0.29	0.96	0.78	0.81	lox_cur	syl_bor	0.120	0.042	8.05

International Journal of application or Innovation in engineering and management

ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

	f_{std}	101	0.25	0.98	0.79	0.81	lox_cur	cor_cor	0.196	0.078	9.93
	f_{mode}	1	0.62	0.99	0.80	0.82	jyn_tor	loc_nae	0.118	0.045	9.33
		101	0.61	0.99	0.82	0.84	den_maj	loc_nae	0.173	0.069	11.24
	f_{mode} Δf_{mode}	1	0.58	0.97	0.79	0.82	tur_pil	loc_nae	0.130	0.056	13.06
		101	0.62	0.98	0.81	0.84	tur_pil	loc_nae	0.197	0.096	15.06
SVM (summ. stats)	f_{mode} Δf_{mode}	-	0.23	0.98	0.75	0.77	lox_cur	cap_eur	0.143	0.036	6.89

Table 1. Summary of per-species AUC performance metric obtained with the ROC and Precision-Recall curves in a one class v. all setup. Species binomial names in rightmost columns are abbreviated to 3 first letters for the genus and the species. The top 7 rows correspond to the 99 species dataset and the lower 7 rows to the 72 species dataset. In the right-most column we also give the accuracy score (in percent values) for each case.

International Journal of application or Innovation in engineering and management

ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

Figures

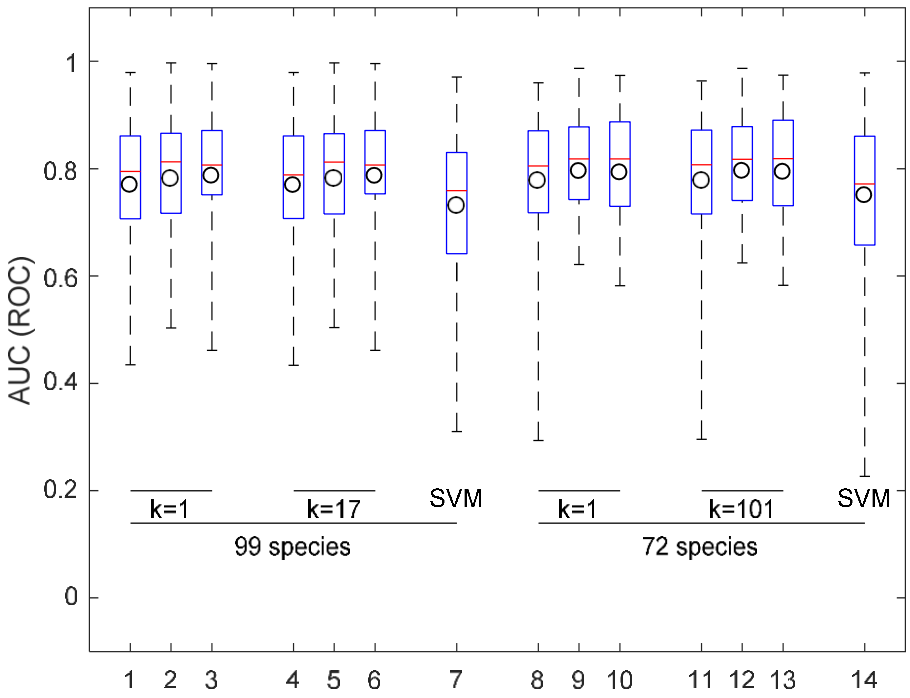

Figure 1: Boxplots are median, interquartile range and whole range of the vector of Area Under the ROC Curve performance for each class against the rest. The black circle is the mean of the per-class AUC vector weighted by the number of test instances in each class. The boxplots appearing in triplets correspond (in order from left to right) to the three cases of frame-level spectral features used in the histogram kNN classifier, namely (i) mean and standard deviation of the frame spectrum, (ii) position of the maximum frequency and (iii) position of maximum frequency and frequency modulation across successive frame couples. For the SVM case we use a 6 dimensional vector of summary statistics.

International Journal of application or Innovation in engineering and management
ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

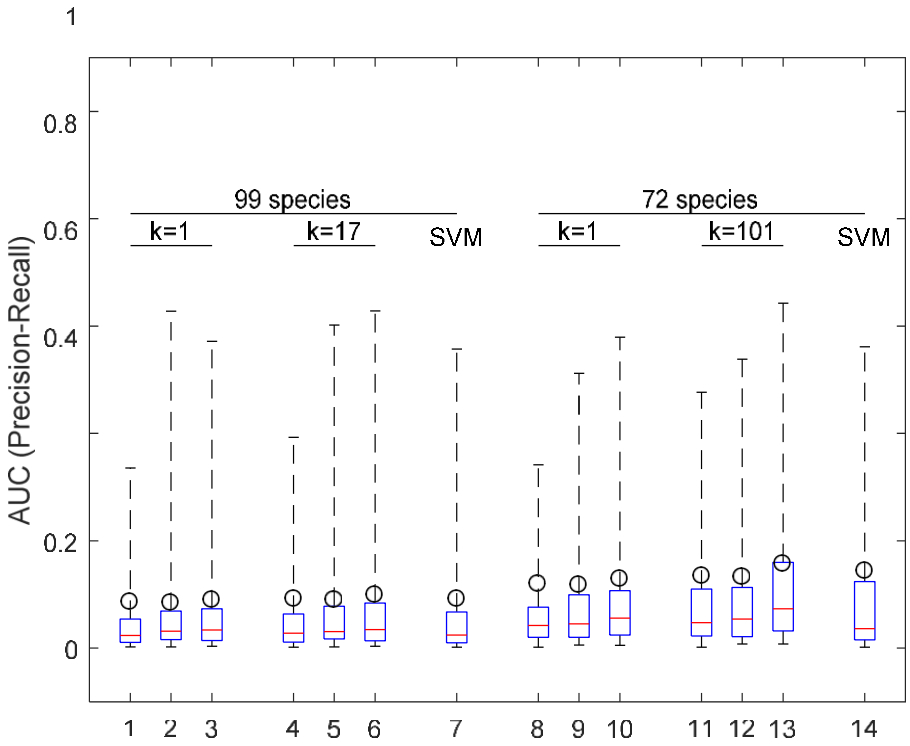

Figure 2: Same as in Figure 1 but for the Area Under the Precision-Recall Curve.

International Journal of application or Innovation in engineering and management
ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847

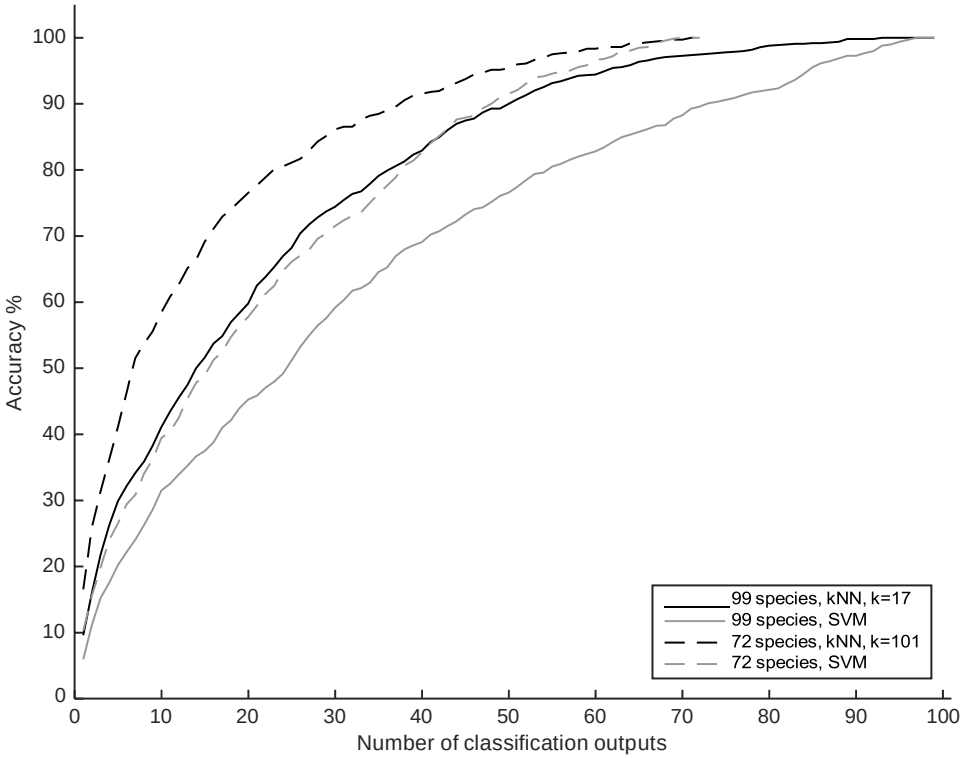

Figure 3: Accuracy in the case where the classifier returns a list of *N* most probable classes. The plotted lines are the percentage of classification results for which the correct class is within the first *N* classification outputs as a function of *N*.

International Journal of application or Innovation in engineering and management

ijaiem.com/Dec 2020/ Volume 9/Issue 1/Article No-1/34-73
ISSN: 2319-4847