Targeting Input Data for Acoustic Bird Species Recognition Using Data Mining and HMMs

Erika Vilches, Iván A. Escobar, Edgar E. Vallejo
Tecnológico de Monterrey, Campus Estado de México
Atizapan de Zaragoza, Edo. de México, 52926, México
{A00461595, A00447647, vallejo}@itesm.mx

Charles E. Taylor
University of California Los Angeles
Los Angeles, CA, 90095, USA
taylor@biology.ucla.edu

Abstract

In this paper we propose the integration of Data Mining with Hidden Markov Models when applied to the problem of acoustic bird species recognition. We first show how each of them is applied on an individual manner, contrast their results and devise a model to combine them for targeted classifications. Previous work has shown that large collections of spectral attributes are needed in order to represent the structure of bird songs, therefore elevating the computational requirements when applied to distributed sensor networks. Data Mining is used to reduce the dimensionality of the spectral attributes and for classification. Hidden Markov models represent a traditional approach and require strong song preprocessing. Our results show that Data Mining can yield efficient results with low requirements and could serve to target HMMs input parameters.

1. Introduction

The monitoring of animal behavior and diversity over a variety of spatial and temporal scales poses many challenges to human observers. A significant amount of knowledge on bird diversity and behavior is the result of field observations made by expert ornithologists. Bird species identification and the study of their interactions have been developed by means of the visual and acoustic abilities of these experts. On the other hand, the identification of individual birds often requires the usage of visual aids, such as color banding.

For this work, the goal of sensor networks [13, 17, 18] is to introduce in a natural environment a certain number of small sensors or motes in order to acquire data from their surroundings. The technology of distributed sensor networks also allows us to detect unusual events that occur in a certain environment, such as the presence of rare or endangered bird species, their communications and social interactions without human intervention.

To be able to explore the full potential of distributed sensor networks we must deal with complex processing challenges. Bird songs, when converted to the frequency domain by means of Fourier transforms, produce large amounts of data that requires processing and discrimination. Implicit relationships must be found; data must be sorted into logical groups and cleaned. Here is where data mining techniques will alleviate the computational needs for these analyses. This will allow us to incorporate these processing technologies into existing energy and processor constrained platforms, such as nodes in a sensor network.

Previous works have used canonical discriminant analysis [14] to demonstrate that invariant features don't actually provide the most important recognition cues, contradicting some common assumptions in the published literature. Hidden Markov Models have also been applied to bird song recognition due to their success in speech recognition, even though they are not appropriate for constrained platforms such as those in a distributed sensor network.

We propose to use data mining techniques to reduce the computational complexity required to classify different bird species by means of attribute reduction and to show that these techniques work better when combined with classical approaches such as Hidden Markov Models.

2. Data Acquisition and Procesing

2.1. Bird Songs

Bird songs and calls for this study were obtained from two different sources. The first was through the Cornell Lab of Ornithology, Macaulay Library [1]. We gathered samples from their collection and did initial testing. The second set of bird songs were collected from our February 2006 trip to the Reserva de la Biosfera de Montes Azules in Chiapas Mexico. We used the later songs to validate some of our

results. We focused on songs from three species of antbirds: great antshrike, *Taraba major* (49 song files); dusky antbird, *Cercomacra tyrannina* (79 song files); and barred antshrike, *Thamnophilus doliatus* (76 song files). Each song file has from a few seconds to several minutes of bird calls, with either one, two or more birds singing on it.

Bird songs are normally more musical and complex than calls. Males usually produce them and they are associated with breeding. "Calls tend to be shorter, simpler and produced by both sexes throughout the year. Unlike songs, calls are less spontaneous and usually occur in particular contexts" [4]. Birds use calls to communicate things to each other and between members of a flock or family.

For this work, we are going to use both songs and calls to extract the most relevant features of the bird's song. We decided to include both sources, since some of their song building blocks have common elements in both songs and calls. For the remainder of this document, we will mention calls and songs as being of equal importance and will be equally used as the basis of bird song data.

The reason to choose these species is that they are abundant in Montes Azules, Chiapas, the ecological reserve where the sensor network will be deployed in the near future. Additionally these tropical bird species do not learn songs, which makes the job of acoustic recognition easier. Finally, the three species share common building blocks in their song organization. This makes the challenge of classification very interesting, since if we compare their spectrograms in a very close up level, we notice that some of their song components are very similar.

2.2. Sound Cleaning

Table 1. Low-pass and High-pass filters per species

Filter	Taraba major	Cercomacra tyrannina	Thamnophilus dolitus
Low-pass	3597 Hz	4200 Hz	3597 Hz
High-pass	517 Hz	920 Hz	686 Hz

Field recordings can be extremely noisy, especially when performed in tropical rainforests. In these types of forests, the vegetation is densely packed causing sound reverberations; there are many different bird species interacting and a huge amount of other animals producing harsh noises. The climate is also a crucial factor, rain can cause significant interference and wind causes leaves to fall and interfere through most of the acoustic frequencies. All these factors limit the quality of the sound recordings making automated bird species recognition a more complicated process and requiring the introduction of different filtering techniques in order to obtain suitable results.

The songs were only preprocessed through low and high pass software filters to facilitate an accurate call and pulse recognition. These filters are species dependant as we can see in Table 1.

3. Methods

3.1. Feature Extraction

Once we finished the cleaning process, we decided to use .wav format files, since they had better resolution and were easily manipulated by computers. For this purpose, used the computer software, Sound Ruler [9]. With this software, we are able to see the oscillogram and spectrogram of the signal and within the oscillogram we are able to locate each call from the recording and each pulse within a call.

The pulse-by-pulse analytical results that Sound Ruler threw were saved as comma delimited files. These files contain the 71 attributes of each pulse from the processed samples, representing the bird's song data. The resulting datasets' size is as follows: *Taraba Major* 21,360 pulse samples, *Cercomacra Tyrannina* 5373 pulse samples, and *Thamnophilus Doliatus* 911 pulse samples.

3.2. Hidden Markov Models Theory

The basic HMM theory was published in a series of classic papers by Baum and his colleagues [5]. The HMM has become one of the most powerful statistical methods for modeling speech signals. Its principles have been successfully used in automatic speech recognition, formant and pitch tracking, speech enhancement, speech synthesis, statistical language modeling, part-of-speech tagging, spoken language understanding, and machine translation [3, 5, 6, 7, 8, 10, 11, 12, 16].

3.3. Data Mining

When working with bird songs, we unfortunately need to deal with information that is represented as raw data. This information may contain valuable records that may be hidden from the naked eye. We have to apply different computational tools in order to extract the information we require from the raw data. The approach we took was to apply different data mining techniques in order to obtain the most relevant information from the raw data. Once the important data is extracted, we can use only the significant information to feed our classifying algorithms in the sensor nodes in order to recognize different bird species based on their song and call production.

3.3.1 Vector Quantization

This algorithm was implemented because of the ID3's and association rules lack of numeric support. Quantization [15] is a process in which numeric to nominal data conversion is possible. The algorithm takes an original numeric vector and returns a quantized equivalent numeric vector, which can be easily represented by nominal values.

Mainly, this process consists of the making of ranges of numbers, and assigning each range a numeric value that represents the whole range: Lets say: 1 to 1000 = 1, 1000 to 2000 = 2, etc., and then interpreting the assignments as labels, instead of numbers. For the specific details, check [15].

In figure 1 we present a plot comparison from a full original signal with values from 0 to 5000 approximately versus a quantized signal with values from 0 to 6. We can clearly appreciate how the relationship among the attribute values is preserved in the quantized set, even though we can appreciate some information loss.
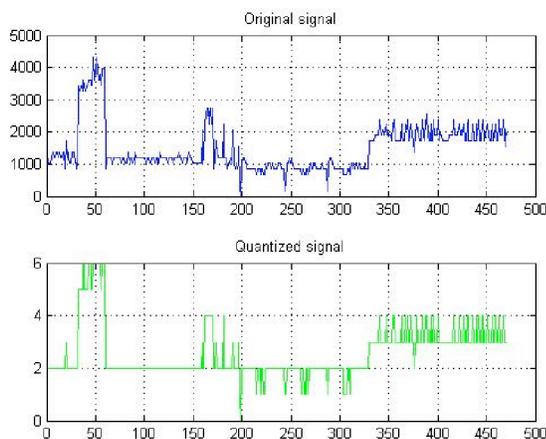
Figure 1. Original vs. quantized signal

3.3.2 ID3

Once we converted the entire species data sets into quantized data, we proceeded to process the information with a decision tree algorithm. The ID3 algorithm was used to generate the decision tree with Weka [21] software. ID3 is a decision tree algorithm that takes all unused attributes and counts their entropy concerning the test samples to be used.

When the tree is completed, the resulting nodes will be the most significant attributes used to classify the different instances of bird species (the leaves of the tree). Once we obtained the corresponding decision tree, we only preserved in our data set the attributes that were used in the nodes of the tree (an attribute can be repeated in many nodes). This reduced data set will be used to attempt a reliable classification with the Nave-Bayes algorithm.

3.3.3 J4.8

This algorithm is an extension of the ID3 algorithm, which solves some deficiencies that the original ID3 algorithm had. Some of the improvements are that J4.8 avoids overfitting, uses a reduced-error pruning focus that is based on the consideration that each node of the tree is a prune candidate, reducing this way the error. Also it performs a rule post-pruning to find high precision hypothesis and numeric attribute handling. The two main advantages that made us select this algorithm are the computational cost savings and the numeric attribute handling. Weka was used to test this algorithm with our original data sets. The surviving attributes in the reduced data sets were also used to attempt a reliable classification with the Nave-Bayes algorithm.

3.3.4 Nave-Bayes

We decided to introduce the Nave-Bayes algorithm usage as a final classifier because of the main disadvantages that decision tree algorithms have. One of them is that they are unstable. In effect, slight variations in the training data can result in different attribute selections at each choice point within the tree. The effect can be significant since attribute choices affect all descendent sub-trees. Another important disadvantage with decision trees is that trees created from numeric data sets can be quite complex since attribute splits for numeric data are binary. Nave-Bayes was executed in Weka, for the original, post-ID3 and post-J4.8 datasets.

In this work, we attempt to eliminate redundancy or dependency in data by means of decision trees (ID3 and J4.8). We use only its surviving attributes to construct the data set that will be fed into Nave-Bayes. This will assert that we are working only with independent attributes and thus assuring that the learning process is being skewed as less as possible by redundancy and that the maximum efficiency is being obtained.

3.3.5 Association Rules

For our experiments we took into consideration the association rules that had a 100% confidence and high instance support. An important factor to consider is that different association rules express different regularities that underlie the dataset, and they generally predict different aspects of the data [21]. This is why we need to analyze and interpret rules separately.

It is important to point out that even though high confidence values might suggest strong association rules, there

might be some deception on these results, since the antecedent or consequent might have high support values giving for this matter high confidence levels even if they are independent. We recall that our goal is to find relationships between the attributes that compose a bird's song.

To obtain information about the Association Rules' concepts used in this work (confidence, support, lift, conviction and leverage), please consult [21].

4. Results

4.1. Hidden Markov Models

Hidden Markov Models were explored in this analysis to contrast their accuracy on bird species' recognition against the data mining recognition approach. For this objective HTK [22] was used as a HMM testing implementation. HMMs [20] have been selected because they represent the traditional approach used nowadays for human speech recognition. For this purpose, whole songs (duration approx. 2-3 sec.) have been manually cut from audio recordings made in the natural reserve of Montes Azules in Chiapas, Mexico. Trifa, [19] a colleague and member of our laboratory, used 25 samples for each of the species of interest and different amounts of background noise. Songs have been selected with high variability's. This choice was motivated by the fact that the goal was to minimize the possibility that the HMMs would model properties of the noise instead of just the bird song itself. From the 25 samples of each species, 15 samples have been used to train the HMMs and 10 samples to test the recognition performance. These samples were selected randomly from the ones available.

The performance metric used to measure the efficiency of the HMMs is the one proposed in the HTK Book [22]. For each species or individual i tested, performance p is measured simply as the ratio of correctly recognized samples C_i over the total amount of samples N_i used for testing, and is defined as:

$$p_i = \frac{C_i}{N_i} \times 100$$

For each experiment, 50 runs were performed using random partitioning of the files. They were also randomly split into training and testing sets. After several runs, a window size of 25ms was selected and an overlapping range for these windows of 15ms. These values were chosen because they obtained better performance results.

As it was expected, the performance of HMMs is slightly superior when the frequency range of the bird's song is bounded according to each species spectral parameters. This happens because the noise located outside this frequency is not taken into consideration by the HMMs.

With these settings, it was found that the average HMM recognition performance for these species is about 93%. Several aspects cause this 7% lack of accuracy, among them it was identified that sometimes, birds do not sing their songs completely, they sing just a part of them. Also it was noted that ambient noise is one of the main error sources.

To find the optimal number of states required to model the species was a complex task. It was difficult, if not impossible, to relate this parameter with any physical property of bird songs. A value of 5 states was selected, since it was noted to work better on average for all tests.

Performance did not change in a significant way when using between 5 and 15 states, but adding more than 15 states actually degraded the performance. We realized that using the standard feature extraction proposed by HTK is a very generic procedure and might not reflect the features relevant for bird species' classification [19].
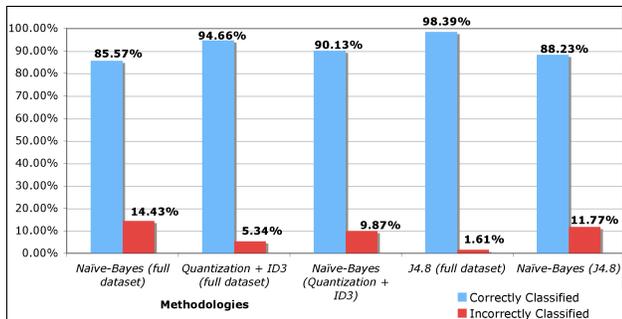
4.2. Data Mining

In figure 2 we can observe that the most accurate algorithm is J4.8 (without Nave-Bayes) obtaining a 98.39% of accuracy. The original attribute number was 71 which this algorithm reduced to 47. We can also appreciate that regarding Nave-Bayes, the reduced data sets produce a slightly better performance, up to 4.5% improvement.

Besides the reliable accuracy preservation, the required processing power is also directly affected from the attribute reduction, since the number of calculations needed to classify in a smaller data set is lower and so is the power consumption.

In the J4.8 tree, the main attribute was pulse dominant frequency, the root of the tree. In the next level we find the width of the dominant frequency peak at half of its height divided by the frequency of the peak. One more level down, we find the maximum of dominant frequency in the pulse, the total number of pulses in the call and the dominant frequency at final 50% peak amplitude. These five attributes which J4.8 identified as the main ones, contrast with the song duration, number of phrases and number of notes identified by Nelson [14] and also the speed, duration, frequency range, and center frequency identified by Bard [2]. The reasons of these disagreements probably are the use of songs from different bird species and different algorithms for attribute selection, such as canonical discriminant analysis.

4.2.1 Association Rules

Our tests were performed considering only results valid for instances with a 100% of confidence and a high degree for support. We knew that there was a high degree of association between some attributes of our instances. From our results, we can confirm that there is a high degree of association among attributes.

Figure 2. Data Mining accuracy percentages plot

Rule A: If the duration of a pulse on the initial axis (0%) of a call is high, the time in which half of the frequency modulation is low, and the second harmonic's relative amplitude is low, then the duration of a pulse during the first 10th% of the call is high.

Rule B: If the duration of a pulse on the initial axis (0%) of a call is high then the pulse duration during the first 10th% of the call is high, the time in which half of the frequency modulation is acquired in the pulse is low and the relative amplitude of the second harmonic is low.

Rule C: If the duration of a pulse on the initial axis (0%) of a call is high, the energy during the 90th% of the initial call is high and the relative amplitude of the second harmonic is low, then the duration of the pulse during the first 10th% of the call is high and the time in which half of the frequency modulation is acquired in the pulse is low.

We can observe that there is a mild degree of association from attributes from the consequent and the antecedent, reflected by a low leverage value. For example in rules A and B only 16% of the total instances (21,360) covered by the antecedent and consequent of those rules are totally independent. We can also observe lift values above one, indicating that the rules generated are valid.

We know that leverage and lift measure similar characteristics in a rule but leverage measures more precisely the level of co-occurrence of X in Y as independent probabilities. In other words, leverage measures the proportion of the cases covered by X and Y over the expected cases of X and Y if they were considered independent from each other. We can also observe that there is a high association among the elements on either side of the rule, noted by the conviction values.

Interesting relationships have been found among the attributes observed such as that a high pulse duration at the edge of a call can imply a high pulse time until the peak of the same call, helping species classification. Finally, the mentioned parameters are the only ones from the 71 attributes that have a high dependency on each other, which can lead us to do further Nave Bayes testing with a dataset, in which these attributes have been removed.

5. Conclusions and Future Work

The increase of performance obtained through the combination of decision tree algorithms and Nave Bayes is due to the elimination of redundant information performed by these algorithms, although an 88.23% or 90.13% are still not enough for reliable classification. Further work is required and a combination of association rules with these algorithms might prove valuable.

We can mention some advantages of data mining over HMMs, which might be crucial in order to obtain high precision reliable results. The first of these advantages is that Hidden Markov Models required us to partition the sound files and to place bird songs close together in order to get efficient results. Different audio files had to be chopped down and cut in order to compose a multi call file for the HMMs. Although this is not an extremely complex procedure it does limit the performance of applying these models on live sensor networks, since the general goal for a near future is to do analysis in real time. These HMMs were also extremely sensitive to ambient noise in the recordings during the training phase.

On the other hand, the data mining approach suffered very little with these inconveniences. Songs were processed in full without cutting or placing calls together. Only low and high pass software filters were applied to the original recordings when used with the data mining algorithms. In this way we removed some of the ambient noise, which left little margin for errors. This is because we were looking for certain attributes from each species, which repeated constantly among the recordings. We also noticed that the computational power to use very large data sets when working with data mining was very low in contrast with the one required by the HMM approach.

As an integration of the two methods, we propose based on our current experiments, to use feature extraction to target HMMs input parameters in order to optimize the recognition process. We propose for future work, to vary the input parameters given to HTK in order to train the HMMs with different data models. These models will be obtained through Data Mining. As a final approach, the recognition results of the HMMs will be compared amongst each other in order to see how efficient is the targeting selection performed by the Data Mining on their input parameters. The proposed model is as follows:

1. Run HTK.

2. Register the results in a small database.

3. A *C* program will perform the feature extraction from the sound signal through the FFTs of the signal instead of using the Perceptual Linear Prediction and MFCC used normally by HTK [22].

4. The *C* program will get the parameters extracted in the previous step and will process them by means of generating decision trees. With the decision trees we will be able to prune down the attributes from the extracted base. The output of this step will include the most significant attributes used for classification.

5. Finally the program will convert the output from the previous step into a binary file to be used with HTK with an optional modification.

Finally we plan to take this work into the field and test our algorithms using adapted beamforming microphones with sensor networks and performing live monitoring and classification, expecting to see if our results hold, considering ambient noise and tropical weather interference.

6. Acknowledgments

References

[1] Cornell lab of ornithology macaulay library. http://birds.cornell.edu/MacaulayLibrary/.

[2] S. C. Bard, M. Hau, M. Wikelski, and J. C. Wingfield. Vocal distinctiveness and response to conspecific playback in the spotted antbird, a neotropical suboscine. In T. C. O. Society, editor, *The Condor*, volume 104, pages 387–394, 2002.

[3] J. P. Campbell. Speaker recognition: a tutorial. In IEEE, editor, *Proceedings of the IEEE*, volume 85, pages 1437–1462, 1997.

[4] C. K. Catchpole and P. J. Slater. *Bird Song Biological Themes and Variations*. Cambridge University Press, April 1995.

[5] E. Chang, J. Zhou, S. Di, C. Huang, and K.-F. Lee. Large vocabulary mandarin speech recognition with different approaches in modeling tones. In *International Conference on Spoken Language Processing, 2000 (ICSLP-2000)*, pages 983–986, Beijing, China, October 2000.

[6] M. Eichner and M. Wolff. Data-driven generation of pronunciation dictionaries in the german verbmobil project - discussion of experimental results. In IEEE, editor, *Acoustics, Speech, and Signal Processing, 2000 (ICASSP '00)*, volume 3, pages 1687–1690, Istanbul, Turkey, June 2000.

[7] W. O. Galitz. *Handbook of Screen Format Design*. Q.E.D. Information Sciences, 1981.

[8] J. Gao, H.-F. Wang, M. Li, and K.-F. Lee. A unified approach to statistical language modeling for chinese. In IEEE, editor, *Acoustics, Speech, and Signal Processing, 2000 (ICASSP '00)*, volume 3, pages 1703–1706, Istanbul, Turkey, June 2000.

[9] M. Gridi. Sound ruler. http://soundruler.sourceforge.net/.

[10] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall PTR, 1st edition, April 2001.

[11] X. Huang and et. al. Mipad: A next generation pda prototype. In *International Conference on Spoken Language Processing, 2000 (ICSLP-2000)*, volume 3, pages 33–36, Beijing, China, October 2000.

[12] K. Laurila and P. Haavisto. Name dialing - how useful is it? In IEEE, editor, *Acoustics, Speech, and Signal Processing, 2000 (ICASSP '00)*, volume 6, pages 3731–3734, Istanbul, Turkey, June 2000.

[13] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson. Wireless sensor networks for habitat monitoring. In ACM, editor, *International Workshop on Wireless Sensor Networks and Applications, 2002 (WSNA'02)*, Atlanta, Georgia, September 2002.

[14] D. A. Nelson. The importance of invariant and distinctive features in species recognition of bird song. In T. C. O. Society, editor, *The Condor*, volume 91, pages 120–130, 1989.

[15] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Englewood Cliffs, New Yersey, 2nd edition, December 2002.

[16] Y. Sagisaka and L. Lee. Speech recognition of asian languages. In IEEE, editor, *Automatic Speech Recognition and Understanging, 1995 (ASRU '95)*, pages 55–57, Snowbird, Utah, December 1995.

[17] R. Szewczyk, E. Osterweil, J. Polastre, M. Hamilton, A. Mainwaring, and D. Estrin. Habitat monitoring with sensor networks. In ACM, editor, *Communications of the ACM*, volume 47, pages 34–40, June 2004.

[18] C. Taylor, E. Stabler, and M. Hamilton. Sensor arrays for acoustic monitoring of bird behavior and diversity. http://www.nsf.gov/awardsearch/showAward.do? AwardNumber=0410438, 2004-2009.

[19] V. Trifa. A framework for bird songs detection, recognition and localization using acoustic sensor networks. Master's thesis, École Polytechnique Fédérale de Lausanne, 2006.

[20] M. Wilde and V. Menon. Bird call recognition using hidden markov models. Technical report, EECS Department Tulane University, 2003.

[21] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, California, 2nd edition, June 2005.

[22] S. Young and et. al. *The HTK Book for HTK Version 3.4*. Cambridge University Engineering Department., December 2006.