Automatic acoustic classification of insect species based on directed acyclic graphs

Stavros Ntalampiras
Department of Computer Science, University of Milan, via Celoria 18, 20133, Milano MI, Italy
stavros.ntalampiras@unimi.it

Abstract: This work presents the design of a directed acyclic graph (DAG) scheme, the nodes of which incorporate hidden Markov models (HMMs) for classifying insect species. Such a DAG scheme is able to limit the problem space, while having the HMMs capture the temporal evolution of Mel-scaled spectrograms extracted out of wingbeat sounds. Interestingly, the proposed approach offers interpretability of the classification process by inspecting the sequence of edges activated in the DAG (path). The dataset encompasses 50 000 wingbeat sounds representing six species, i.e., *Ae. aegypti* (male and female), *Cx. quinquefasciatus* (male and female), *Cx. stigmatosoma* (male and female), *Cx. tarsalis* (male and female), *Musca domestica*, and *Drosophila simulans*, and is publicly available at https://sites.google.com/site/insectclassification/. Thorough species classification experiments showed that the proposed solution outperforms state-of-the-art approaches.

Date Received: April 22, 2019 **Date Accepted:** May 31, 2019

1. Introduction

Classification of insect species may assist pest management and control significantly not only in terms of biodiversity assessment and cataloging, but also financially since the presence/absence of certain species may lead to catastrophic economic losses (Potamitis *et al.*, 2015). Importantly, mosquitoes comprise a great animal threat to human health as they directly contribute to the transmission of deadly diseases such as yellow fever, malaria, dengue, and the Zika virus (Pile, 2018). Such relevance has driven many researchers to develop mechanisms automatizing the process of identification (Chen *et al.*, 2014; Zhang *et al.*, 2017). Interestingly, the work reported in Chen *et al.* (2014) apart from presenting a Bayesian classifier for insect species, presents a publicly available dataset focusing on Diptera, thus facilitating research in the field. The dataset encompasses 50 000 wingbeat sounds representing six species, i.e., *Ae. aegypti* (male and female), *Cx. quinquefasciatus* (male and female), *Cx. stigmatosoma* (male and female), *Cx. tarsalis* (male and female), *Musca domestica*, and *Drosophila simulans*. The creators employed optical sensors to record the "sound" of insect flight; such types of recordings remain completely unaffected by potential environmental sound/noise interferences. Following the same line of thought, this letter proposes a directed acyclic graph (DAG) suitably dividing the problem space, while modeling the temporal patterns existing in the spectral content of wingbeat sounds. The proposed solution encompasses the following steps: (a) feature extraction, (b) DAG construction including its topological ordering, (c) establishment of the node-based classifiers, and (d) insect species identification based on DAG's operation.

2. Feature set

In essence, we used a Mel-scaled spectrogram to represent the available signals. To this end, we employed a triangular Mel filterbank for extracting 23 log-energies. At first, the audio signal is windowed and the short-time Fourier transform (STFT) is computed with respect to each window. The outcome of the STFT passes through the filterbank and the logarithm is computed to adequately space the data. We avoid using the discrete cosine transform, which might lead to information loss, thus we exploit the entire content of each wingbeat Mel-spectrum. Figure 1 demonstrates Mel-scaled spectrograms extracted out of recordings coming from all available classes.

3. DAG-based scheme

The proposed framework relies on the DAG logic, i.e., the classification scheme is a graph denoted as $\mathcal{G} = \{N, L\}$, where $N = \{n_1, \ldots, n_m\}$ represents the nodes and $L = \{l_1, \ldots, l_k\}$ represents the links associating the nodes. Each node in N is responsible for a binary classification task conducted via a set of hidden Markov models (HMMs)

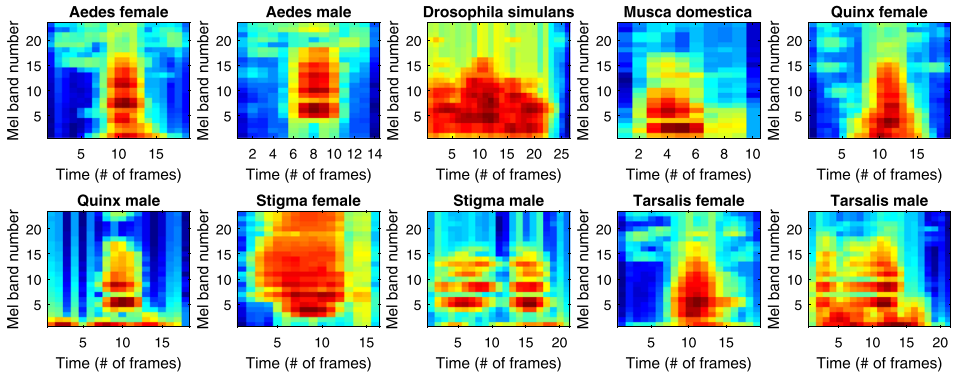

Fig. 1. (Color online) Mel-scaled spectrograms extracted out of recordings coming from all available classes [*Ae. aegypti* (male and female), *Drosophila simulans*, *Musca domestica*, *Cx. quinquefasciatus* (male and female), *Cx. stigmatosoma* (male and female), and *Cx. tarsalis* (male and female)].

which fit well the specifications of audio pattern recognition tasks (Ntalampiras, 2014), thus the DAG-HMM notation.

The motivation behind creating such a graph-based classification system is that in this way, one is able to limit the problem space and design classification algorithms for two mutually-exclusive classes than having to deal with the entirety of the different classes at the same time. Essentially, the proposed methodology breaks down any C_m-class classification problem into a series of 2-class classification problems.

DAGs can be seen as a generalization of the class of Decision Trees, while the redundancies and repetitions that may occur in different branches of the tree can be observed more efficiently since different decision paths might be merged. In addition, DAGs are able to collect and conduct a series of tasks in an ordered manner, subject to constraints that certain tasks must be performed earlier than others. The sequential execution of tasks is particularly important and directly related to the efficacy with which the overall task is addressed (VanderWeele and Robins, 2010).

The DAG-HMM architecture used in this paper includes $m(m-1)/2$ nodes, each one associated with a 2-class classification problem. The connections between the different nodes in $\mathcal{G}$ have only one orientation without any kind of loop(s). As a result, each node of such a so-called *rooted* DAG has either 0 or 2 leaving arcs.

Sections 4 and 5 provide a detailed analysis of the way the DAG-HMM is constructed and subsequently operates. The principal issue associated with the design of every DAG is the *topological ordering*, i.e., ordering the nodes in a way that the starting endpoints of every edge occur earlier than the corresponding ending endpoints. In the following, we describe how such a topological ordering is discovered based on the Kullback–Leibler divergence.

4. Determining the topological ordering of the DAG-HMM

Naturally, one would expect that the performance of the DAG-HMM depends on the order in which the different classification tasks are conducted. This was also evident from early experiments. This observation motivated the construction of the DAG-HMM so that "simple" tasks are executed earlier in the graph. In other words, these are placed in the top nodes of the DAG-HMM, in a way that classes responsible for a high amount of misclassifications are discarded early in the graph operation. In order to get an early indication of the degree of difficulty of a classification task, we employed the metric representing the distance of the involved classes in the probabilistic space, i.e., the Kullback–Leibler Divergence (KLD) between per-class Gaussian mixture models (GMMs) in the feature space. The basic motivation is to place early in the DAG-HMM tasks concerning the classification of classes with large KLD, as they could be completed with high accuracy. The scheme determining the topological ordering is illustrated in Fig. 2. There, a simplified two-dimensional feature space (facilitating demonstration purposes) is used where the KLD distances are computed and subsequently sorted.

The KLD between two J-dimensional probability distributions A and B is defined as (Taylor, 2006)

$$D(A||B) = \int_{R^J} p(X|A) \log \frac{p(X|A)}{p(X|B)} dx. \tag{1}$$

KLD provides an indication of how distant two models are in the probabilistic space. It is important to note that KLD as given in Eq. (1) comprises an asymmetric

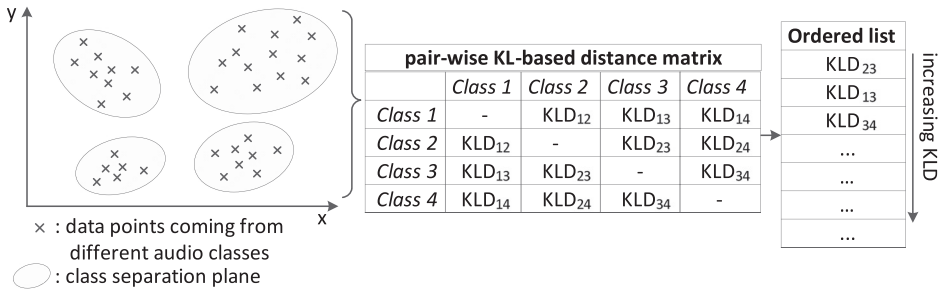

Fig. 2. The determination of the topological ordering (for simplicity, only four classes are considered).

quantity. The symmetrical form can be inferred by simply adding the integrals in both directions, i.e.,

$$D_s(A||B) = D(A||B) + D(B||A). \tag{2}$$

In the special case where both A and B are Gaussian mixture models KLD can be defined as follows:

$$\mathrm{KLD}(A||B) = \int A(x)\log\frac{B(x)}{A(x)}dx. \tag{3}$$

Unfortunately, there is not a closed-form solution for Eq. (3), thus we employed the empirical mean as follows:

$$\mathrm{KLD}(A||B) \approx \frac{1}{n}\sum_{i=1}^{n}\log\frac{B(x_i)}{A(x_i)}, \tag{4}$$

given that the number of Monte Carlo draws is sufficiently large. During our experiments we set $n = 2000$.

It should be noted the KLD between HMMs was not used since computing distances between HMMs of unequal lengths, which might be common in this work as HMMs representing different classes might have different number of states, can be significantly more computationally demanding without a corresponding gain in modeling accuracy (Liu *et al.*, 2007; Zhao *et al.*, 2007).

After computing the KLD for the different pairs of classes, i.e., reach the second stage depicted in Fig. 2, the KLD distances are sorted in an increasing manner. This way the topological ordering of the DAG-HMM is revealed, placing the classification tasks of low difficulty on its top. Each node removes a class from the candidate list until there is only one class left, which comprises the DAG-HMM prediction. The elements of the distance matrix could be seen as early performance indicators of the task carried out by the corresponding node. The proposed topological ordering places tasks likely to produce misclassifications at the bottom of the graph. This process outputs a *unique* solution for the topological sorting problem, as it is usually met in the graph theory literature (Cook, 1985).

5. The DAG-HMM operation

The operation of the proposed DAG-HMM scheme is the following: after extracting the features of the unknown audio signal, the first/root node is activated. More precisely, the feature sequence is fed to the HMMs, which produce two log-likelihoods showing the degree of resemblance between the training data of each HMM and the unknown one. These are compared and the graph flow continues on the larger log-likelihood path. It should be stressed that the HMMs are optimized (in terms of the number of states and Gaussian components) so that they address the task of each node optimally. That said, it is possible that a specific class is represented by HMMs with different parameters when it comes to different nodes of the DAG-HMM.

An example of a DAG-HMM addressing a problem with four classes is illustrated in Fig. 3. The remaining classes for testing are mentioned beside each node. Digging inside each node, Fig. 3 shows the HMM-based sound classifier responsible for activating the path of the maximum log-likelihood.

The operation of the DAG-HMM may be parallelized with that of investigating a list of classes, where each level eliminates one class from the list. In more detail, in the beginning the list includes all the potential audio classes. At each node the feature sequence is matched against the respective HMMs and the model with the lowest

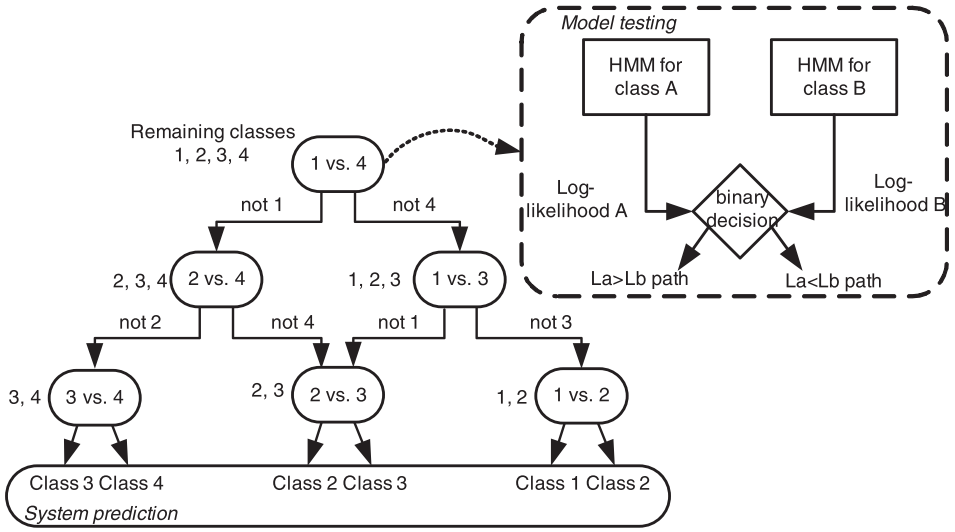

Fig. 3. An example of a DAG-HMM addressing a problem with four classes. The binary operation carried out in each node of the DAG-HMM is also depicted.

log-likelihood is erased from the list, while the DAG-HMM proceeds to the part of the topology without the discarded class. This process terminates when only one class remains in the list, which comprises the system's prediction. Hence, in case the problem deals with m different classes, the DAG's decision will be made after the evaluation of $m - 1$ nodes.

6. Experimental setup and results

In this section, we analyze the: (a) wingbeat dataset including ten classes, (b) parametrization of both DAG-HMM and feature extraction module, (c) contrasted approaches, and (d) we present and comment the achieved results.

The low-level feature extraction window is 30 ms with 10 ms overlap, so that the system is robust against possible misalignments. The sampled data are Hamming windowed to smooth potential discontinuities while the fast Fourier transform (FFT) size is 512. Standard normalization techniques, i.e., mean removal and variance scaling, were applied.

The HMMs of each node are optimized in terms of number of states and nodes following the Expectation-Maximization and Baum Welch algorithms (Rabiner, 1989). As the considered sound events are characterized by a distinct time evolution, we employed HMMs with left–right topology, i.e., only left to right state transitions are permitted. Moreover, the distribution of each state is approximated by a Gaussian mixture model of diagonal covariance, which may be equally effective to a full one at a much lower computational cost (Reynolds and Rose, 1995).

The maximum number of k-means iterations for cluster initialization was set to 50 while the Baum–Welch algorithm used to estimate the transition matrix was bounded to 25 iterations with a threshold of 0.001 between subsequent iterations. The number of explored states ranges from 3 to 7 while the number of Gaussian components used to build the GMM belongs to the {2, 4, 8, 16, 32, 64, 128, 256, and 512} set. The final parameters were selected based on the maximum recognition rate criterion. The machine learning package Torch (freely available at http://torch.ch/) was used to construct and evaluate GMMs and HMMs. Furthermore, MATLAB was employed for extracting the Mel-spectrograms and setting up the DAG.

Here, we address two crucial points toward comparability and reproducibility of the proposed approach: (a) we employed a publicly available dataset (Chen *et al.*, 2014), and (b) our experiments were carried out based on the cross-validation protocol adopted in Chen *et al.* (2014) and Zhang *et al.* (2017).

Table 1. The recognition rates for the proposed and contrasted methods on the species identification task.

Approach	Recognition rate (%)
AlexNet-BN + LIBSVM (Chen *et al.*, 2014)	76.12
FFT + LIBSVM (Zhang *et al.*, 2017)	74.74
DAG-HMM	**80.7**

Table 2. The confusion matrix (in %) achieved by the DAG-HMM approach.

Presented \ Responded	*aedes* (F)	*aedes* (M)	*drosophila*	*quinx* (F)	*quinx* (M)	*stigma* (F)	*stigma* (M)	*tarsalis* (F)	*tarsalis* (F)	*tarsalis* (F)
aedes(F)	**86.4**	0.8	0.2	0.4	6.2	—	3	2.2	—	0.8
aedes(M)	—	**98**	0.4	0.2	—	0.4	0.4	—	—	0.6
drosophila	0.4	—	**85.2**	9.6	1	0.2	2	0.2	1.2	0.2
musca dom.	0.2	0.2	3.6	**86.8**	0.2	1	6.6	0.2	1	0.2
quinx (F)	20.2	0.2	0.6	—	**69.4**	—	8.6	—	0.4	0.6
quinx(M)	0.4	2.8	0.4	2.4	—	**66**	0.6	1	0.6	25.8
stigma(F)	—	—	4.4	8.8	3.2	0.8	**71.4**	3.6	6	1.8
stigma(M)	1	—	0.6	0.8	—	0.4	0.6	**94.2**	—	2.4
tarsalis(F)	0.8	0.6	0.4	4	0.6	0.4	8.2	—	**85**	—
tarsalis(M)	—	2.4	1.2	1.2	—	26.6	0.4	3.2	0.4	64.6

Preprocessing of the wingbeat sounds includes removal of any background noise and filling the silent parts with the value 0. Each insect flight sound lasts 1 s.

Classification of insect flight sounds is typically focused on two tasks, i.e., (a) gender and (b) species classification. After the excellent results reported in Zhang *et al.* (2017) regarding the first task, this work concentrates on the second one, which comprises a ten-class problem. The dataset includes six insect species, i.e., *Ae. aegypti* (male and female), *Cx. quinquefasciatus* (male and female), *Cx. stigmatosoma* (male and female), *Cx. tarsalis* (male and female), *Musca domestica*, and *Drosophila simulans*. Gender information is included only with respect to four species. Overall, the dataset encompasses 50.000 wingbeat sounds distributed equally among the classes and is publicly available for research purposes at https://sites.google.com/site/insectclassification/.

The resulting topological ordering is presented at https://sites.google.com/site/stavrosntalampiras/demos/insect-species-classification. Table 1 tabulates the average recognition rates achieved by the proposed solution and the ones reported in the literature. As we can see, the one obtained here is the highest one surpassing by 4.58%/2290 recordings and 5.96%/2980 recordings the solutions reported in Zhang *et al.* (2017) and Chen *et al.* (2014), respectively (Table 1). Toward a more thorough analysis of the results, in Table 2 we present the confusion matrix offered after applying the proposed solution. We observe that the highest rate is achieved for the *Aedes male* class (98%), while the worst one for the *Tarsalis male* (64.6%). The latter is misclassified for *Quinx male* 26.6% of the time. Other strong misclassifications concern the pairs *Quinx female-Aedes female* and *Stigma female-House flies*.

7. Conclusion

This paper presented an effective solution for insect species classification using sounds of their wingbeats. Its cornerstone dividing the problem space via a directed graph, the nodes of which carry out binary classification tasks using HMMs. Interestingly, the proposed solution is a comprehensive classification scheme, since its operation does not follow the black-box logic, while one is able to "open" the classifier, and by inspecting the misclassifications, understand the reasons leading to the specific errors. This is a crucial advantage over deep nets, where it is hard to interpret their operation, let alone explain their errors. Finally, new insect classes can be easily incorporated in the proposed scheme as long as the respective data become available.

References and links

Chen, Y., Why, A., Batista, G., Mafra-Neto, A., and Keogh, E. (**2014**). "Flying insect classification with inexpensive sensors," J. Insect Beh. **27**(5), 657–677.

Cook, S. A. (**1985**). "A taxonomy of problems with fast parallel algorithms," Inf. Control **64**(1), 2–22.

Liu, P., Soong, F. K., and Zhou, J. L. (**2007**). "Divergence-based similarity measure for spoken document retrieval," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 4, pp. IV-89–IV-92.

Ntalampiras, S. (**2014**). "Directed acyclic graphs for content based sound, musical genre, and speech emotion classification," J. New Music Res. **43**(2), 173–182.

Pile, D. F. P. (**2018**). "Monitoring mosquitoes," Nature Photonics **12**(5), 254–254.

Potamitis, I., Rigakis, I., and Fysarakis, K. (**2015**). "Insect biometrics: Optoacoustic signal processing and its applications to remote monitoring of McPhail type traps," PLoS One **10**(11), e0140474.

Rabiner, L. R. (**1989**). "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE **77**(2), 257–286.

Reynolds, D. A., and Rose, R. C. (**1995**). "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Process. **3**(1), 72–83.

Taylor, P. (**2006**). "The target cost formulation in unit selection speech synthesis," in *INTERSPEECH*, Pittsburgh, PA (September 17–21).

VanderWeele, T. J., and Robins, J. M. (**2010**). "Signed directed acyclic graphs for causal inference," J. R. Stat. Soc. **72**(1), 111–127.

Zhang, C., Wang, P., Guo, H., Fan, G., Chen, K., and Kämäräinen, J.-K. (**2017**). "Turning wingbeat sounds into spectrum images for acoustic insect classification," Electron. Lett. **53**(25), 1674–1676.

Zhao, Y., Zhang, C., Soong, F. K., Chu, M., and Xiao, X. (**2007**). "Measuring attribute dissimilarity with hmm kl-divergence for speech synthesis," in *6th ISCA Workshop on Speech Synthesis*, Bonn, Germany (August 23–24), pp. 206–210.