Animal Sound Classification using Sequential Classifiers

Javier Romero, Amalia Luque* and Alejandro Carrasco
Escuela Politécnica Superior, Universidad de Sevilla, Sevilla, Spain

Keywords: Sound Classification, Data Mining, Sequential Classifiers, Habitat Monitoring.

Abstract: Several authors have shown that the sounds of anurans can be used as an indicator of climate change. But the recording, storage and further processing of a huge number of anuran's sounds, distributed in time and space, are required to obtain this indicator. It is therefore highly desirable to have algorithms and tools for the automatic classification of the different classes of sounds. In this paper five different classification methods are proposed, all of them based on the data mining domain, which try to take advantage of the sound sequential behaviour. Its definition and comparison is undertaken using several approaches. The sequential classifiers have revealed that they can obtain a better performance than their non-sequential counterpart. The sliding window with an underlying decision tree has reached the best results in our tests, even overwhelming the Hidden Markov Models usually employed in similar applications. A quite remarkable overall classification performance has been obtained, a result even more relevant considering the low quality of the analysed sounds.

1 INTRODUCTION

Climate change is becoming one of the most demanding concerns for the whole humanity. For this reason, many indicators are being defined and used trying to monitor the global warming evolution. Some of these indicators have to do with the warming impact in the biosphere, measuring the change in some animal species population.

Indeed, sound production in ectotherms animals is strongly influenced by the ambient temperature (Márquez and Bosch, 1995) which can affect various features of their acoustic communication system. In fact the ambient temperature, once exceeded a certain threshold, can restrict the physiological processes associated to the sound production even inhibiting behaviour call. As a result, the temperature may significantly affect the patterns of calling songs modifying the beginning, duration and intensity of calling episodes and, consequently, the anuran reproductive activity.

Therefore, the analysis and classification of the sounds produced by certain animal species have revealed as a strong indicator of temperature changes and therefore the possible existence of climate change. Particularly interesting are the results provided by anuran sounds analysis (Llusia et al., 2013).

In previous works (Luque et al., 2016) it has been proposed a non-sequential method for sound classification. According to this procedure, the sound is split up in 10 msec. frames. Next, every frame is featured using 18 MPEG-7 parameters: a vector in $\mathbb{R}^{18}$ (ISO, 2001). Then the frame features are compared to some frame patterns belonging to known species, assigning a class label to each frame. Eventually the sound is classified by frame voting, i.e., the most frequent frame class is assigned to the whole sound.

Comparing frame features to frame patterns is called a supervised classification in the data mining realm. Up to 9 different algorithms have been proposed in (Romero et al., 2016) to make this classification.

However, sounds are inherently made up of sequential data. So, if the frame sequence information is added to the classification process, better classification results should be expected.

*Corresponding author

Romero J., Luque A. and Carrasco A.
Animal Sound Classification using Sequential Classifiers.
DOI: 10.5220/0006246002420247
In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017)*, pages 242-247
ISBN: 978-989-758-212-7

2 METHODS

To introduce the sequential information several methods are proposed.

- Temporal parameters construction (TP) (Schaidnagel et al., 2014). For every frame, a "segment" is considered using the closest neighbour frames. And for every parameter, a new parameter is constructed: the interquartile range in the segment. In this way, up to 36 parameters (a vector in $\mathbb{R}^{36}$) are now identifying a frame, 18 of them including some kind of sequence information. A 10 frame segment is proposed in this study.
- Sliding windows (SW) (Aggarwal, 2007). A short window (e.g. with 5 frames), centred in each frame, is considered. Now the parameters featuring each frame (e.g. 5 parameters) are those corresponding to all the frames under the window. In the example, each frame is featured using 5x5 parameters (a vector in $\mathbb{R}^{5x5}$).
- Recursive sliding windows (RSW) (Dietterich, 2002). It is a method similar to the previous one, but now the classifier considers not only the parameters of the frame under the window, but also their classification results.
- Hidden Markov Models (HMM) (Rabiner, 1989). It is a genuine sequential classifier. The sequence of frame parameters is considered to be obtained as the result of an HMM made up of hidden states emitting observed data. This is the classifier recommended in the MPEG-7 standard.
- Autoregressive integrated moving average models (ARIMA) (Box et al., 2011). It is also a genuine sequential classifier. The sequence of frame parameters is considered the result of an ARIMA time series. For a certain sound, the coefficients of the time series are computed and, in a second step, these coefficients are classified using non-sequential classifiers.

Most of this sequential classifier (all except the HMM) are relying on an underlying non-sequential classifier. A broad and representative selection of them has been used through this paper:

- Minimum distance (Wacker and Landgrebe, 1971);
- Maximum likelihood (Le Cam, 1979);
- Decision trees (Rokach et al., 2008);
- k-nearest neighbour (Cover and Hart, 1967);
- Support vector machine (SVM) (Cristianini and Shawe-Taylor, 2000);
- Logistic regression (Dobson and Barnett, 2008);
- Neural networks (Du and Swamy, 2013);
- Discriminant function (Härdle and Simar, 2012);
- Bayesian classifiers (Hastie et al., 2005).

Sequential classifiers significantly increases the number of parameters required. To cope with this drawback, a reduction on the number of original MPEG-7 parameters is proposed, considering the 5 most significant features (leading to a vector in $\mathbb{R}^5$). Feature selection procedures are employed to determine this reduced set.

For comparison reasons, 2 non-sequential methods are also considered:

- Non-sequential classification based on 18 MPEG-7 parameters (NS-18).
- Non-sequential classification based on the 5 most relevant MPEG-7 parameters (NS-5).

To compare the results obtained for every classifier, several metrics for the performance of a classifier can be defined (Sokolova and Lapalme, 2009), all of them based on the confusion matrix. The most relevant indicators and their definitions are the following:

- Accuracy: Overall effectiveness of a classifier;
- Error rate: Classification error;
- Precision: Class agreement of the data labels with the positive labels given by the classifier;
- Sensitivity: Effectiveness of a classifier to identify positive labels;
- Specificity: How effectively a classifier identifies negative labels.

Additionally, a graphical way to compare classifiers will be used, representing their performance in the Receiver Operating Characteristic (ROC) space (Powers, 2011), where the True Positive Rate (sensitivitiy) of a classifier is plotted versus its False Positive Rate (defined as one minus the specificity).

3 RESULTS

For testing purposes, sound files provided by the Zoological Sound Library (Fonozoo, 2016) have been used, corresponding to 2 species, the epidalea calamita (natterjack toad) and alytes obstetricans (common midwife toad), with a total of 63 recordings containing 3 classes of sounds:

- Epidalea calamita; mating call (23 records)
- Epidalea calamita; release call (10 records)

- Alytes obstetricans (30 records)

In total 6,053 seconds (1h:40':53") of recording have been analysed, with a 96 seconds (1':36") average file length, and a 53 seconds median.

A common feature of all recordings is that they have been made in the natural habitat, with very significant surrounding noise (wind, water, rain, traffic, voice...), which means an additional challenge in the signal processing.

The results obtained by the non-sequential classifiers based on 18 MPEG-7 parameters are compared using the ROC analysis which is depicted in Figure 1. The best result corresponds to the decision tree classifier with an accuracy of 91.53%.

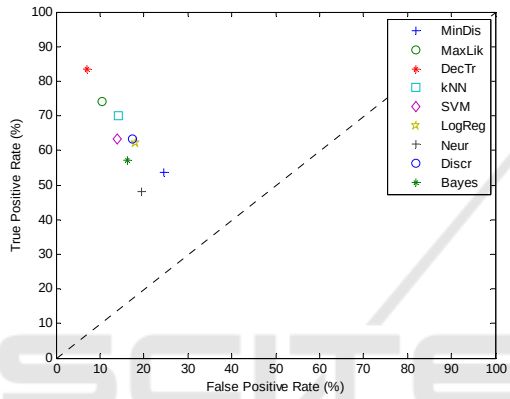

Figure 1: ROC analysis for non-sequential classifiers based on 18 MPEG-7 parameters.

The results obtained by the non-sequential classifiers based on the 5 most relevant MPEG-7 parameters are also compared using the ROC analysis which is depicted in Figure 2. The best result corresponds to the decision tree classifier with an accuracy of 89.42%.

The results obtained by the temporal parameter construction approach are now considered. The corresponding ROC analysis is depicted in Figure 3. The best result corresponds to the decision tree classifier with an accuracy of 91.53%.

Focusing now on the sliding window method, its results are compared through the ROC analysis and presented in Figure 4. The best result corresponds to the decision tree classifier with an accuracy of 90.48%.

Once more, the results obtained by the recursive sliding window method are compared using the ROC analysis which is portrayed in Figure 5. The best result corresponds to the decision tree classifier with an accuracy of 73.54%.

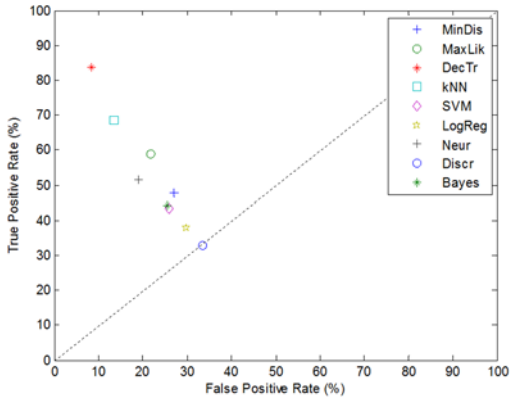

Figure 2: ROC analysis for non-sequential classifiers based on the 5 most relevant MPEG-7 parameters.

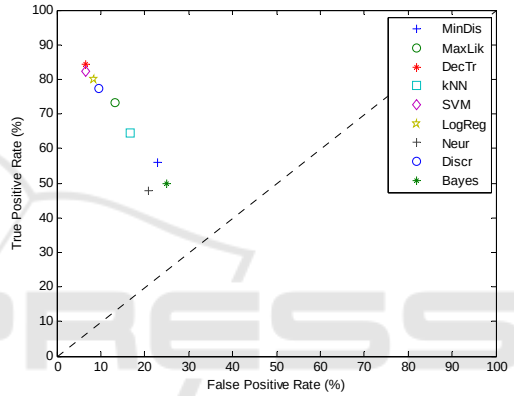

Figure 3: ROC analysis for sequential classifiers using temporal parameter construction.

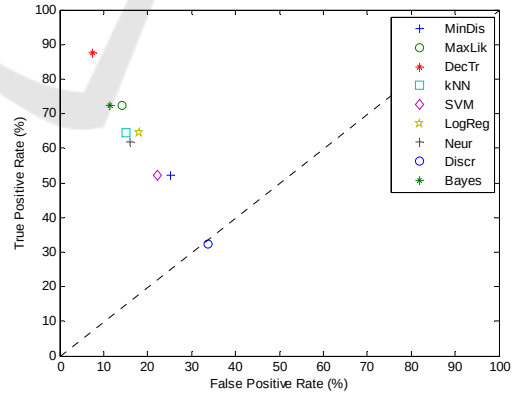

Figure 4: ROC analysis for sequential classifiers using sliding windows.

Table 1: Performance metrics.

Method		Number of features	ACC	PRC	SNS	SPC
Non-sequential		18	91.53%	87.05%	83.43%	93.01%
		5	89.42%	85.40%	83.77%	91.52%
Temporal parameters		36	91.53%	91.40%	84.44%	93.33%
Sliding windows		5x5	90.48%	87.23%	87.44%	92.53%
Recursive sliding windows		5x5	73.54%	51.57%	62.22%	74.92%
HMM	Over the file	5	84.13%	70.24%	61.64%	85.58%
	Over ROI length segments	5	59.79%	-	40.00%	68.54%
	Over 5 sliding windows	5	69.31%	56.42%	52.75%	78.13%
ARIMA		5	70.37%	63.17%	52.66%	77.48%

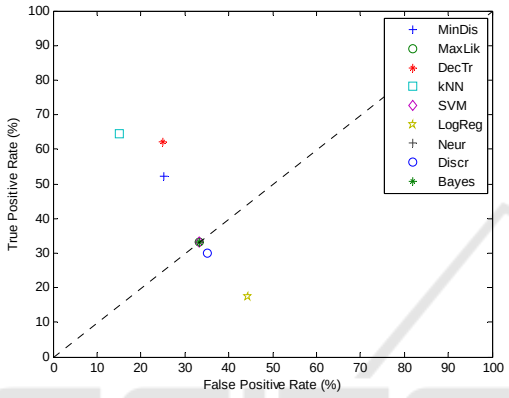

Figure 5: ROC analysis for sequential classifiers using recursive sliding windows.

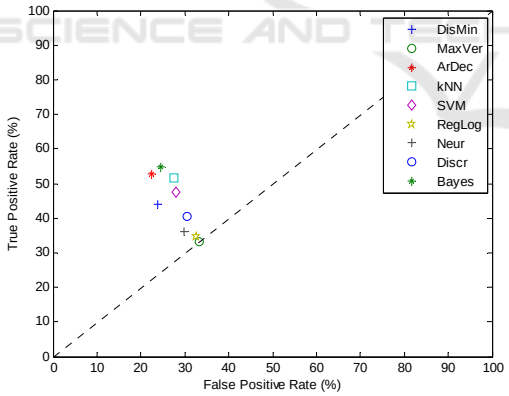

Figure 6: ROC analysis for sequential classifiers using ARIMA models.

The HMM is the only of the proposed methods not relaying in other classifiers. Therefore, it is not a ROC analysis to compare the non-existent classifiers. The HMM takes a sound segment and try to classify it as a whole, not framing it. When a sound file has to be classified 3 alternatives have been explored to determine the segment length:

- The full file (HMM-F). This approach obtains an accuracy of 84.13%.
- A segment with the same length that the ROI mean length (HMM-ROI). The Regions Of Interest (ROIs) are the segments of the sound patterns containing a valid sound (no silence or noise). This approach obtains an accuracy of 59.79%.
- A segment defined by a sliding window of a certain length (HMM-SW). It is proposed to use 5 frames in the analysis. This approach obtains an accuracy of 69.31%.

Eventually, the results obtained by the ARIMA method are also compared using the ROC analysis which is illustrated in Figure 6. The best result corresponds to the decision tree classifier with an accuracy of 70.37%.

Until now, partial results have been presented for every sequential method. To obtain an overall perspective, a comparison of the different methods proposed for sequential classifiers is presented in Table 1, where the non-sequential classifiers are also considered for contrasting reasons. Additionally, a ROC analysis has also been accomplished for every method and its results are depicted in Figure 7.

Considering now the best sequential method (sliding window) and the underlying best classifier (decision tree) the overall results can be examined. The confusion matrix is presented in Table 2.

Additionally the classification profile is depicted in Figure 8. It can be seen that an overall success rate of 85.71% is obtained, with quite balanced values for every class. These results are considered a good performance, furthermore when considering the low quality of the sound records.

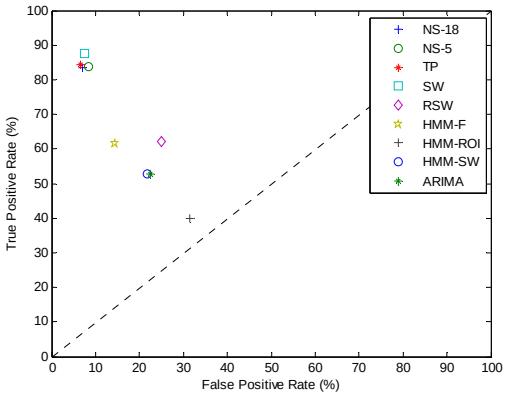

Figure 7: ROC analysis for sequential methods.

Table 2: Confusion matrix for the sliding window method with decision tree classifier.

		Classification obtained		
		1	**2**	**3**
Data class	**1**	20	0	3
	2	3	7	0
	3	1	1	28

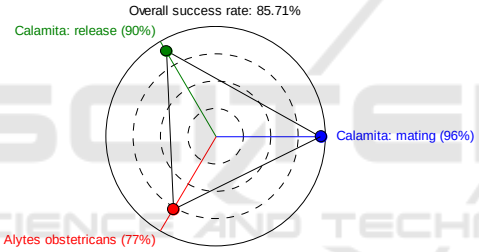

Figure 8: Classification profile for the sliding window method with decision tree classifier.

4 CONCLUSIONS

From the above results several conclusions can be reached. Firstly, they have shown that at least 2 sequential methods, the temporal parameter construction and the sliding window, slightly enhance the non-sequential classification, at least in terms of ROC analysis.

The sliding window appears as the most efficient approach and show the best metrics among the classifiers based on the same number of parameters (5 parameters).

When the sequential method relies on another classifier, the decision tree algorithm stands out in almost every case. The only exception is the recursive sliding window method, where decision tree is the second best classifier (after the k-nearest neighbour).

The sliding window with an underlying decision tree classifier reaches a remarkable overall success rate (85.71%), a figure even more relevant considering the low quality of the analysed sounds. Its results clearly overcome the performance obtained, in any of the studied variants, through the Hidden Markov Models, the MPEG-7 proposed technique.

ACKNOWLEDGEMENTS

This work has been supported by the Consejería de Innovación, Ciencia y Empresa, Junta de Andalucía, Spain, through the excellence eSAPIENS (reference number TIC-5705). The authors would like to thank Rafael Ignacio Marquez Martinez de Orense (Museo Nacional de Ciencias Naturales) and Juan Francisco Beltrán Gala (Faculty of Biology, University of Seville) for their collaboration and support.

REFERENCES

Aggarwal, C. C. (2007). *Data streams: models and algorithms* (Vol. 31). Springer Science and Business Media.

Box, G. E., Jenkins, G. M., Reinsel, G. C. (2011). *Time series analysis: forecasting and control* (Vol. 734). John Wiley and Sons.

Cover, T. M., Hart, P. E. (1967). Nearest neighbor pattern classification. Information Theory, IEEE Transactions on, 13(1), 21-27.

Cristianini, N., Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press.

Dietterich, T. G. (2002). Machine learning for sequential data: A review. In *Structural, syntactic, and statistical pattern recognition* (pp. 15-30). Springer Berlin Heidelberg.

Dobson, A. J., Barnett, A. (2008). An introduction to generalized linear models. CRC press.

Du, K. L., Swamy, M. N. S. (2013). Neural Networks and Statistical Learning. Springer Science and Business Media.

Fonozoo.com (2016). Retrieved from http://www.fonozoo.com/

Härdle, W. K., Simar, L. (2012). Applied multivariate statistical analysis. Springer Science and Business Media.

Hastie, T., Tibshirani, R., Friedman, J. (2005). The elements of statistical learning: data mining, inference and prediction. Springer-Verlag.

ISO (2001). *ISO 15938-4:2001 (MPEG-7: Multimedia Content Description Interface), Part 4: Audio*. ISO.

Le Cam, L. M. (1979). Maximum likelihood: an introduction. Statistics Branch, Department of

Mathematics, University of Maryland.

Llusia, D., Márquez, R., Beltrán, J. F., Benitez, M., Do Amaral, J. P. (2013). Calling be-haviour under climate change: geographical and seasonal variation of calling temperatures in ectotherms. *Global change biology*, 19(9), 2655-2674.

Luque, J., Larios, D. F., Personal, E., Barbancho, J., León, C. (2016). Evaluation of MPEG-7-Based Audio Descriptors for Animal Voice Recognition over Wireless Acoustic Sensor Networks. *Sensors*, 16(5), 717.

Márquez, R., Bosch, J. (1995). Advertisement calls of the midwife toads Alytes (Amphibia, Anura, Discoglossidae) in continental Spain. *Journal of Zoological Systematics and Evolutionary Research*, 33(3-4), 185-192.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37-63.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286.

Rokach, Lior, Maimon, O. (2008). Data mining with decision trees: theory and applications. World Scientific Pub Co Inc.

Romero, J., Luque, A., Carrasco, A. (2016). Anuran sound classification using MPEG-7 frame descriptors. *XVII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA)*, 801-810.

Schaidnagel, M., Connolly, T., Laux, F. (2014). Automated feature construction for classification of time ordered data sequences. *International Journal on Advances in Software*, 7(3 and 4), 632-641.

Sokolova, M., Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4), 427-437.

Wacker, A. G., Landgrebe, D. A. (1971). The minimum distance approach to classification. Purdue University. Information Note 100771.